

1

Forensic handwriting examination

2

Automatic handwriting and signature analysis

3

Differences between disciplines and their implications

4

Pathways to integration

1

Forensic handwriting examination

2

Automatic handwriting and signature analysis

3

Differences between disciplines and their implications

4

Pathways to integration

PART 3

Differences between disciplines and their implications

Data · Output · Evaluation · Reporting · What **FHE** need from *your discipline*

ICDAR 2026 Tutorial From Ink to AI: Innovations and Integration in Forensic Handwriting
Examination

Vienna · August 30, 2026

FRAMING

Two disciplines, two ways of asking the same question

THE QUESTIONS ARE POSED BY COURT/LAWYER

Was this questioned handwriting written by this person?

Writership and authentication of one specific document or one signature in one specific case with a reference set.

AUTOMATIC WRITER RECOGNITION (AWR)

Does this sample match the reference (1:1)? / is in the reference set (1:N)?

Statistical pattern-matching against a reference sample, trained for generalization.

Both disciplines are **pattern-recognition** based: both ask the same underlying question and both seek to express certainty about the answer.

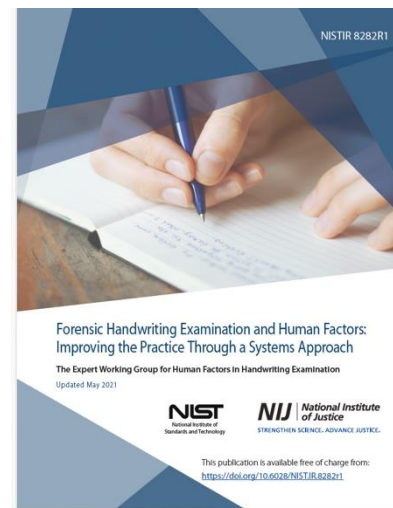
FOUNDATIONS

Established foundations

ENFSI: European Network of Forensic Science, 2022



NIST National Institute of Standards & Technology, 2021



Australia: Modular Forensic Handwriting Method, 2024

WILEY



In Wiley Interdisciplinary Reviews: Forensic Science

Source: ENFSI Best Practice Manual · NIST IR 8282r1 · Bird, Jones & Ballantyne, WIREs Forensic Science (2024)

NIST agrees

This publication is available free of charge from: <https://doi.org/10.6028/NIST.IR.8282r1>

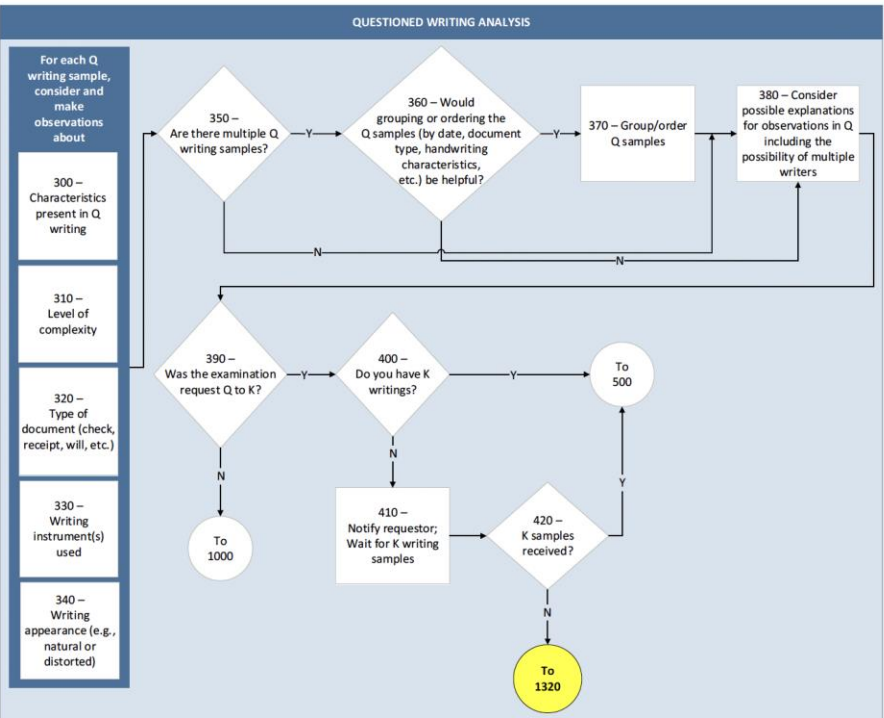
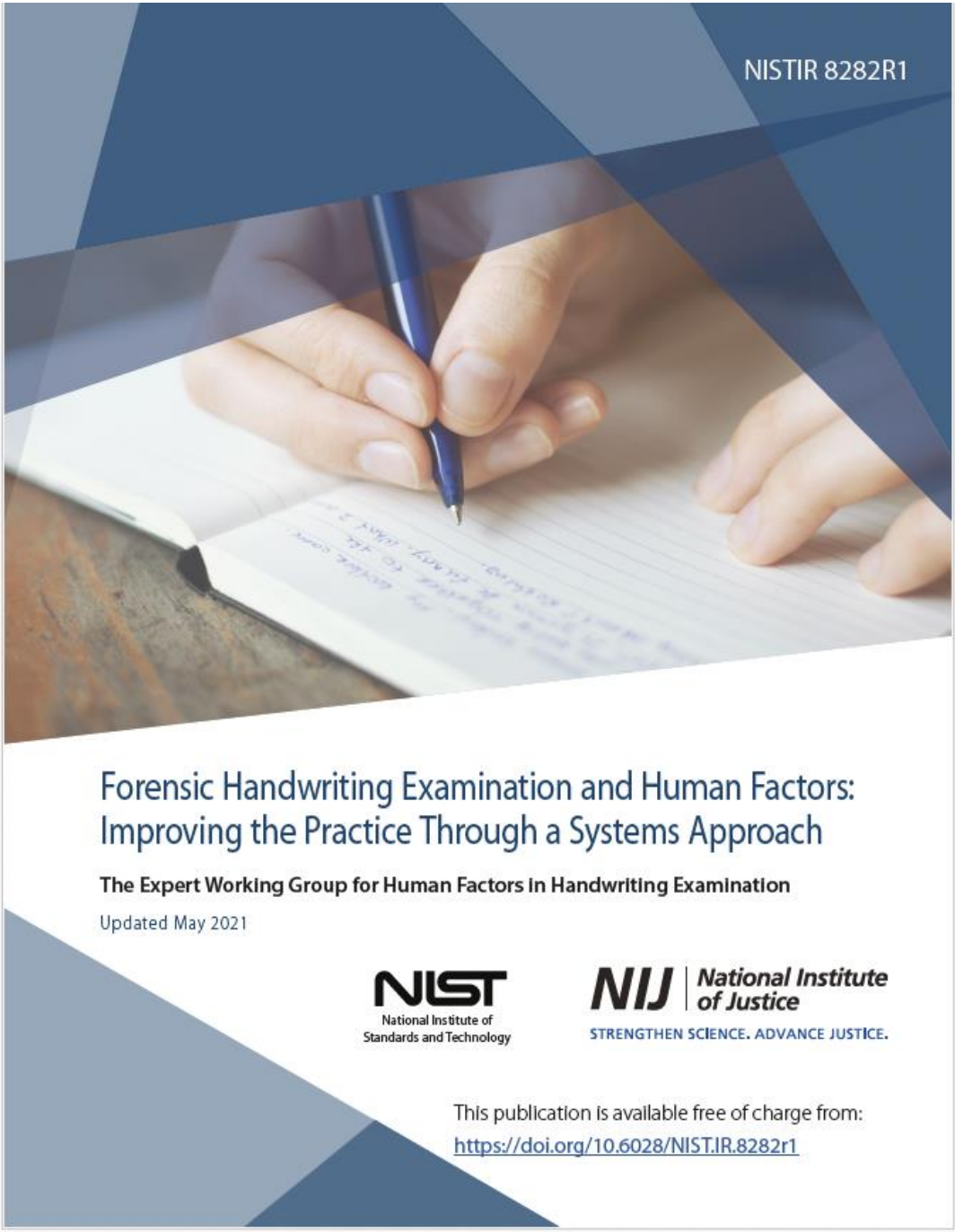


Figure 1.1: Handwriting examination process map (Continued)

Recommendation 2.8: The forensic document examiner community should collaborate with the computer science and engineering communities to develop and validate applicable, user-friendly automated systems.

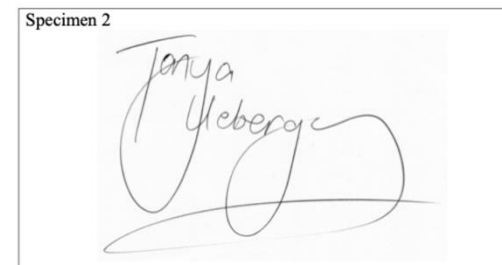
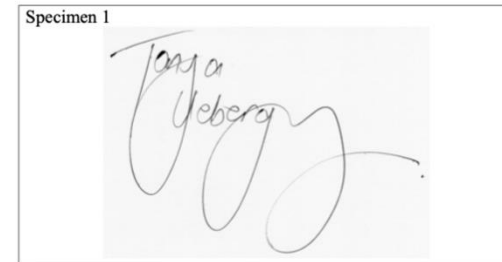


FOUNDATIONS

What is proficiency testing for FHE?

A proficiency test (PT) presents an FHE with material resembling real casework, where the ground truth 'who really wrote what' is known to the test organizer but not the examiner. The FHEs conclusions are then scored against that known truth.

- Error rates are calculated across false-positive, false-negative, and inconclusive responses
- An inconclusive answer is a legitimate, scoreable outcome, **not a missing data point**
- Results depend on who volunteers to participate, and on which items go unanswered



This is the closest FHE equivalent to benchmarking a model on a labeled test set.

Source: Crown et al., Forensic Sciences Research (2024)

SHARED HISTORY

Decades of shared ground

CEDAR-FOX · from 2002

Save

welcome to cedar

center of excellence
for document analysis
and recognition

about cedar
research
publications
patents
in the news
people
academics
resources
contact us



 **University of Buffalo** The State University of New York
Department of Computer Science & Engineering

WANDA · 2004

FISH-new:
A Common Ground for Writer Identification

Katrin Franke,
Fraunhofer IPK,
Berlin, Germany

Lambert Schomaker,
Rijksuniversiteit Groningen,
The Netherlands

Louis Vuurpijl,
Nijmegen University,
The Netherlands

Stefan Giesler,
Freiburg, Germany



Forensic Information System Handwriting

Two independent systems, **each explicitly designed around forensic requirements rather than pattern-recognition benchmarks alone.**

 Zurich Forensic Science Institute

 Friedrich-Alexander-Universität Erlangen-Nürnberg

Amsterdam University of Applied Sciences 

SHARED HISTORY

Decades of shared ground – FHE datasets

Venue	Competition name	What was tested	# systems tested in competition	Data tested
ICDAR 2009	SigComp09	Signatures, skilled forgeries	18 systems	WANDA signatures & NFI signatures
ICFHR 2010	4NSigComp10	Signatures, skilled forgeries, disguised signatures	6 systems	Australian LaTrobe signatures
ICDAR 2011	SigComp11	Signatures, skilled forgeries	13 systems	Chinese & Dutch NFI signatures
ICFHR 2012	4NSigComp12	Signatures, skilled forgeries, disguised signatures	5 systems	Australian LaTrobe signatures
ICDAR 2013	SigWiComp13	Signatures & HW	23 systems	Japanese & Dutch NFI data
ICDAR 2015	SigWiComp15	Signatures & HW	40 systems	German, Bengali, Italian data
ICDAR 2021	SVC 2021 SVC - onGoing	Dynamic & static signatures, skilled forgeries SVC – ongoing provides a complete panorama of the state-of-the art in the online signature verification field under realistic scenarios https://sites.google.com/view/SVC2021/home		



Same words, different meanings

TERM	FHE MEANING	AWR MEANING
Verification	Confirming an examiner's conclusion via independent review by a second examiner (the 'V' in ACE-V).	Confirming a claimed identity via a 1:1 match against a reference sample, based on score and threshold, YES/NO answer.
Genuine	Genuineness is the hypothesis that gets tested, not a given.	A sample labeled as authentic in a training/testing dataset.
Score	No score is calculated; the examiner attaches a probability-based gradation directly to a subjectively formed assessment.	A numerical output (distance, similarity, or likelihood) computed directly by the algorithm.
Error rate	Only measurable through controlled proficiency testing, tied to a specific task and sample, never one universal figure.	Computed directly from the test set (e.g. EER), a fixed property of the model and benchmark.

One case versus thousands of writers

FHE CASEWORK

- Individual cases, often with a small number of known samples
- Natural variation: pen, surface, writing conditions
- Forensically relevant conditions: disguise, simulation, tremor, or illness
- Material arrives as it is, without controlled acquisition

AWR BENCHMARKS

- Hundreds to thousands of writers per dataset
- Controlled acquisition: online + offline, fixed protocols
- Goal: generalizable models, not one individual case
- Increasingly supplemented with synthetic or augmented samples to scale training data

2 · DATA CHARACTERISTICS

Skilled forgery, different definitions

Granularity within 'forgery': AWR benchmarks do separate random from skilled forgeries for scoring (e.g. distinct EER values) but rarely go further. Acquisition conditions such as writing surface, time pressure, or posture are usually not captured as separate labels. FHE distinguishes at that finer level: a signature written under time pressure or on an unfamiliar surface is not a deliberate forgery.

A shared challenge, not just an AWR gap: both disciplines struggle to reliably distinguish, let alone score, high-quality skilled forgeries. AWR systems show measurably higher error rates against skilled forgeries than random ones; FHE likewise find well-executed simulations among the hardest cases to evaluate with confidence.

Neither discipline has a solved answer for skilled forgery; this is common ground for future collaboration, not a one-sided weakness.

3 · OUTPUT FORMATS

A verbal likelihood ratio conclusion versus a score

FHE CONCLUSION

The findings of the investigation provide

- extremely strong support ...
- very strong support...
- moderate support ...
- weak support ... for one proposition over the other
- equal support for both propositions
- elimination

Proposition-based evaluation: a verbal scale, with a Bayesian (LR) approach.

AWR OUTPUT

- Similarity score
- Score-based likelihood ratio
- Ranked list of candidate writers

Numeric, continuous, or ranked, no fixed verbal scale.

3 · OUTPUT FORMATS

When does a score become a likelihood ratio (LR) and when not?

Forensic LR framework: a structured, evidence-based reasoning method: The findings are weighed against two competing hypotheses.

A score only becomes an LR once it is calibrated against a representative population, otherwise it is a distance measure wearing an LR label.

In the best practices, there is no prescribed reporting format or standard for the implementation of AWR: a methodological gap, not a deliberate choice.

'Reproducible' does not mean the same thing in both worlds

	FHE	AWR
Repeatability	Same examiner, same method, same conditions.	Same model, same test set, same run, identical result (e.g. same EER value).
Reproducibility	Different examiners and/or conditions, same method: do they reach the same conclusion?	Reproducing the same result on the same fixed dataset.
Known weakness	Proficiency test results depend on who volunteers to participate, and on unanswered items within the test, both can lead to underestimated error rates.	EER is sensitive to how representative the test set is of forensic conditions (disguise, quality).

ACE-V versus classifier pipeline



The report always includes limitations, and discusses the findings under both propositions, where uncertainties are openly presented. Conclusions convey the strength of the evidence.



Implicit assumption in many AWR systems: one run = the answer without built-in second-reader step.

AWR produces a score or a yes/no outcome, but not a step-by-step, auditable reasoning process that aligns with ACE-V for court.

TOWARD PART 4

The gap is real .. and bridgeable

Data · Output · Evaluation · Reporting · What FHE needs from CS

1

Non-representative
datasets

2

Limited
explainability

3

Mismatched output
formats

4

Weak workflow
integration

Four bottlenecks we consider as the reason for the gap between benchmark performance and forensic application.

Proof that a bridge is possible: explainable AWR systems (Diaz et al., 2024) show transparent features, understandable distance measures, and output aligned with how FHE already think.

What FHE needs from AWR

1

Verify

AWR that can verify the subjectively formed FHE interpretation.

2

Quantify

AWR that can quantify the strength of evidence, expressed as a likelihood ratio.

3

Explain

AWR that produce explainable output.

4

Update

Hypotheses that can be revised as new evidence becomes available.

5

Respect fundamentals

AWR that account for fundamental aspects of handwriting (e.g. natural variation).

→ Part 4 will build on this: benchmarking, validation, reproducible experiments, shared tools.

Trace over these letter shapes. Use a different colour pencil crayon each time.

ai ai ai ai ai ai

ai ai ai ai ai ai

Now practise writing these letters on the lines below.



ai ai _____