Developments in ISO 19795-10: Measuring Demographic Performance Across Demographic Groups

European Association for Biometrics, Workshop on Demographic Fairness in Biometric Systems, 3/30/2021

Homeland Security
Science and Technology

Jacob Hasselgren, John Howard

The Maryland Test Facility

Arun Vemury

Biometric and identity Technology Center, Director

Disclaimer

- Support for this effort is funded by the U.S. Department of Homeland Security, Science and Technology Directorate on contract number 70RSAT18CB0000034

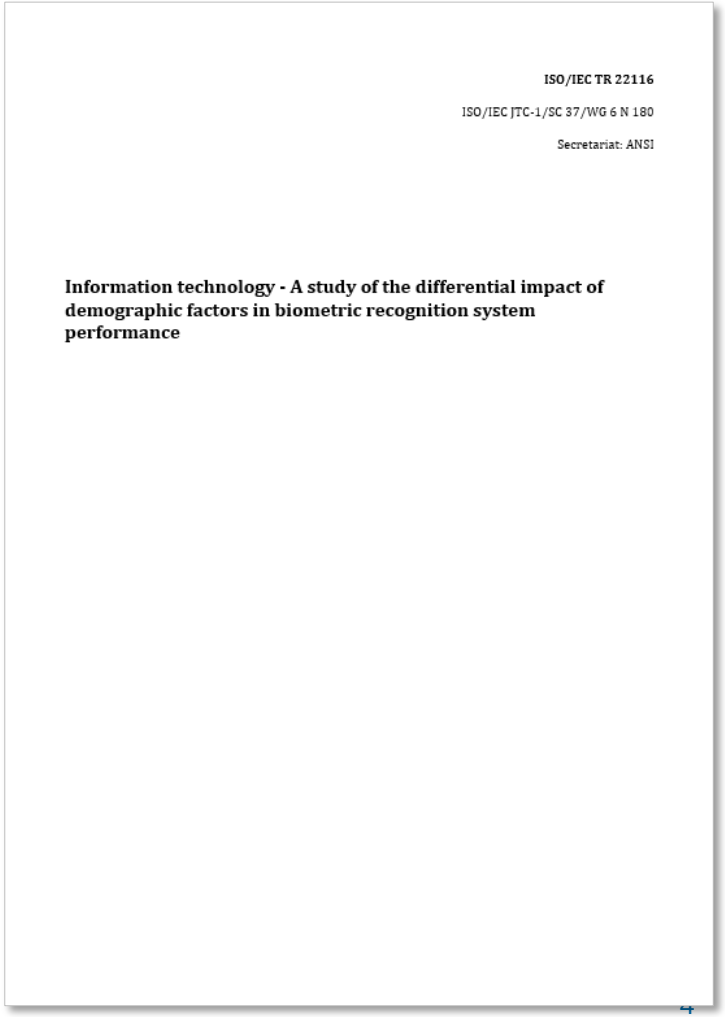
- Any opinions provided today are those of the authors and do not represent those of the Department of Homeland Security, the U.S. Government, or their employers

Standards History

- **ISO 19795-1 (2006) - Information technology — Biometric performance testing and reporting — Part 1: Principles and framework**
  - Establishes general principles for testing the performance of biometric systems in terms of error rates and throughput rates
  - Specifies performance metrics, requirements for recording of test data, and requirements on test protocols
  - Provides definitions for performance metrics, such as false-negative and false-positive identification rates
  - Currently under a five year review. Expected to be republished in 2021
- **ISO 2382-37 (2012) - Information technology — Vocabulary — Part 37: Biometrics**
  - Provides systematic descriptions of concepts in the field of biometrics pertaining to recognition of human beings
  - A normative reference for 19795-1
  - Most recent version republished in 2017

Standards History – ISO/IEC Technical Report 22116

- ISO IEC Joint Technical Committee 1 (Information Technology)
  - Subcommittee 37 (Biometrics)
    - Working Group 6 (Cross Jurisdictional and Societal Aspects of Biometrics)
- Scope
  - Terms and definitions
  - Where performance variation can exist in a biometric system
  - Literature review
- Approved for publication in January 2021


ISO/IEC TR 22116

ISO/IEC JTC-1/SC 37/WG 6 N 180

Secretariat: ANSI

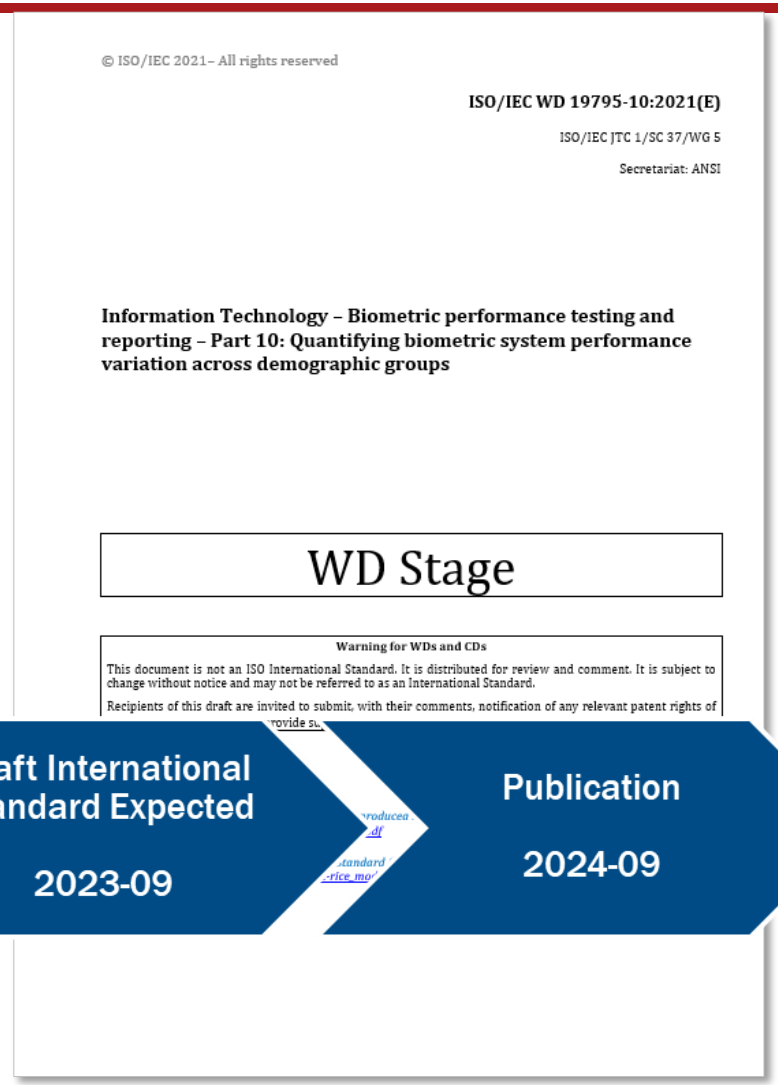
Information technology - A study of the differential impact of demographic factors in biometric recognition system performance

Current Need to Standardize How we Measure and Talk about Demographic Fairness

- Growing numbers of deployments (law enforcement, border control, private)
- Increased public awareness and concerns
- Concern amongst policy-makers:
  - USS.3284 – Ethical Use of Facial Recognition Act
  - USS.4084 - Facial Recognition and Biometric Technology Moratorium Act of 2020
  - Australian Identity Matching Services Bill 2019
  - European Commission Ethics Guidelines for Trustworthy AI
- Inconsistency amongst researchers:
  - Bridges v. South Wales Police
  - “Bias” versus “Differential”
  - Sources of differentials (training, historical, process, etc.) and how we test for them

ISO/IEC WD 19795-10

- Quantifying biometric system performance across demographic groups
- New work item, approved in 2020
- First draft summer 2021
- Anticipated publication in 2023 - 2024



Homeland Security
Science and Technology

19795-10 Current Challenges

- **Scope**
- Definitions and nomenclature
- Categorical versus phenotypical measures and studies
- Statistical versus practical equivalence & uncertainty estimates
- How, where, and when to test
- What to report when you do test

Scope

- From the approved new work item proposal, this standard will:

  … establish requirements for estimating and reporting of performance variations observed when cohorts belonging to different demographic groups engage with biometric enrollment and recognition systems

- Within Scope:
  - **guidance** on establishing demographic group membership
  - **guidance** on using phenotypic measures
  - establish **terms and definitions** to be used when reporting performance variation across demographic groups
  - **requirements** on reporting of tests
  - **requirements** for stating statistical uncertainty estimates

Homeland Security
Science and Technology

Scope

- **Demographics** – statistical characteristics of human populations (Merriam-Webster)
  - Populations, plural - i.e., groups of people
  - Can be based on:
    - Biological Characteristics: Sex, age, weight, height, skin tone, etc.
    - Geography: Birthplace, country of residence, city of residence, neighborhood, etc.
    - Social Constructs: Race, ethnicity, gender, marital status, income, education, employment, shopping habits, etc.
  - Very broad

- Important to determine which groups to address explicitly:
  - Groups important to biometric performance?
  - Groups with legal protections?

Scope

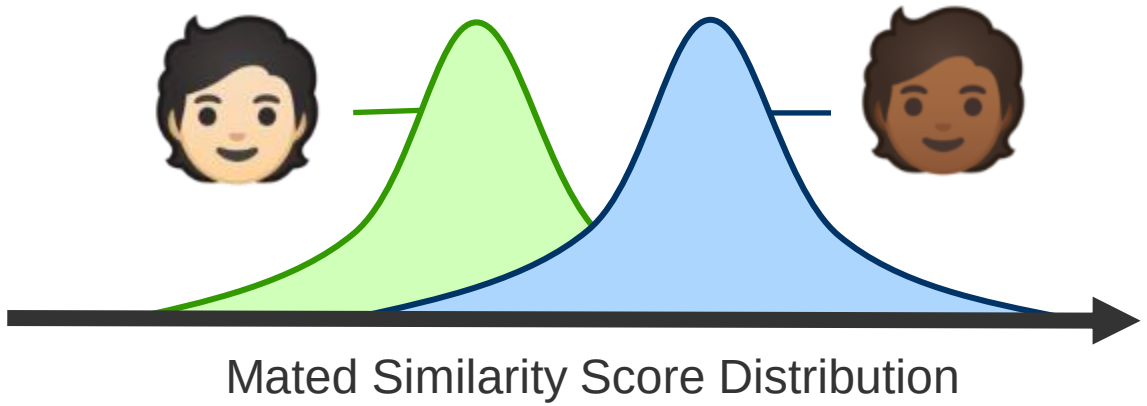
- Excluded from scope (not explicit in the new work item):
  - Biometric “non-recognition”, i.e., analysis
    - Biometric Sample Quality
    - Emotion, gender, or age estimation
  - Demographic groupings based on traits, not states
    - Makeup - makeup is not a biological demographic
    - Mask wearing - masks are not a biological demographic
  - Medical conditions
    - Eye surgery, cataracts, vision correction
    - Stroke, cleft lip, Apert’s syndrome
    - Missing digits

19795-10 Current Challenges

- Scope
- **Definitions and nomenclature**
- Categorical versus phenotypical measures and studies
- Statistical versus practical equivalence & uncertainty estimates
- How, where, and when to test
- What to report when you do test

Definitions and Nomenclature

Differential Performance:



Differential Outcomes:

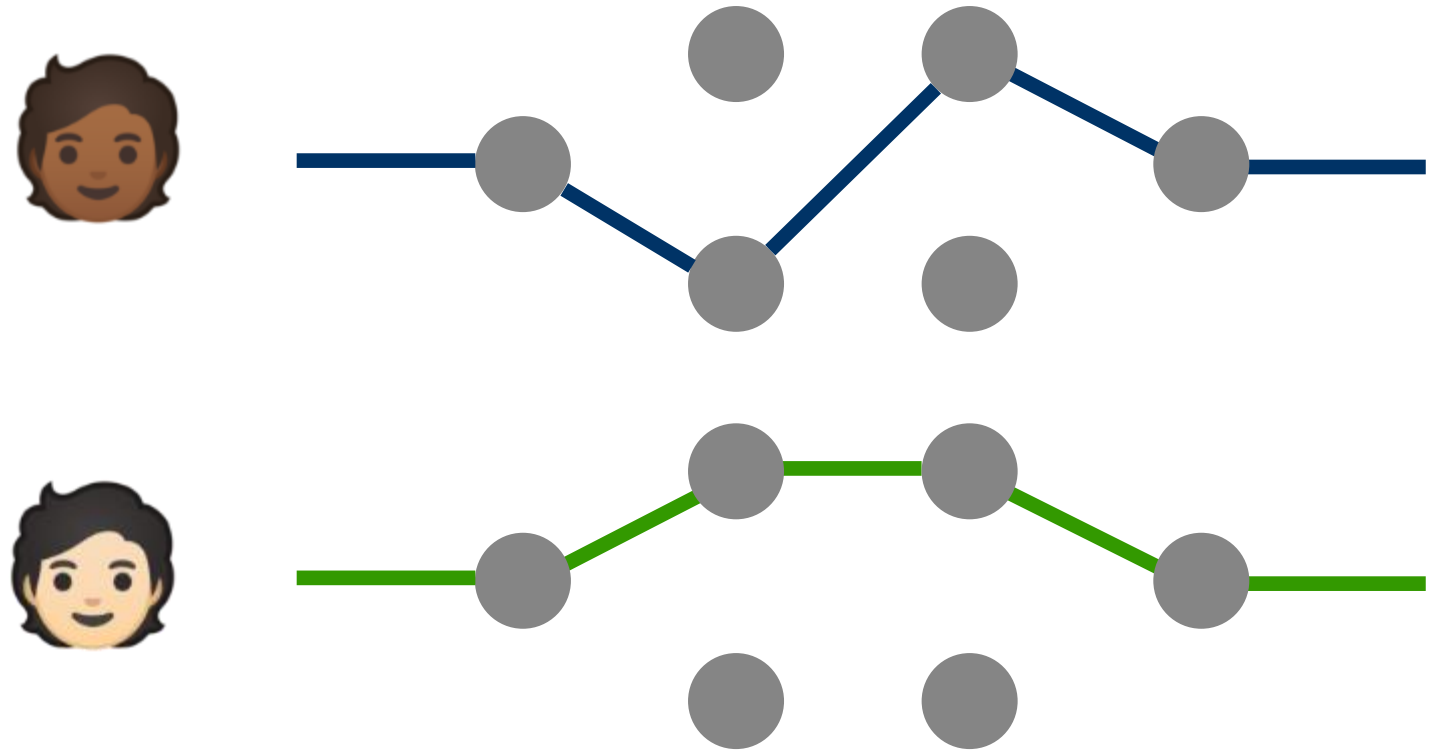
FNMR: 5%

FNMR: 3%

Homeland Security
Science and Technology

Definitions and Nomenclature

Differential Treatment:



Definitions and Nomenclature

- **False negative differentials** - tendency for **mated biometric samples** from subjects in one demographic group not to match relative to another demographic group

- **False positive differentials** - tendency for **non-mated biometric samples** from one demographic group to falsely match relative to another demographic group, or a tendency for this effect to occur across demographic groups

- Each differential can be described separately

- Standard may include guidance on identifying the differential(s) of concern across use-cases

Definitions and Nomenclature

- **Summative Measures** - measures that combine multiple error rates or performance metrics
  - Differentials may be observed in summative measures (e.g., Accuracy, DCF, HTER)

- **Fairness Measures** - summative performance measures that have been proposed as fairness metrics that combine differentials (FDR, NIST Inequity)

- Standard may leave choice of metrics open

$$\frac{TP + TN}{TP + TN + FP + FN} \qquad \beta * FPR + \alpha * FNR \qquad \frac{FPR + FNR}{2}$$

$$A(\tau) = \max(|\mathrm{FMR}^{d_i}(\tau) - \mathrm{FMR}^{d_j}(\tau)|) \quad \forall d_i, d_j \in \mathcal{D}$$

$$B(\tau) = \max(|\mathrm{FNMR}^{d_i}(\tau) - \mathrm{FNMR}^{d_j}(\tau)|) \quad \forall d_i, d_j \in \mathcal{D}$$

$$FDR(\tau) = 1 - (\alpha A(\tau) + (1 - \alpha)B(\tau))$$

$$\mathrm{A}(\tau) = \frac{\max_{d_i} \mathrm{FMR}^{d_i}(\tau)}{\min_{d_j} \mathrm{FMR}^{d_j}(\tau)} \quad \forall d_i, d_j \in \mathcal{D}$$

$$\mathrm{B}(\tau) = \frac{\max_{d_i} \mathrm{FNMR}^{d_i}(\tau)}{\min_{d_j} \mathrm{FNMR}^{d_j}(\tau)} \quad \forall d_i, d_j \in \mathcal{D}$$

$$\mathrm{INEQUITY} = A(\tau)^{\alpha} B(\tau)^{\beta}$$

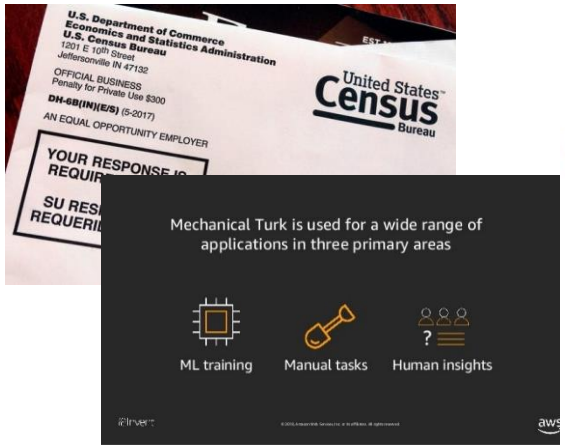
Homeland Security
Science and Technology

19795-10 Current Challenges

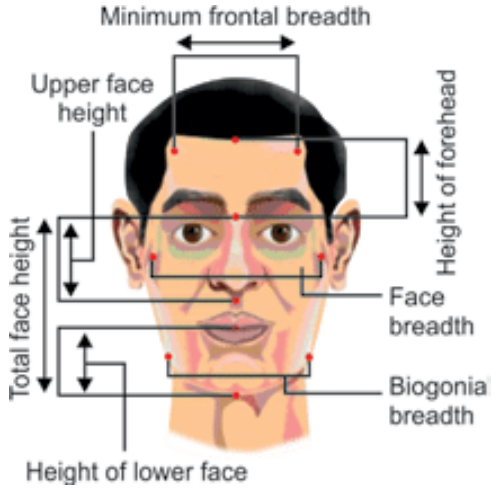
- Scope
- Definitions and nomenclature
- **Categorical versus phenotypical measures and studies**
- Statistical versus practical equivalence & uncertainty estimates
- How, where, and when to test
- What to report when you do test

Categorical versus Phenotypical

- **Categorical**
  - Subjective categories
  - Self reported or assigned
  - Discrete



- **Phenotypical**
  - Observable characteristics
  - Measurable
  - Can be continuous



Categorical versus Phenotypical Measures

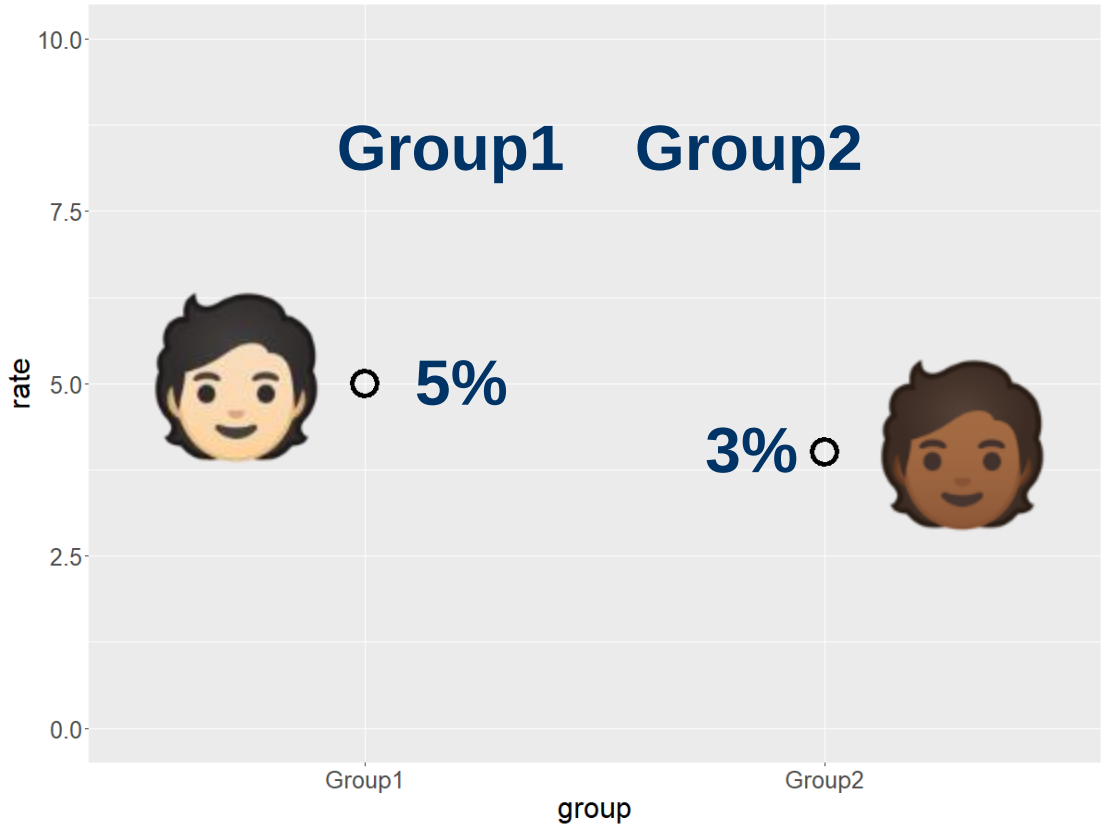
Categorical	Phenotypes
Cons: • Rely on (potentially) socially defined or locale specific definitions • Can be poor explainers of the variability in a dataset. “Black or Asian” describes people from diverse racial backgrounds.	Cons: • Can be difficult to collect without access to the subject (Fitzpatrick, skin tone in general) • Often attempted from the actual biometric sample, which introduces sampling error to both measurement and outcome
Pros: • In some locales, categorical variables can be legally protected classes • May be required to show fairness across categorical category in evaluations	Pros: • Don’t rely on social constructs • Possibly a better explainer of the outcome variable • Often easier to arrive at engineering solutions given phenotypic explanations

19795-10 Current Challenges

- Scope
- Definitions and nomenclature
- Categorical versus phenotypical measures and studies
- **Statistical versus practical equivalence & uncertainty estimates**
- How, where, and when to test
- What to report when you do test

Statistical versus Practical Equivalence

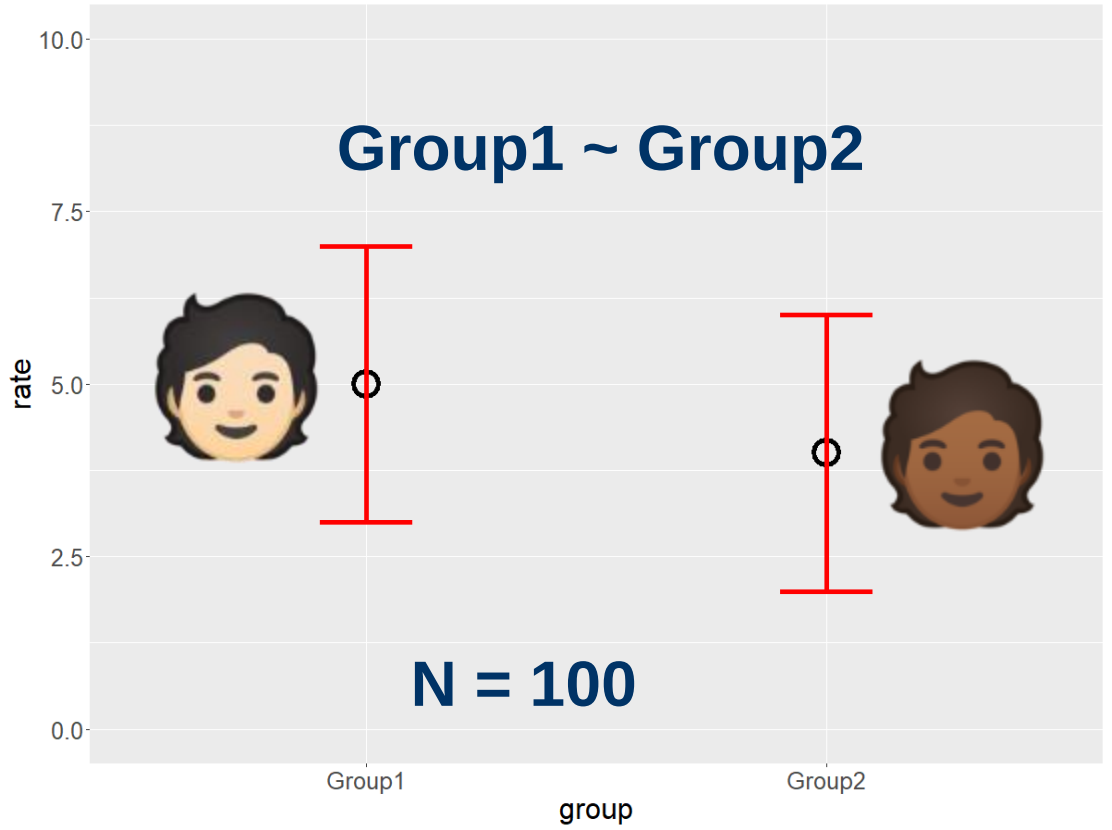
- Standard may include requirements for reporting of statistical uncertainty in differentials
- What do we mean when we say two rates are equal?
- Precisely equal? 95.21% != 95.22%
- Statistically equal?
  - Sampling a population introduces error
  - That error is based, in part, on how much of the population you sampled



Homeland Security
Science and Technology

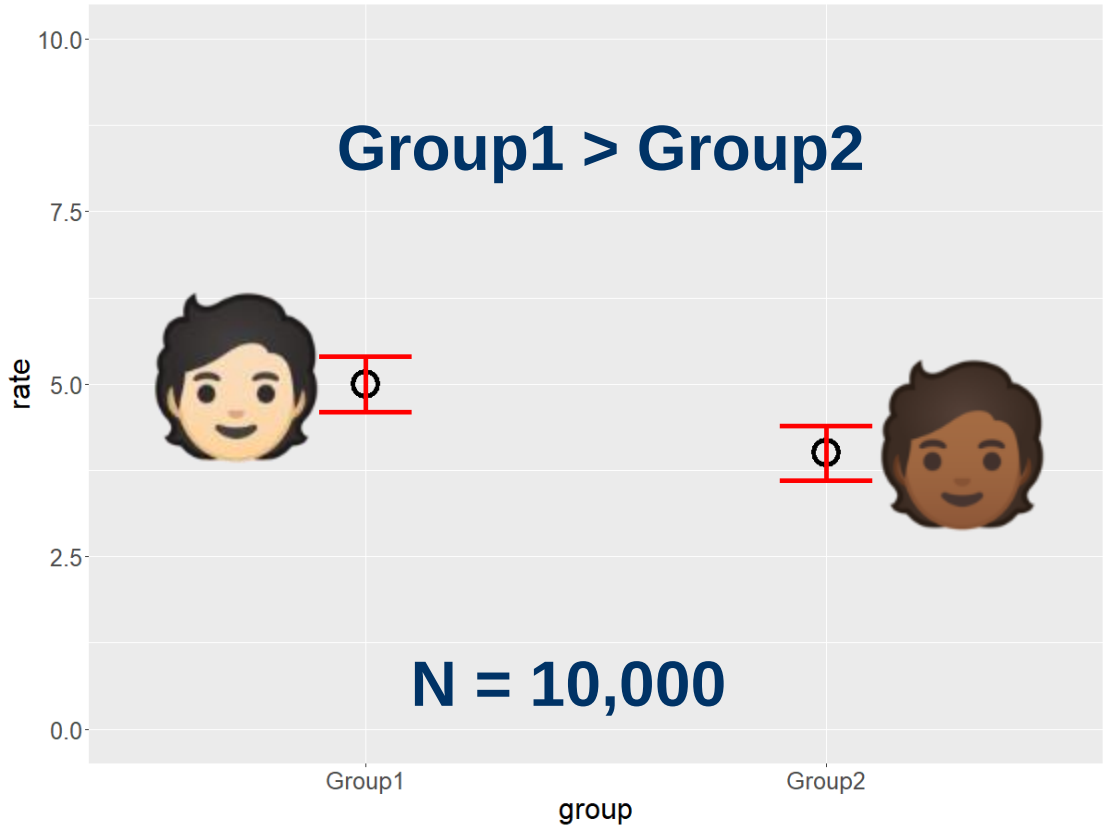
Statistical versus Practical Equivalence

- Standard may include requirements for reporting of statistical uncertainty in differentials
- What do we mean when we say two rates are equal?
- Precisely equal? 95.21% != 95.22%
- Statistically equal?
  - Sampling a population introduces error
  - That error is based, in part, on how much of the population you sampled
  - As number of observations ⬆ confidence intervals ⬇



Statistical versus Practical Equivalence

- Standard may include requirements for reporting of statistical uncertainty in differentials
- What do we mean when we say two rates are equal?
- Precisely equal? 95.21% != 95.22%
- Statistically equal?
  - Sampling a large population introduces error
  - That error is based, in part, on how much of the population you sampled
  - As number of observations ⬆ confidence intervals ⬇
  - This has a downside – at some level N there is always a **statistical** difference. *Minimum detectable effect*.



Homeland Security
Science and Technology

Statistical versus Practical Equivalence

- NIST FRVT Part 3 numbers of subjects in each demographic category
  - 3 million imposter comparisons within each group
- At this population size (N), it is likely that even small differences in error rates between groups will be statistically significant
- Standard may include requirements for reporting statistical uncertainty estimates based on the sample sizes used in the evaluation

	Race Label	Sex Label	Mated Comparison Count	Impostor Comparison Count
1	A	F	10 995	3 000 000
2	A	M	139 342	3 000 001
3	B	F	263 910	3 000 007
4	B	M	1 954 864	3 000 009
5	I	F	26 699	3 000 000
6	I	M	268 364	3 000 006
7	W	F	362 816	3 000 012
8	W	M	1 033 237	3 000 017
9	Total		4061227	108000690

Statistical versus Practical Equivalence

- Lets pretend a false match rate of 10 in 100,000 tries (1e-4) for black males

- If a false match happens 12 in 100,000 times for white males, is that equal?

```
> data.frame(ovp2) %>% mutate( success.rate = success / (success + false.match))
  success false.match success.rate
1     300     2999709 0.0000999997
2     360     2999657 0.0001199993
>
> prop.test(ovp2)

        2-sample test for equality of proportions with continuity correction

data:  ovp2
X-squared = 5.2747, df = 1, p-value = 0.02164
alternative hypothesis: two.sided
95 percent confidence interval:
 -3.711608e-05 -2.883165e-06
sample estimates:
      prop 1       prop 2
0.0000999997 0.0001199993
```

- P < 0.05, yes, a statistical difference exists

- Caution:
  - Minimum effect of interest >> Minimum detectable effect

- Standard may include guidance on interpretation of statistical differences

Homeland Security
Science and Technology

Statistical versus Practical Equivalence

- Observable differences are based on 1) differences in error rates and 2) volume of biometric operations

- Very few existing definitions of what that *allowable* difference in observed error **rate** or observed **errors** can be
  - Based on a proportion? (US Equal Employment Opportunity Commission)
  - Based on a finite percentage? (Minimum effect of interest)
  - Others?

FNMR: 5%

FNMR: 3%

Homeland Security
Science and Technology

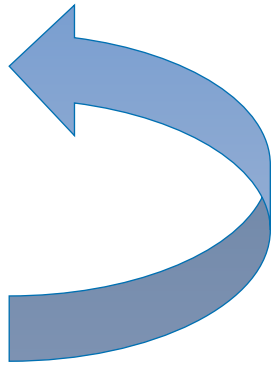
19795-10 Current Challenges

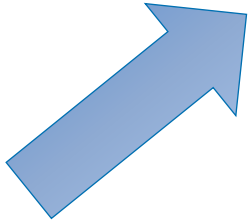
- Scope
- Definitions and nomenclature
- Categorical versus phenotypical measures and studies
- Statistical versus practical equivalence & uncertainty estimates
- **How, where, and when to test**
- What to report when you do test

How, Where, and When to Test

Operational Testing

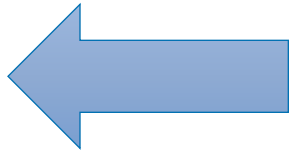
ISO 19795-6: Biometric performance testing and reporting – Part 6: Testing methodologies for operations evaluation





Scenario Testing

ISO 19795-2: Biometric performance testing and reporting — Part 2: Testing methodologies for technology **and scenario** evaluation



Technology Testing

ISO 19795-2: Biometric performance testing and reporting — Part 2: Testing methodologies for technology evaluation

Homeland Security
Science and Technology

How, Where, and When to Test

- Technology test:
  - Good for motivating progress from industry
  - Tracking progress (same dataset over time)
  - Very large N allows very good capability to distinguish technologies
- Scenario test:
  - Good for finding issues in whole systems (poor camera, poor camera height, poor signage)
  - Good for in-depth demographic studies
  - Small N allows for less differentiation
- Operational test
  - Neither technology or scenario tests can be fully predictive of operational performance
  - Things change: database, environment, population, masks
  - Collecting ground-truth information about “subjects” in an operational test can be a challenge

19795-10 Current Challenges

- Scope
- Definitions and nomenclature
- Categorical versus phenotypical measures and studies
- Statistical versus practical equivalence & uncertainty estimates
- How, where, and when to test
- **What to report when you do test**

What to Report

- Different use cases have different “primary error(s) of concern”. Therefore, different use cases may have different reporting criteria for demographic differentials.

- Factors:
  - Kind of test (technology, scenario, and operational)
  - Kind of operation (1:1, 1:N-allow, 1:N-deny, etc.)

- Operational test of a 1:N-deny system:
  - Gallery composition
  - False positive identification rate (positives / non-gallery searches), across demographics
  - False discovery rate (false positive / positives)

- Laboratory test of a 1:N or a 1:1 system:
  - Level of specific and broad homogeneity across demographic groups of interest
  - False non-match rate across phenotypes -- skin tone

Conclusions

- ISO/IEC 19795-10 will standardize how we quantify biometric system performance across demographic groups
- This will help address questions regarding “demographic fairness” in biometric system performance
- Development is underway. Now soliciting contributions
- Major areas of development:
  - Scope
  - Definitions and nomenclature
  - Categorical versus phenotypical measures and studies
  - Statistical versus practical equivalence
  - How Where, and When to Test
  - What to report

Questions & Next Steps

- jacob@mdtf.org
- john@mdtf.org
- jerry@mdtf.org

- Find out more at https://mdtf.org/

- **arun.vemury@hq.dhs.gov**

Homeland Security
Science and Technology