U.S. Department of Homeland Security

SCIENCE AND TECHNOLOGY DIRECTORATE

Operational Testing of Face Recognition Systems: Challenges, New Approaches and Developing International Standards

Science and Technology

Arun Vemury
Senior Advisor

Biometric & Identity Technology Center

John J. Howard
Chief Data Scientist,

The Identity and Data Sciences Lab, Maryland Test Facility

April 2025

Disclaimer

- This research was funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034.

- This work was performed by the Identity and Data Sciences Laboratory team at the Maryland Test Facility.

- The views presented here are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers.

- The data used in this research was acquired under IRB protocol.

Science and Technology

Agenda

- Kinds of Testing
- Classic Operational Testing
- The Need for New Approaches
- New Approaches
- Recent use in Execution of DHS Management Directive 026-11
- The Need to Standardize Internationally

Science and Technology

Biometric & Identity Technology Center

The Science & Technology Directorate (S&T) conducts foundational research to ensure advancements in science and technology are harnessed in the development of cutting-edge solutions to new and emerging operational challenges.

- Drive biometric and identity innovation at the Department of Homeland Security (DHS) through Research, Development, Test, and Evaluation (RDT&E) capabilities.
- Facilitate and accelerate understanding of biometrics and identity technologies for new, DHS use cases.
- Drive efficiencies by supporting cross-cutting methods, best practices and solutions across programs.
- Deliver subject matter expertise across the DHS enterprise.
- Engage industry and provide feedback.
- Encourage innovation across industry and academia.

AI Testing in 2025

- **A Surge in AI System Deployments**
  - In 2025, industry analysts, including Gartner, predict a major shift from AI pilot programs to **full-scale AI system deployments**.

- **AI-Driven Acceleration of Product Timelines**
  - AI-powered development significantly reduces product development timelines:
    - Medicine – Rapid drug discovery
    - Software – Shortened development cycles
    - Virtual Assistants & AI Agents – Greater automation
    - Identity & Security Products – More use-cases

Science and Technology

AI Testing in 2025

- **System Testing: The Emerging Bottleneck**
  - With more **AI-powered products reaching deployment faster than ever**, the primary challenge shifts from development to **validating and testing** these systems in real-world conditions to ensure effectiveness.

- **Options to Meet Current/Future Demand:**
  - Option 1: Scale **pre-launch** testing considerably (good, fast, or cheap – pick two). Will take time and resources.
  - Option 2: Move some testing activities to the **post-launch** phase. Be <u>in the loop, not in the way</u>.

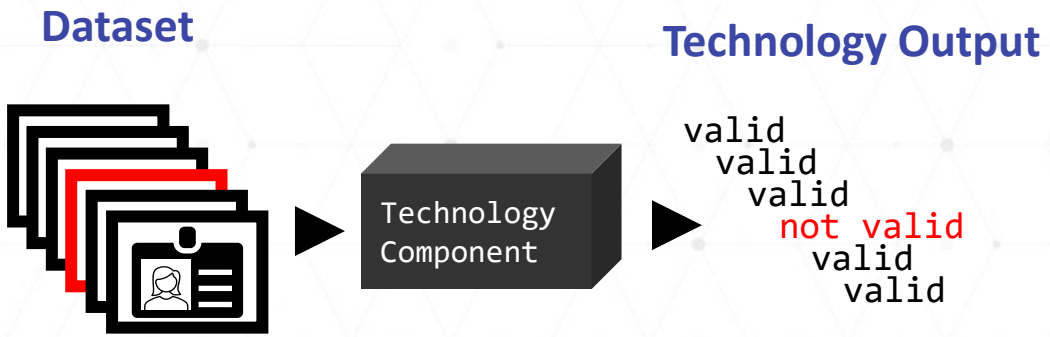
Science and Technology

Kinds of Biometric Evaluations

- **Technology Testing:**
  - Focused on a specific system component
  - Re-use of biometric datasets, often sequestered
  - Larger sample size
  - Enables the highest level of control, easily repeatable
  - Low cost.

  - Answers questions about how technologies advance or perform relative to each other.
  - Answers questions about the limits of a technology's performance.

  - E.g., What is the minimum false match rate achievable by face recognition technology?



Science and Technology

Kinds of Biometric Evaluations

- **Scenario Testing:**
  - Focused on a use-case
  - Full biometric system
  - Gathers new biometric samples
  - Can enable high levels of control
  - Smaller sample size. Important to delineate the effect size you can reliably detect.
  - Costly to repeat
  - Answers questions about technology performance and suitability in an intended use.
  - E.g., How will face recognition perform in a high-throughput unattended scenario? Are user's able to safely interact with the technology.

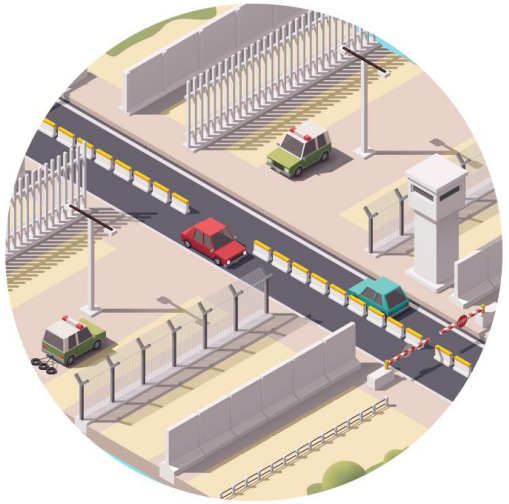
Science and Technology

Kinds of Biometric Evaluations

- **Operational Testing:**
  - Focused on a specific operational system,
  - Full biometric system,
  - Gathering new biometric samples, from the deployed system
  - Has less control and ground-truth information,
  - Larger sample size.

  - Answers questions about how an operational system performs and how that performance may vary under different operating conditions / locations.

  - E.g., How well does face recognition system X perform at IAD versus at LAX?



Science and Technology

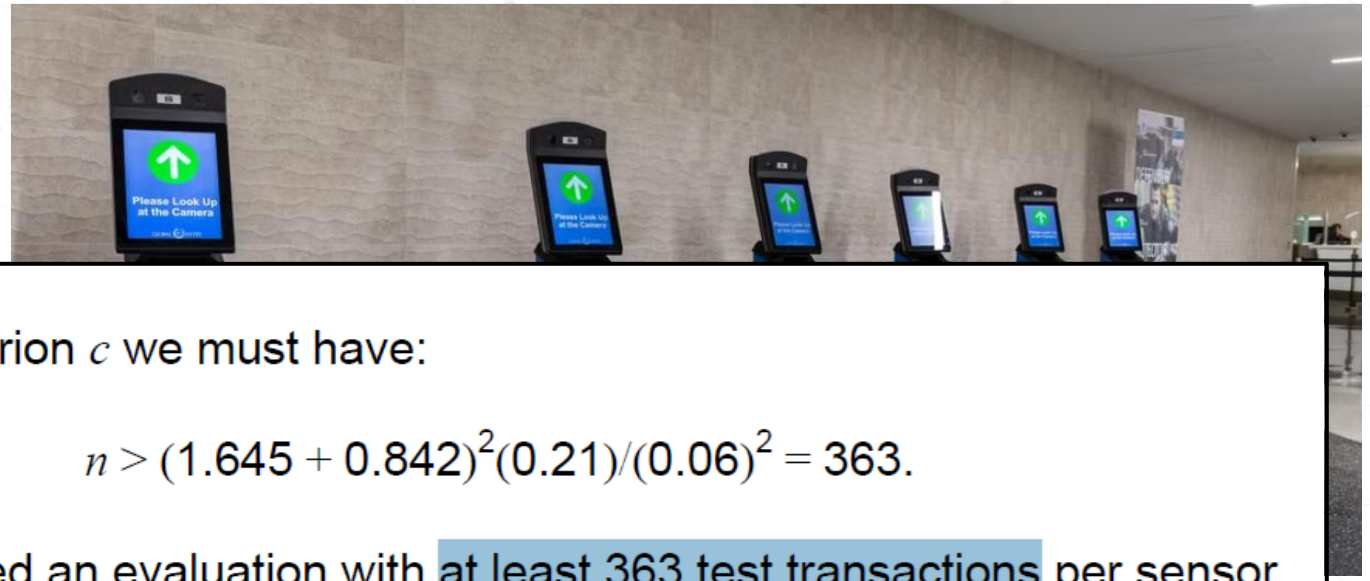
Traditional Approaches to Operational Evaluations

- Identify the system to test

Science and Technology

Traditional Approaches to Operational Evaluations

- Identify the system to test
- Identify the test size
  - Depending on the effect size of interest, and the expected error rates this can be large.



criterion c we must have:

$$n > (1.645 + 0.842)^2(0.21)/(0.06)^2 = 363.$$

need an evaluation with at least 363 test transactions per sensor.

mber of test transaction for comparison of proportions (for a two-tai
is

Traditional Approaches to Operational Evaluations

- Identify the system to test

- Identify the test size
  - Depending on the effect size of interest, and the expected error rates this can be large.

- Travel to site with known test crew
- Perform:
  - X mated transactions
    - Straight-forward
    - Expected error rates are 1-5% (low N)
  - Y non-mated transactions
    - Requires specialized personnel
    - Expected error rates are 1/1000 (high N)

- Report on:
  - Throughput
  - Enrollment rates (optional)
  - Recognition rates

criterion c we must have:

$$n > (1.645 + 0.842)^2(0.21)/(0.06)^2 = 363.$$

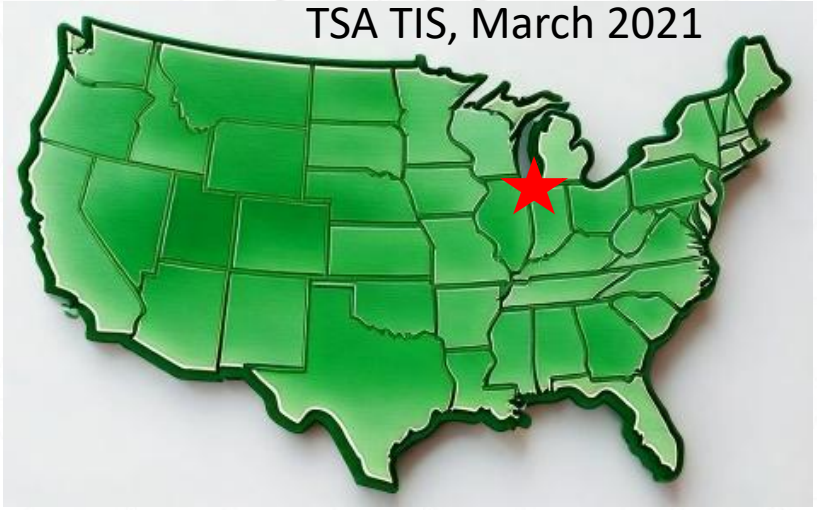
need an evaluation with at least 363 test transactions per sensor.

s (for a two-tai

Science and Technology

The Need for New Approaches

- This may work for single site deployments

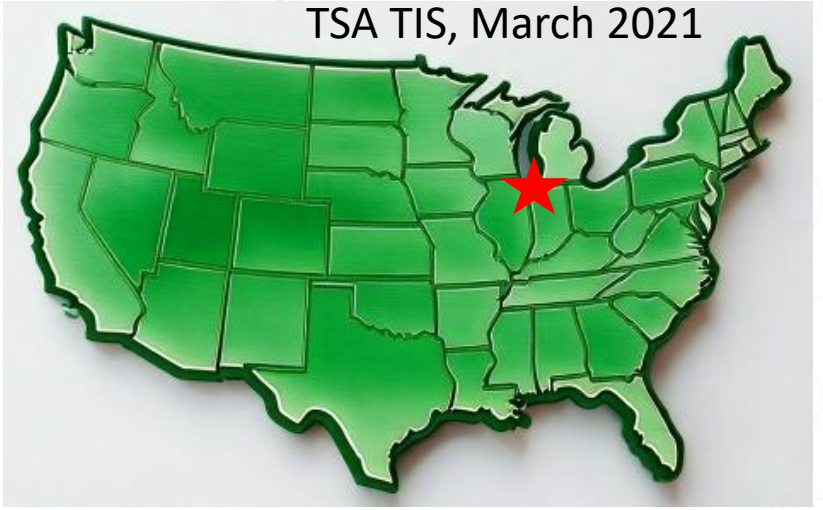


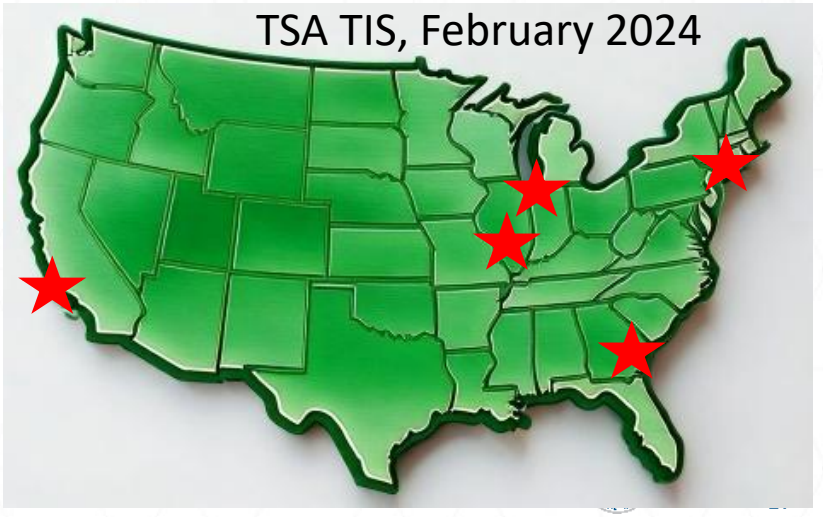
Science and Technology

The Need for New Approaches

- This may work for single site deployments

- But is more challenging to scale to multisite, multi-environment deployments





The Need for New Approaches

- At the same time increasing numbers and size of FR deployments has made operational testing more challenging.
- ... it is becoming more called for:
  - EO 13960 – Promoting the Use of Trustworthy AI in the Federal Government:
    - Principals for Use of AI in Government: Agencies shall ensure their AI applications are regularly tested and mechanisms should be maintained to deactivate existing applications of AI that are inconsistent with their intended use.
  - U.S. Commission on Civil Rights (USCCR) – The Civil Rights Implications of the Federal Use of Facial Recognition Technology:
    - "Congress should direct and empower NIST to develop an operational testing protocol that agencies can use to assess how effective, equitable, and accurate their FRT systems are when actually deployed"

The Need for New Approaches

- At the same time increasing numbers and size of FR deployments has made operational testing more challenging.
- ... it is becoming more called for:
  - U.S. Dept. of Homeland Security, Dept. of Justice, White House Office of Science and Technology Policy (OSTP) – Biometric Technology Report:
    - Section 7 – Best Practices and Guidelines – ".. agencies should specify and disclose in public documentation any independent assessments and benchmarks of biometric systems, which should be measured using standardized methodologies in as close to an operational context as possible"
  - H.R. 4609 – The National Institute of Standards and Technology for the Future Act – "Standards and guidelines required shall include: .. performance standards and guidelines for high-risk biometric identification systems, including facial recognition systems, accounting for various use cases, type of biometric identification systems, and relevant operational conditions"

Scaling Operational Testing

- Scaling operational testing requires removing the requirement for the use of a “test crew”
- Reframing from Operational Testing to Testing of an Operational System

- Use the real operational populations and perform:
  - Offline computation of performance using samples
  - Offline computation of performance using logs

- What about imposter transactions?
  - Offline reuse of genuine transaction data
  - Scenario testing

- What about failure to acquire?
  - In the field observations by a tester
  - Scenario testing

A Recent Example

- Background on FR/FC:
  - In 2023, DHS issued management directive 026-11 on the Use of Face Recognition (FR) and Face Capture (FC) Technologies
  - The directive tasked DHS S&T to:
    - Develop accuracy and performance metrics and T&E procedures
    - Provide guidance, technical expertise, and oversight for testing and evaluation of DHS uses of FR/FC technologies
  - In 2024, over 32 different FR/FC use cases were identified that may be subject to this directive
  - Many of these are deployed to:
    - Tens - hundreds of locations
    - In U.S and abroad.
    - In challenging environments (southern border, remote airfields / ports of entry, etc.)

A Recent Example

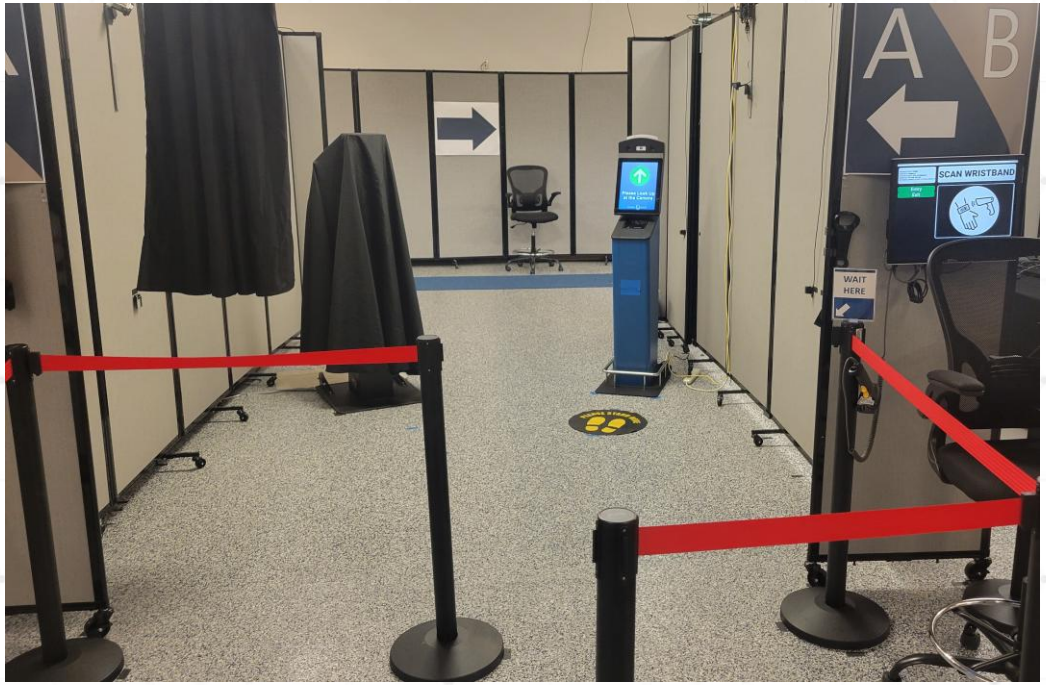
- One of those systems was GE Portal
  - Deployed at over 75 airports, worldwide
  - Thousands of portals
  - Millions of users

- Needed to evaluate:
  - Failure-to-acquire ← Requires observing user interaction
  - False negative identification rate ← Operational data / Operational logs / Scenario testing
  - False positive identification rate ← Should be small, requires large number of samples
  - Demographic differentials ← Requires demographic labels, not present in operational data

A Hybrid Scenario + Operational Evaluation

- Scenario test, September 2024 at the Maryland Test Facility.
  - 634 test volunteers used a GE Portal
  - Measured efficiency, failure-to-acquire, and satisfaction
  - Demographic differentials (full meta data)

- Operational test, November – December 2024 in DHS S&T's Cloud Based Analytic Environment (CBAE)
  - Received a GE Portal gallery of 156,332
  - 12,430 GE Portal probes from five ports of entry
  - Over 1.9 billion individual comparisons
  - Measured false positive identification rates
  - Demographic differentials (limited meta data)



Science and Technology

A Hybrid Scenario + Operational Evaluation

- Produced a standardized **“report card”** for the GE Portal system
  - Describe the evaluation
  - Describe the system
  - Describe the results
  - Describe the data
  - Provide recommendations
- Allows for repeatable communication of testing activities
- Performed a similar activity for other operational systems

FACE VERIFICATION SYSTEM TEST CARD

DHS Component: CBP **System Name: GE Portal**

Evaluation

Date	9/9/2024 – 12/10/2024
Type	Scenario + Operational Data Test
Scope	☑ Face Capture ☑ Face Recognition

Description

DHS S&T managed an evaluation of the GE Portal…

System

Characteristic	Version
Configuration ID	X
Face Capture System	X
Face Recognition System	X

Use Case

The GE Portal uses face capture and face recognition technology to…

Image Source

The probe imagery is captured via the GE Process

Role of DHS Staff

GE Portals are unstaffed. CBP officers review match results before entry is granted

Performance

Metric	Pass/Fail	Objective	Measured
Efficiency	-	-	X
Failure to acquire Rate (probe)	-	-	X
Failure to acquire (reference)	-	-	X
False negative rate (FNIR)	✓	X	X
False positive Rate	✓	X	X

Performance Notes

Dataset

Scenario Test Dataset	Number
Unique Subjects / Acquisition Attempts	634 / 634
Mated Identifications / Gallery Size	632 / 5349
Operational Test Dataset	
Non-mated Identifications / Gallery Size	12,437 / 156,317

Statistical Considerations

Demographic Groups

Recommendation

Developments in International Standards

- We think this is a good model to increase the **efficiency** of operational testing
- Might be the only way to meet the increased need for testing in the context of increasing system deployments
- Currently, the international standard for operational testing is ISO/IEC 19795-6

Current Copy of 19795-6

- **19795-6:** Biometric performance testing and reporting, Part 6: Testing methodologies for operational evaluation

- Edition 1 (and only) published in **2012**

- Since then:
  - Notable increase in the pace of **operational deployments**
  - And notable developments in the **understanding** of biometric systems:
    - 19795-1: Principals (2021)
    - 2382-37: Vocabulary (2022)
    - 30107-X: PAD (2023)
    - 19795-10: Demographic differentials (2024)
    - 29794-X: Sample Quality (pending IS, 2025)

Updates to 19795-6

- **Topic 1:** Reframe the standard from “Operational testing” to “Testing of operational systems”

- **Topic 2:** Currently the standard envisions many of these concepts but leaves out a framework for combining them (as we did in FCFR)
  - Adding this could notably reduce the cost and time associated with conducting a standards compliant operational test

- **Topic 3:** Formalize the kinds of entities that might perform a test
  - First party, second party, third party
  - Increasingly crowded space with public and private entities
  - Who conducts the test has implications on cost, timeline, internal/external validity

Updates to 19795-6

- **Topic 4:** Common approaches to make systems testable
  - SDKs versus APIs versus logging
  - Earlier adoption of a testing harness is a key driver of testability later on

- **Topic 5:** Describe the role of the human
  - There are systems on the market today that have a 100% false positive rate (by design)
  - Human review of biometric outcomes is seen as a fail safe
  - It could be in some cases (orthogonal information) but likely isn't in all cases
  - In order to understand the performance of the full system, a tester needs to account for human performance.

Updated to 19795-6

- **Topic 7:** Address the testing of “face adjacent” technologies, such as quality and PAD
  - Operational systems combine many different software solutions
  - 19795-1 formalizes the concept of a False Rejection Rate that accounts for classic failure to acquire
  - Need to extend these out to include failures to proceed with the process because of:
    - Rejection of genuine users from a PAD subsystem (BPCER)
    - Rejections of good quality users from a quality subsystem (error discard)
    - Rejections of genuine users from a biometric injection attack detection subsystem (??)
    - Others?
- **Topic 8-?:** Call for contributions!

Questions & Answers

- Contact information:
  - peoplescreening@hq.dhs.gov
  - arun.vemury@dhs.gov
  - jhoward@idslabs.org

- Visit our websites for additional information.
  - To see additional work DHS S&T supports, visit www.dhs.gov/science-and-technology.
  - For information about this and other DHS S&T technology evaluations, visit https://mdtf.org.

Science & Technology