

# Biometric Testing: It's Not as Easy as You Think

Valorie S. Valencia, PhD  
Chief Executive Officer



*... providing scientific consulting, evaluation, and  
training services in all areas of authentication*

# Experts in the field do not agree on high-level concepts and terminology (1/3)

## Example 1: Different ways to categorize biometric applications

- Taxonomy 1. Questions Asked
  - Is this person in my database? Verification
  - Who is this person? Identification
- Taxonomy 2. User Claim
  - Positive identification – user makes overt or implied claim to be known to the system (access control, benefit claims)
  - Negative identification – user makes direct or implied claim to NOT be known to the system (benefit enrollment, overt watchlist)
  - Surveillance – no claim is made (covert watch list)
- Taxonomy 3. System Knowledge
  - Positive identification – user makes overt or implied claim to be known to the system (access control, benefit claims)
  - Negative identification – user makes direct or obscure claim to NOT be known to the system (benefit enrollment, surveillance, watch list)
- Taxonomy N-1. Generalized Watch List
  - Every application is a special case of watch list

Experts in the field do not agree on high-level concepts and terminology (2/3)

Example 2: International Biometric Standards Committee (SC37) Harmonized Vocabulary Working Group cannot agree of definition of the term “biometrics”

- Working Version 4: [Automated] recognition of [living] persons (beings) based on observation of behavioral and biological (anatomical and physiological) characteristics (traits).

# Experts in the field do not agree on high-level concepts and terminology (3/3)

- Strong disconnect between biometric “experts”
  - Experts come from wide variety of backgrounds – computer science, physics, pattern matching, image processing, statistics, engineering, etc.
- Vendors (and everybody else) sometimes do not understand terminology
  - For example: What is a “gallery”?
  - Can we blame them if various testers use terms differently?
- Analogy: Physicists and chemists often “violently agree” – using completely different terminology
- . . . Biometric Sciences is an emerging discipline

# Testing theory not well understood

- Testing paradigms are controversial
- Statistical models are complicated by real world
- Performance prediction models are not available

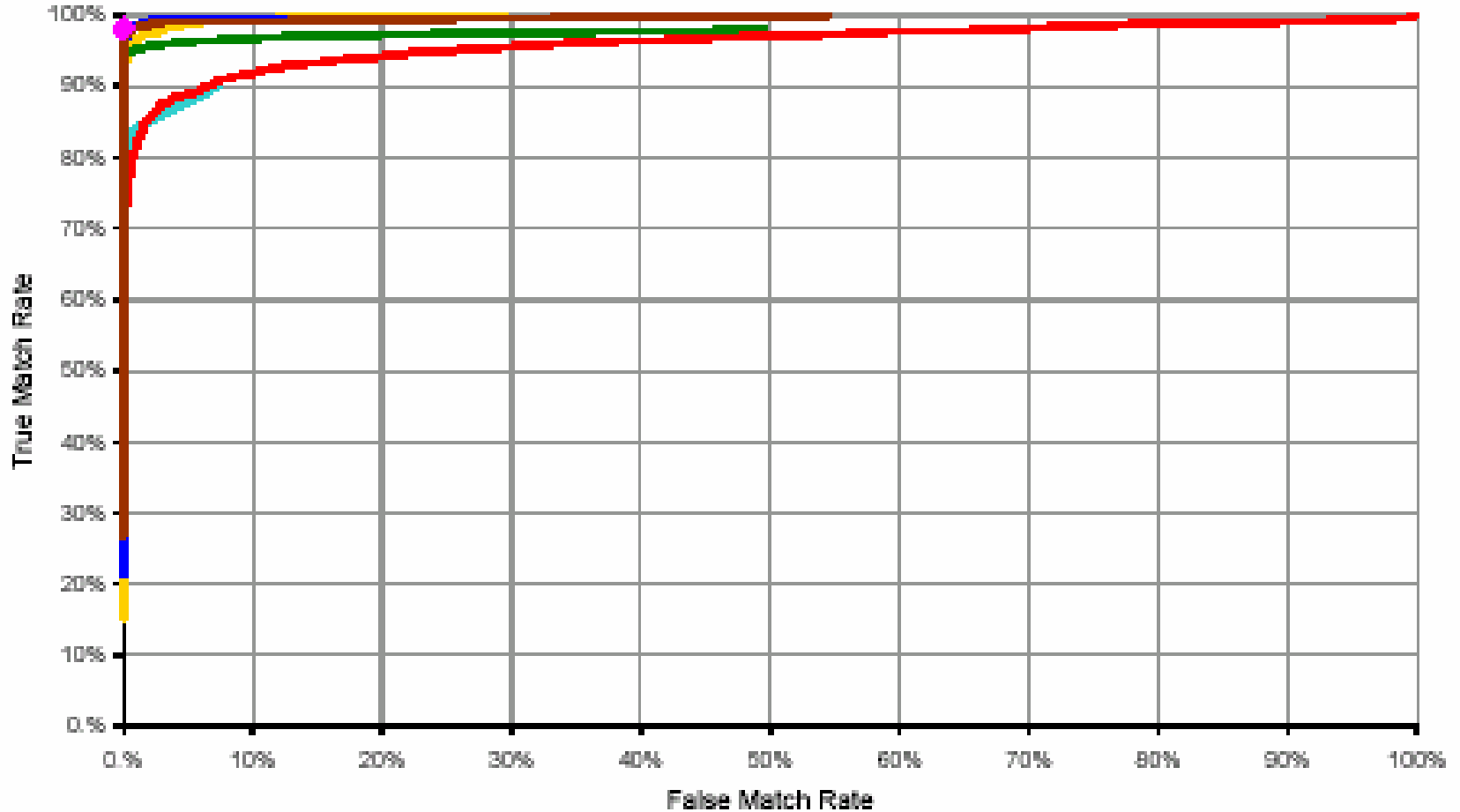
# Testing paradigms controversial

- Two Examples:
  1. Open-set versus closed-set testing
  2. Technology, Scenario, and Operational testing

# 1. Open-Set versus Closed-Set Testing

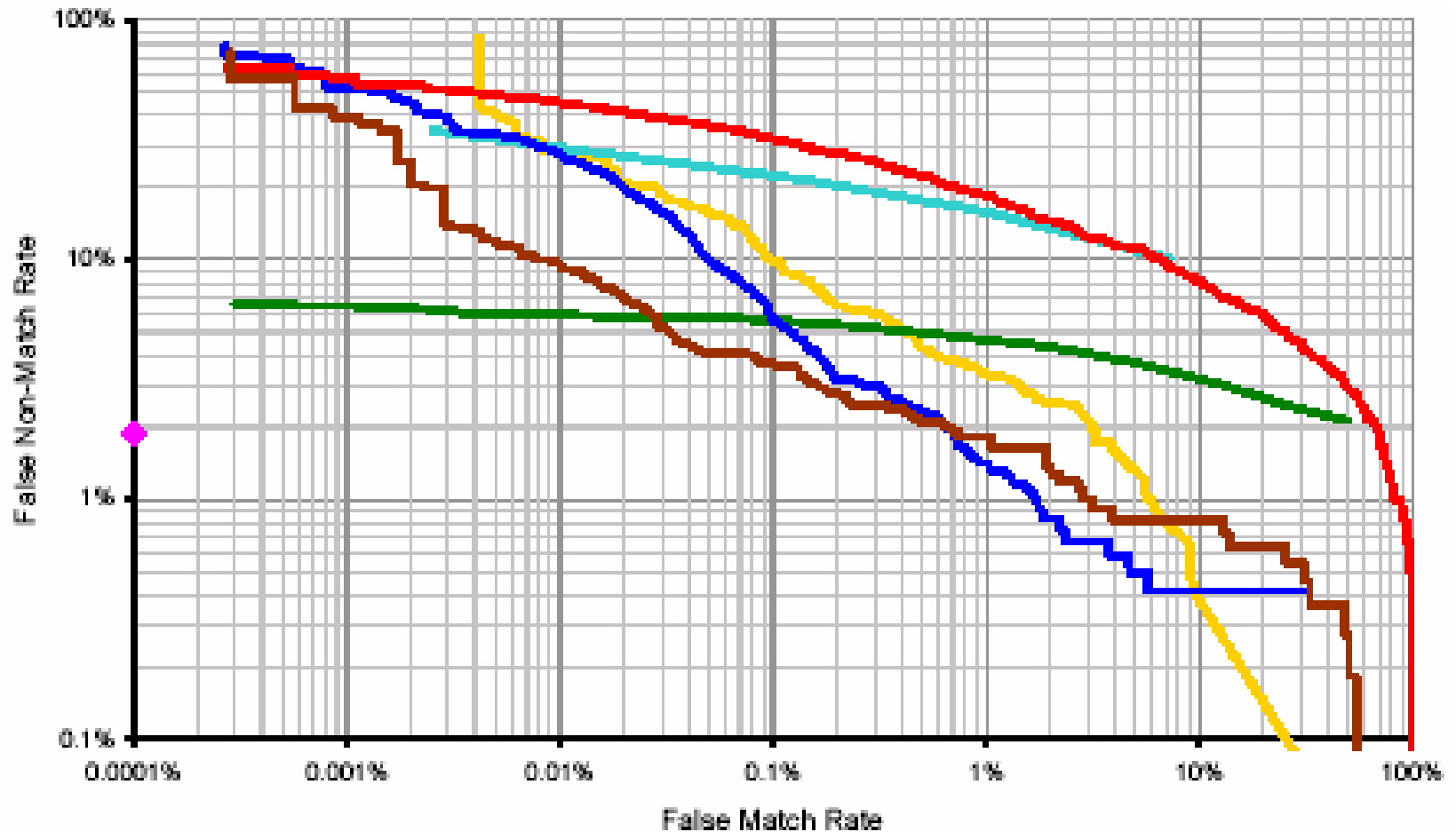
<b>Open-Set Testing</b>	<b>Closed-Set Testing</b>
Do I know you? Who are you? (Unknown impostor)	I know you. Who are you?
Threshold/rank dependent	Threshold independent
False match rates (FMR) and false non-match rates (FNMR) Receiver Operator Curve (ROC) Detection Error Trade-off Curve (DET)	Rank reporting  Cumulative Match Curve (CMC)
Consistent estimator (converges with test size)	Inconsistent estimator (does not converge with test size)
Generate histograms	Do not generate histograms

# Sample Receiver Operator Characteristic (ROC) Curves



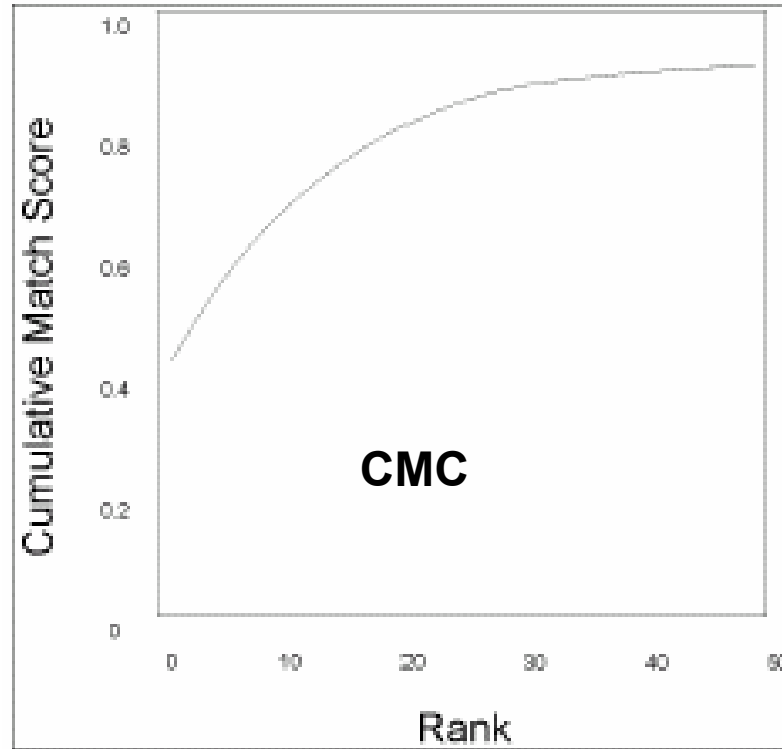
From "Best Practices," Mansfield and Wayman

# Sample Detection Error Trade-off (DET) Curves



From "Best Practices," Mansfield and Wayman

# Sample Cumulative Match Curve (CMC)



From "FRVT 2000," Blackburn, Bone, & Phillips

# Testing paradigms controversial

## 1. Open-set versus closed-set testing

- Verification → agreement between experts
  - Use **Open** set testing approach
- Identification → disagreement between experts
  - When use **Open** set versus when use **Closed** set testing approaches?

Type of Test	Technology (in vitro)	Scenario (in situ)	Operational (in vivo)
<b>Database</b>	Typically pre-collected, usually for testing multiple components	Gathered with system under test	Gathered with system under test
<b>Data Comparisons</b>	Offline	Online and/or Offline	Online (may have offline component)
<b>Object of Testing</b>	Biometric <b>component</b> (e.g., algorithm or sensor)	Biometric <b>system</b>	Biometric <b>system</b>
<b>Physical Environment</b>	Controlled or uncontrolled when biometric data recorded, Not applicable during testing	Controlled and/or recorded	Not controlled, preferably recorded
<b>User Interaction</b>	Maybe recorded when biometric data recorded, Not applicable during testing	Recorded	Recorded during enrollment, Maybe recorded during verification/identification
<b>User Behavior</b>	Controlled and/or Uncontrolled when biometric data recorded, Not applicable during testing	Controlled	Uncontrolled
<b>Results</b>	Internally consistent	Compromise between internal and external consistency	Externally consistent
<b>Repeatability of Results</b>	Repeatable (database fixed)	Quasi-repeatable (if test scenario and population controlled)	Non-repeatable
<b>Typical Results Reported</b>	Comparison of biometric components or versions of components (e.g., algorithms or sensors) Determine critical performance factors	Compare biometric systems Determine critical performance factors Predict simulated performance	Measure performance in an operational environment
<b>Constraints</b>	Appropriate test database, e.g., gathered with a universal sensor	Operational, instrumented system	Operational, instrumented system ( typically only decisions available, scores preferable)
<b>Human Test Population</b>	Recorded	Live	Live

Type of Test	Technology (in vitro)	Scenario (in situ)	Operational (in vivo)
<b>Database</b>	Typically pre-collected, usually for testing multiple components	Gathered with system under test	Gathered with system under test
<b>Data Comparisons</b>	Offline	Online and/or Offline	Online
<b>Object of Testing</b>	Biometric <b>component</b> (e.g., algorithm or sensor)	Biometric <b>system</b>	Biometric <b>system</b>
<b>Physical Environment</b>	Controlled or simulated	Controlled or simulated	Controlled or simulated
<b>User Interaction</b>	Controlled	Controlled	Recorded
<b>User Behavior</b>	Controlled	Controlled	Uncontrolled
<b>Results</b>	Consistent	Consistent between internal and external consistency	Externally consistent
<b>Repeatability</b>	Repeatable	Quasi-repeatable (if test scenario and population controlled)	Non-repeatable
<b>Typical Results</b>	Compare biometric components or versions of components (e.g., algorithms or sensors) Determine critical performance factors	Compare biometric systems Determine critical performance factors Predict simulated performance	Measure performance in an operational environment
<b>Constraints</b>	Appropriate test database, e.g., gathered with a universal sensor	Operational, instrumented system	Operational, instrumented system (typically only decisions available, scores preferable)
<b>Human Test Population</b>	Recorded	Live	Live

**BE CAREFUL WITH CAUTION**

# Testing paradigms controversial

- ## 2. Technology, Scenario, and Operational testing
- In reality, this taxonomy represents a continuum of testing as opposed to three discrete types of testing
  - Many hybrid test have been performed to date
    - Example: “Face Recognition at a Chokepoint” test by Mike Bone and Duane Blackburn at NIST
      - Data acquired from live humans in a controlled environment (scenario testing)
      - Data analyzed off line using technology testing infrastructure
      - Technology test or scenario test? Yes.

# Statistical models complicated by real world (1/6)

- Straightforward statistical approach assumes *independent, identically distributed* (iid) data samples:
  1. *Independent* → no correlation errors
  2. *Identically distributed* → no variation in statistics for each individual, no “Doddington’s Zoo”

But these two assumptions are typically not valid!

- Correct sample size and confidence intervals determined only after data acquisition and analysis!

# Statistical models complicated by real world (2/6)

## 1. Correlation errors

- Data samples not necessarily independent
- Error rates don't simply multiply
  - for example,  
odds of two errors in a row  $\neq$  (odds of one error)<sup>2</sup>
- Don't yet fully understand the influence of non-independent non-identically distributed samples
- Active area of research

# Statistical models complicated by real world (3/6)

## 2. Doddington's Zoo

- The Zoo:
  - Sheep – easily accepted by the system (common people)
  - Goats – exceptionally unsuccessful at being accepted (chronically high false rejecters, typically because the biometric data pattern is outside the range recognized by the system)
  - Lambs – exceptionally vulnerable to impersonation (good false matchees)
  - Wolves – exceptionally successful at impersonation (good false matchers)
- The existence of goats/sheep or lambs/sheep means that 3 errors in 100 trials does not equate to 100 people each with a 3% error rate.
  - Can get 3 errors in 100 trials with, for example
    - 97 people with 0% error and 6 people with 50% error
    - 50 people with 0% error and 50 people with 6% error
    - etc.
  - Each case leads to a different large-scale performance estimate

# Statistical models complicated by real world (4/6)

Recent large-scale tests (indirect FRVT 2002) indicate that  
Doddington's Zoo is alive and well and biting!

**Bottom Line:** Every person has their own error rate. This undermines large-scale estimates, reduces our ability to assess confidence intervals, and prevents us from answering the question “How big should the test be?”

And it makes the overall statistics really hard!

# Statistical models complicated by real world (5/6)

- **Sample size: How big should the test be?**
  - Rules of 3 and 30 provide lower bounds to number of attempts needed for a given level of accuracy, but the rules are over-optimistic
  - **Rule of 30:** Test until get 30 errors then 90% confident that observed error probability (e.g., FMR) is within  $\pm 30\%$  of true value
  - **Rule of 3:** If no errors in  $N$  independent tests then have 95% confidence level that error probability  $P < 3/N$ 
    - “What is the lowest error rate that can be statistically established with a given number  $N$  of independent identically distributed comparisons?”

# Statistical models complicated by real world (6/6)

- Confidence interval estimates
  - Confidence interval (uncertainty in the observed error rates, level of accuracy) depend on number of test subjects and probability of errors (e.g., FMR)
  - Once data has been collected and the comparisons have been made, then the confidence intervals can be estimated.
  - **Only then do we know whether the test was large enough!**
- Development of statistical models is a high priority research area

# Performance prediction models not available

- Performance in the lab has turned out to be a very poor indicator of performance in the field
  - The performance of biometrics is influenced by many factors, including environment, user behavior, and application
  - Currently do not have predictive theories to model these variables and to predict real-world performance of biometric systems
  - Cannot currently produce results analytically
- So . . . forced to use brute force, ignorant approach to obtain valid results
  - Testing a large number of people
- Needed research:
  - Models to predict operational performance from laboratory testing
  - Models to predict large database performance from small database performance

# Lack of accepted testing & reporting protocols (1/3)

- Currently, each biometric test is an **experiment** that (should) follow scientific method
- Why? Because multiple factors must be taken into account before an effective test protocol can be designed. For example:
  - What is the goal of the test?
    - To compare biometric products or components?
    - To estimate real-world performance?, etc.
  - What are the hypotheses?
    - Is this algorithm race or gender neutral?
    - Will this product meet the client's requirements in this specific climactic and human environment?, etc.
  - What are the **application**, **environment**, and **population** factors?
    - **These factors profoundly influence the performance of a biometric system and thus the testing protocol**
  - Which test variables should be independent, controlled, randomized, neglected?
- **Many** questions must be asked and answered and scientific method must be followed to obtain valid test results

# Lack of accepted testing & reporting protocols (2/3)

- If scientific method is not followed:
  - Reported results may be meaningless
  - Partial and sometimes misleading results may be reported
  - But . . . it is very hard to convince the general public that these results are meaningless and/or misleading
- Currently there is no scientifically-based, preferred general biometric testing protocol to follow that ensures valid test results
  - Too many variables
  - Little agreement to date on:
    - What are most important questions to be answered?
    - What are those “things” that are required for a biometric test to be considered reasonable?

# Lack of accepted testing & reporting protocols (3/3)

- Moving from experimental testing to **standardized** (UL-like) testing
  - To develop standards-based testing protocols, need to:
    - Agree on some standard hypotheses/questions that the testing will address
    - Apply scientific method to a few general, high-priority application profiles
    - Prescribe detailed reporting requirements
  - Some experts claim that it is premature to develop standard testing and reporting protocols until the science of biometric testing is more mature
  - However, governments and private industry are counting on biometrics to help improve border crossing security and personal identification programs
  - These potential consumers of biometric products must have reliable information to make appropriate procurement decisions
- International biometric performance testing and reporting standards efforts underway . . .

# Humans are involved

- Biometric technology is profoundly influenced by human behavior
  - Biometric input is not always the same
  - Difficult to adapt to input variations
- Typically testing human more than device
  - Need to quantify modes of human operation – human factors – that we do not understand
  - Need help from sociologists and psychologists
- Difficult to ensure participation from human test subjects to achieve good results
  - People don't always follow directions

# We are currently experiencing technical difficulties (1/3)

- Process errors dominate
  - Errors induced during the test processes and procedures are potentially larger than the errors being measured
  - Data entry errors have been found to be about 1/1000.
  - Biometric performance may be better than 1/1000, but we won't be able to tell.
- Scenario testing with a human test population
  - Logistically, max population for online scenario 200-250 people
    - Not because of statistical significance but because two expert testers can handle this many people
    - If use unskilled testers to acquire data, data errors increase and data quality decreases
- Technology testing timing
  - Everybody's hardware works on different platforms at different speeds

# We are currently experiencing technical difficulties (2/3)

- Data quality
  - Must continually monitor data, “meta-data” and testers to ensure that quality data is being gathered
  - But do not currently have metrics for biometric data quality
- Data acquisition and analysis
  - Design and acquisition of biometric databases is difficult
  - Do not have automated data acquisition and analysis tools

# We are currently experiencing technical difficulties (3/3)

- Basically, the world is not static
  - We cannot perfectly control the environment
    - For example, get different fingerprint results in the summer and winter because people adapt to the different climates, and their skin changes accordingly
    - In addition, the differences in the climate alone can change the performance of the human/sensor interface
  - We cannot perfectly control user behavior
    - Biometric input is not always the same
    - Difficult to adapt to input variations

# Hard to get representative data (1/3)

- Hard to get data
  - Harder to get right data
    - Even harder to get truly representative data

# Hard to get representative data (2/3)

- Need representative test population
  - Set of people using equipment in operational setting
  - Operational population
    - Example 1: Female fingers and ridge patterns are smaller and tend to crack. If target user population is mostly men, using an over-represented female population yields unrepresentative results
    - Example 2: Data taken from older people in a stressful situation by unskilled testers will not represent actual performance of a well-trained older population in a low-stress environment
- Need representative operational environment
  - A few influencing factors:
    - Lighting – face
    - Climate (dry/humid, temperature) – fingerprints
    - Occupation (dirty hands, worn ridges) – fingerprints
    - Background noise – voice

# Hard to get representative data (3/3)

- Difficult to ensure that equipment is tuned according to specs during testing (data acquisition) and during operational use
- Difficult to ensure consistent operational use by population
  - Quality of user interaction with device good at first
  - Then interaction quality regresses
    - Users get sloppy, no longer paying attention to detail
    - Good to incorporate end user re-training program and to monitor data quality
- Test results taken with tuned equipment and well-trained population may not be applicable over life of system

# Test results typically lack general applicability

- Results are typically valid only for the specific situation tested
  - Specific application
  - Specific environment
  - Specific population
  - Specific test goal
  - Specific . . .
- Test results typically have limited applicability to other scenarios

# Test results often lack statistical significance

- Need enough subjects to get statistically significant results, but how good is good enough?
  - Consider how variation in results (e.g., FAR) will influence overall application
  - A 3-4% change in performance may be acceptable depending on the size of the operation
    - For an expensive IAFIS system, small variations in performance will have significant cost consequences
    - For PC logical access control, small variation is likely quite acceptable
    - Tolerances (confidence intervals) → Cost!
  - Tolerances impact design, cost, and throughput of the system
    - How wrong can you afford to be?

# No one wants to pay (1/3)

- Consumers of test reports typically want testing without knowing why they want testing
  - Consumers want both internal (repeatable but not indicative of real-life performance) and external (indicative of real-life performance but not repeatable) consistency, but testing protocols are mutually exclusive
  - Consumers want results that are hardest (most expensive) to obtain at minimal cost
  - Consumers want a single figure of merit but none exists today
  - Consumers want the impossible
- Difficult to present test results in an understandable way to the general public

# No one wants to pay (2/3)

- The “best” biometric system just meets the users requirements but does not over satisfy those requirements
  - The consumer may not need a comprehensive test to determine if the system meets their requirements
- Once the “good” biometrics have been separated from the “beta” biometrics, other implementation issues, such as exception handling procedures, system security issues, user attitudes, etc., can influence system performance more than the biometric system

# No one wants to pay (3/3)

- Products change very quickly, and testing goes slowly – a product tested today may not be there tomorrow
- Product testing may not always be a good idea
- Testing costs a lot
  - in time, people, and resources
- Biometric testing has not “panned out” for anybody . . . yet

# Vendors are sensitive to test results

- Fundamental tension between those that supply biometric systems and those that test biometric systems
  - No strong vendor representation in SC37 Biometric Testing and Reporting Working Group
- Some vendors don't want to know results
  - Would rather have “illusion” of good performance
- Biometric evaluation is an area of research

# Biometric Sciences is an emerging discipline (1/2)

- There is a lack of general agreement amongst the experts on several biometric testing issues, perhaps due to the multi-disciplinary nature of the technology
  - Example: Experts cannot agree whether to count failure to enroll (FTE) as a false non-match (reject) when computing false non-match rate (FNMR)? Some say yes, because this is statistically correct; some say no because people who can't enroll will not use system. Who is right? It depends.
- The science of biometrics is not yet far enough advanced to definitively harmonize these differences
- Need enhanced interactions between the various biometric communities to share discovery at the fundamental level
- Biometrics is an emerging field of science and technology

# Biometric Testing: It's Not as Easy as You Think (1/2)

- Experts in the field do not agree on high-level concepts and terminology
- Testing theory not well understood
  - Testing paradigms controversial
  - Statistical models complicated by real world
  - Performance prediction models not available
- Lack of accepted testing & reporting protocols
- Humans are involved

# Biometric Testing: It's Not as Easy as You Think (2/2)

- We are currently experiencing technical difficulties
- Hard to get representative data
- Test results typically lack general applicability
- Test results often lack statistical significance
- No one wants to pay
- Vendors are sensitive to test results
- Biometric Sciences is an emerging discipline

# Acknowledgements

- Joe Campbell, MIT, USA
- John Campbell, 3-M AIT, Canada & USA
- Patrick Grother, NIST, USA
- George Keibuzinski, Mitretek, USA
- Hakil Kim, Inha University, Korea
- Tony Mansfield, NPL, UK
- Jonathon Phillips, DARPA & NIST, USA
- Colin Soutar, Bioscript, Canada & USA
- Philip Statham, CESG, UK
- Michael Thieme, IBG, USA
- Kaoru Uchida, NEC Corporation, Japan
- James Wayman, USA & UK

# Biometric Testing: It's Not as Easy as You Think

Valorie S. Valencia

valorie@authenti-corp.com



*... providing scientific consulting, evaluation, and training services in all areas of authentication*