Validity, Reliability, Interpretation, and Opinion Scales as presented in the NIST/NIJ Expert Working Group for Human Factors in Forensic Handwriting Examination Report

Reinoud D. Stoel[1], Christopher P. Saunders[2], Mara L. Merlino[3], Nikola K. P. Osborne[4], Jonathan Jackson-Morris[5], Michael P. Caligiuri[6]

Reference: Reinoud D. Stoel, Christopher P. Sanders, Mara L. Merlino, Nikola K.P. Osborne, Jonathan Jackson-Morris, Michael P. Caligiuri (2021). SPECIAL ISSUE: Introduction to the 2020 NIST/NIJ Expert Working Group for Human Factors in Handwriting Examination Report. J. Forensic Document Examination, Vol 30,19-26.

Introduction

The *Forensic Handwriting Examination and Human Factors: Improving the Practice Through a Systems Approach* Report (HFHE Report [1]), produced by the NIST/NIJ Expert Working Group on Human Factors in Forensic Handwriting Examination (the Working Group), addressed whether the underlying scientific principles, the measurement validity, and reliability of the analytical methods employed by Forensic Document Examiners (FDEs) are sufficient to withstand the challenges posed by the *Daubert* factors and Federal Rule of Evidence (FRE) 702. While judges, litigants, legal scholars, and forensic scientists may differ in what each view as acceptable scientific validity, the question remains whether FDEs can *demonstrate* that the methods used were scientifically derived and appropriately applied to the evidence in question. In this summary, we first examine the durability of the scientific principles underlying handwriting examination. We then synthesize the current state of measurement reliability and validity as applied to handwriting evidence, judgments, and opinions of expert FDEs. This article provides a brief overview of these topics as presented in chapters 2 and 3 of the HFHE Report [1]. Readers should also refer to the introductory article of this special series [2].

1 Statistics Netherlands, The Hague, The Netherlands
2 South Dakota State University, Brookings, SD
3 Kentucky State University, Kentucky
4 Human Factors Training and Consultancy, Auckland, New Zealand
5 Forensic Services SPA, Scotland
6 University of California San Diego, San Diego, CA

Shifting from Uniqueness based to Evidence-based Principles of Handwriting Examination

Two traditional principles or axioms of handwriting examination remain strong in the FDE's psyche. The first is the principle of individuality that "no two writers share the same combination of handwriting characteristics given sufficient quantity and quality of writing to compare." [3]. The second is the principle "that no two writings by the same person are identical" [4. p. 27]. These principles formed the basis decades ago for the development, application, and interpretation of feature comparison methods in forensic handwriting examination. While it may be that every instance of handwriting is "unique" in that distinctive features characterize it, claiming uniqueness in a source document is not useful for writer identification. "Uniqueness" and "individualization" in forensic science no longer correspond to the traditional, strict interpretation of these terms and can lead to exaggerated strength of the evidence. Statistical reasoning does not support one source attribution to the exclusion of all others. The Working Group endorses a shift to evidence-based principles of handwriting examination to account for sources and range of natural variation between and within individual writers.

The nature of intra- and inter-writer variation has deep roots in motor control theory. Motor control theory treats the handwritten stroke to be the base unit. The temporal and spatial-geometric characteristics of handwritten strokes are programmed, sequenced, and executed by the central nervous system. Over time,

an individual learns or habituates complex sequences of motor commands, thus reducing the demands placed on memory and motor systems during natural writing [5]. As complex sequences of handwriting movements become habituated over time, feature variability decreases within an individual writer while the flexibility to adapt to changing environmental or physical constraints increases. These properties of motor control theory (e.g., motor equivalence) lead to predictions about natural variation, including the prediction that certain features of handwriting remain invariant throughout changes in writing surface, orientation, or whether the individual wrote with the dominant or non-dominant hand.

Empirical research driven by handwriting motor control theory can potentially shift the foundation of forensic handwriting examination from the assumptions of individualization to a neurobiological approach. For example, Brault and Plamondon [6] developed a coefficient to measure the difficulty of forging a signature based on a formulation that models the complex processes of perception, memorization, and muscle coordination that the forger employs to execute a simulation. Found et al.[7] and others [8] applied the "kinematic approach" to signatures and observed that the number of turning points and line intersections or retraces best explain the FDE's assessment of signature complexity.

It is important to demonstrate that analytic methods of handwriting and signature examination and interpretation based on these methods are both reliable and valid. Chapter 2 of the HFHE Report aims to distinguish between foundational validity and validity in practice relying heavily on modern scientific principles of handwriting motor control. In this synopsis, we offer a brief review of these psychometric properties of measurement science and how they impact FDE's evaluation and interpretation of handwriting evidence.

Measurement Reliability and Validity

The central aim of section 2.2 of the HFHE Report [1] was to offer a conceptual review of human factors impacting the reliability and validity of methods used in forensic handwriting and signature examination and their interpretation. The HFHE Report did not summarize the results from reliability or validity studies per se. Instead, we refer the reader to Merlino [9], which summarizes findings supporting the construct and discriminant validity of handwriting and signature verification.

In its discussion of reliability and validity, the HFHE Report describes the properties of these and other measurement instruments. These are reproduced in Box 1. When considering reliability in handwriting examination, both repeatability and reproducibility should be measured. The same FDE should rate similar cases in similar ways (inter-rater reliability/repeatability), and different FDEs should rate similar cases in similar ways (intra-rater reliability/reproducibility). These judgments should be compared on ground-truth-known samples.

Validation research involving handwriting often considers the decisional process and opinions of the examiner as a "black-box". The FDE's decisional processes are not transparent to the researcher, hence the term black-box. The aim of this research is to measure the agreement between the opinions expressed by the FDE and ground truth. Proficiency studies may be considered black-box studies. Conversely, "white-box" studies attempt to validate detailed methods or algorithms of the decisional process. In handwriting and signature verification, white box studies assess the degree to which methods or algorithms, transparent to the researchers, lead to writership determinations that align with ground truth. The HFHE Report recommends that the document examination community collaborate with researchers to design "black box" and "white box" studies to assess reliability in forensic handwriting examination ([1] Recommendation 2.4). The HFHE Report also summarizes descriptions in several guidance documents of measurement validity and reliability of analysis methods:

- 2009 National Research Council of the National Academy of Sciences (NAS) report on strengthening forensic science in the United States (NRC report, [12])
- European Network of Forensic Science Institutes (ENFSI) *Best Practice Manual for the Forensic Examination of Handwriting* [13]
- *Latent Print Examination and Human*

Box 1: Properties of good measurement instruments ([1] Box 2.2, page 58)

Reliability [10]: How often single or multiple examiners reach the same answer under specified tasks and constant conditions. Reliability is related to the instrument's degree of random measurement error, which can include the examiner. The smaller the random error, the more reliable the instrument, and vice versa. Two ways to measure reliability are repeatability and reproducibility.

Repeatability: A measure of reliability using the same examiner and the same instrument under precisely the same conditions.

Reproducibility: A measure of reliability using different examiners and/or differing conditions with the same measurement instrument to arrive at the same conclusion or result.

Precision: Equivalent to reliability. A reflection of the degree of random error in a measurement. A precise measurement instrument has minimal random error.

Validity [11]: The absence of both random and systematic measurement errors; can exist in degrees. A measure can be reliable and not valid, but not vice versa. In other words, reliability is necessary but not sufficient for validity. If a measurement instrument is valid, it is also reliable.

Accuracy: Similar to validity; however, accuracy refers only to systematic measurement error.

Factors report (Latent Print Report, [14])

- *Fundamentals of Probability and Statistical Evidence in Criminal Proceedings*, published by the Royal Statistical Society [15]
- 2016 President's Council of Advisors on Science and Technology (PCAST) report on ensuring scientific validity of feature comparison methods [16]

Due to its focus on human factors, the views expressed in the HFHE Report most closely align with the Latent Print Report produced by the NIST/NIJ Expert Working Group on Human Factors in Latent Print Analysis produced [14]. In that report, they discussed measurement validity, error rates, and the appropriateness of measurements and judgments made in latent print analysis. Not surprisingly, the HFHE Report agrees with both the Latent Print Report and the ENSFI criteria stating that to be valid, both a procedure and the measurement tool used to assess the procedure's validity should be fit for purpose. The HFHE Report also agrees that "validity" is a relative term. In an excerpt from the HFHE Report, "demonstrating that comparison procedures may be valid to evaluate the evidence given one set of propositions does not imply that the same procedures are valid for evaluating the evidence given other propositions. For example, the extent to which a measurement is considered a valid representation of an attribute will depend on whether the methods are designed to serve a specific purpose (e.g., measuring attributes of genuine versus simulated signatures might not be valid for hand-printed material)." ([1], p. 59 – 60). For further discussion, see sections 2.2.1 and 2.2.2 of the HFHE Report.

Interpreting Handwriting Evidence

The Working Group discussed in detail the interpretation phase of forensic handwriting examination. Different approaches to interpretation

exist, and the HFHE Report reflects on this discussion in Section 2.3. The HFHE Report focused on two common approaches, including the traditional or two-stage approach and the likelihood-ratio or Bayesian approach to interpreting and presenting handwriting evidence.

In the classical approach, the examiner reaches a conclusion utilizing propositions such as "the signature was produced by the person of that name," or "the threatening letter was (or was not) written by the suspect." Such propositions are denoted as H1 (and H2) for brevity and the putative writer as W1. Conventionally, an FDE might opine that the writer is individualized with a high degree of certainty, based on the classical premise that no one else in the relevant population could have signed the name or written the words on the questioned document.

In a variant of this approach, the FDE will first decide whether the suspect could have written the questioned document based on the similarities and dissimilarities observed between the questioned document and the known writing samples. If the suspect writer cannot be excluded as the writer of the questioned document, then the examiner considers the rate at which alternative writers cannot be excluded as the source of the questioned document. This rate is the "coincidence probability." [17] If the suspect cannot be excluded and the coincidence probability is sufficiently low, then the evidence favors H1; the larger the coincidence probability, the weaker the evidence becomes. Some literature on forensic statistics debates the reasonableness of the coincidence probability [18] which, in the handwriting examination context, corresponds to the rate at which alternative sources "match" the questioned document. A further variant is to map these coincidence probabilities to a reporting scale with a set of ordered categories such as "true," "false," or "inconclusive," perhaps adding terms such as "strong probability," "probable," and "indications." [19] Even though the coincidence probability is defined as a frequentist probability, it is typically estimated subjectively based on the FDE's experience and then mapped to a conclusion scale.

These variants share a common thread: they presuppose that the FDE's task is to give some opinion in support of any proposition, here referred to as H1 (if the samples are adequate to perform an examination). However, the usefulness and appropriateness of this conventional interpretative framework have been questioned. In particular, one can question the premise that the expert should reach *any* decision (qualified or otherwise) about H1. Although expert opinions about matters that a judge or jury must ultimately resolve are generally permissible, no rule of law or scientific principle requires this. The expert need not proffer an opinion about H1—or be compelled to do so—to contribute scientific information to the resolution of a case. The strengths and weaknesses of this interpretative framework have been debated in the literature [20] and throughout the HFHE Report [1].

Expert testimony should indicate the degree of correspondence the FDE identifies between samples. The FDE's findings should report the extent of support for the proposition that the suspect produced the questioned writing (H1) relative to the alternate proposition that the suspect did not produce the questioned writing (H2). Increasing consensus indicates that this would be most helpful to the court or jury. Section 2.3.2.1. of the HFHE Report details aspects of propositions that experts should consider in practice. Interpreting evidence in a relative manner is described in papers and books [11, 21, 22] as the "Bayesian approach" or the "Likelihood Ratio approach" in the forensic community. A growing number of forensic laboratories worldwide have adopted this approach for simple forms of evidence where competing propositions can be explicitly stated. This approach asks the expert to limit evaluative conclusions to the degree of relative support that the evidence provides for H_1 compared to the alternative H_2. In using this approach, the FDE must find a proper way to "measure" the findings' support for each proposition ([1] See box 2.3). Although some researchers have criticized this approach, many advocate that subjective probability is the best method [23].

The HFHE Report discusses both frequentist and subjective approaches to probability and the discussion among forensic statisticians and forensic scientists about the concept of probability and how to apply it in forensic science. This discussion has deep roots in statistical and mathematical science and may never reach a solution that satisfies all those contributing to the discussion. The HFHE Report states that the

fundamental discussion between the frequentist and subjective approaches should not prompt stakeholders to dismiss probability as the core concept in forensic science evidence evaluation. The HFHE Report suggests that the current definition of probability used in forensics is vague. The HFHE Report stresses that, given the complexity of using probabilistic reasoning to interpret handwriting evidence, FDEs will require a basic and precise knowledge of the differences and uses of the two types of probability. The ENFSI *Guideline for Evaluative Reporting in Forensic Science* [24] provides recommendations for implementing the subjective likelihood ratio approach. The American Statistical Association provides additional guidance on the use of probabilistic methods for the presentation of evidence for the US judicial systems [25].

The HFHE Report urges the use of objective estimates of frequency of occurrence and the construction of databases to support interpretation. Such databases are valuable tools for estimating the frequency of occurrence of inter-writer and intra-writer features and combinations of features. These databases will need to contain a large amount of writing, where all features of interest in the writing have been measured, provide insight into, and estimates of, the frequencies and interdependences of salient features in various populations. They should also contain demographic data so that stakeholders will understand the gaps of empirical data used in the methods validated with said databases.

Furthermore, the HFHE Report identifies some key priorities for feature interpretation research studies. Such studies should focus on writing complexity, the development of methods of quantifying and measuring inter-writer and intra-writer variability, the amount of writing required to reach a conclusion about the writership of the questioned writings, comparability of types of writing, relevant information (features) identified in writing samples, and the extent of the consistencies in interpreting such information.

Opinion Scales

While the HFHE Report discusses the merits of different approaches to interpreting forensic evidence in general ([1], Section 3.3), the opinion expressed and the opinion scale used should match the examiner's working approach. In current practice, different opinion scales are used, and here we present an overview of these scales. The predominant opinion scales in practice today are verbal, but in principle, examiners working from the Likelihood Ratio approach could use a quantitative scale. Opinions are mainly reported using ordinal scales ranging from as few as three to as many as thirteen levels. The laboratory or FDE ultimately determines the formation and use of any scale.

There are many different scales that represent efforts made by different parts of the FDE community. The Scientific Working Group for Forensic Document Examination (SWGDOC) published Standard Terminology for Expressing Conclusions of Forensic Document Examiners, providing nine opinions (and associated descriptions) that an FDE may express. The Federal Bureau of Investigation (FBI) laboratory uses five categories that collapse SWGDOC opinions (2) through (4) into "may have (qualified opinion)" and opinions (6) through (8) into "may not have (qualified opinion)." Collaborative Testing Services (CTS), an examination proficiency test provider, uses another 5-category scale. An even simpler scale treats the FDE judgment as a binary (yes/no) decision. In this case, a positive association indicates the questioned writing is the subject's, or a negative association indicates the questioned writing is not. Examiners may reserve judgment by stating that the information in the samples is inconclusive.

Figure 3.1 in the HFHE Report ([1], p. 93) presents examples of the different opinion scales used globally in the practice of forensic handwriting examination. Although opinion scales are not scientifically rigorous, FDEs and the courts often view conclusion terminology as ordinal scales. This view has some inherent problems. An ordinal scale arises from the function of rank-ordering [26] and can demonstrate a gradation of the FDE's opinion strength. However, the gradation levels between the opinion levels are not quantified (except in the likelihood ratio scale). There may be variance between examiners in how they view the degree of difference between the scale levels. For example, when using the subjective posterior probabilities, it is not possible for an examiner to clearly define the degree of difference between "highly probable" and "probable." The examiners can only say that probable is the weaker or less strong of the two opinion levels.

Commonly Used Opinion Scales			
ENFSI LR scale	**SWGDOC**	**ENFHEX**	**Modular Approach**
Extremely strong support H1 over H2	Identification	Extremely strong support H1 over H2	Very strong support for H1 over H2
Very strong support H1 over H2	Strong probability (was written by)	Strong support H1 over H2	
Strong support H1 over H2			
Moderately strong support H1 over H2	Probable (written by)	Moderate support H1 over H2	Qualified support for H1 over H2
Moderate support H1 over H2			
Weak support H1 over H2	Indications (written by)	Limited support H1 over H2	
Findings do not support H1 over H2	No conclusion	Inconclusive	Approximately equal support for H1 over H2
Weak support H2 over H1	Indications (not written by)	Limited support H2 over H1	
Moderate support H2 over H1			Qualified support for H2 over H1
Moderately strong support H2 over H1	Probable (not written by)	Moderate support H2 over H1	
Strong support H2 over H1			
Very strong support H2 over H1	Strong probability (was written by)	Strong support H2 over H1	Very strong support for H2 over H1
Extremely strong support H2 over H1	Elimination	Extremely strong support H1 over H2	

Figure 1. Potential equivalency levels for various commonly used opinion scales

In the traditional scales (5-, 7-, and 9-point), the FDE expresses opinions corresponding to the traditional approach to handwriting analysis. While examiners may state these opinions in probabilistic terms (e.g., probably wrote), their precise meaning may be inconsistent across FDEs. For example, some FDEs may render an opinion based on the rarity of features, while others base their opinions on perceived evidential strength. When presenting evidence using the traditional scales, there is always a step where the FDE decides whether the writer of the known writing samples could have written the questioned document. In contrast, when using the modular [27] and likelihood ratio approaches, the FDE is expressing the strength of the evidence in terms of two or more mutually exclusive propositions or hypotheses without first considering the typicality of the questioned document given what is known about the suspect writer. This opinion is generally expressed as the strength of support for one proposition or hypothesis over one or more mutually competing propositions.

The three overall consistent levels across the traditional "scales" FDEs currently use are identification, inconclusive, and elimination. There are no identification or elimination opinions in the modular approach and the Likelihood Ratio scale. There is some criticism on whether categorical conclusions like identification can be provided by a forensic examiner at all. This criticism is related to the question of whether the examiner can make such a 'leap of faith' and whether it is the task of the examiner to do so.

Figure 1 displays potentially equivalency levels for commonly used opinion scales. However, it is important to note that it is difficult to map or relate the different types of scales because:

1. The traditional scales address the probability of the proposition, while the modular and likelihood ratio approaches focus on the probability of the findings given the proposition; thus, the traditional scale scales do not equate to the other approaches.
2. All scales lack sufficient study and empirical evaluation; therefore, the consistency of application across examiners is not well understood.
3. There would be fundamental mathematical issues in mapping the discrete categories in the different scales unless there was some common reference point or "anchor" between each scale.

The definitive conclusions (identification and elimination) on all conventional scales appear to have consistent application across the FDE community. The scales also share the center point of the inconclusive category but not the range. While the different scales might share the same meaning for identification, elimination, or possibly inconclusive, the sufficiency of the evidence that an individual FDE may use to support that conclusion may not be equivalent. FDEs have reported[7] that the actual category boundaries of the scale are subjectively determined during the evaluation and depend on the extent of perceived differences or similarities among the questioned and known writings and the limitations of the materials examined.

Finally, the HFHE Report encourages the FDE community to move toward a unified, standard approach for expressing conclusions to address some of the issues above and making the different approaches to the presentation of evidence more scientifically rigorous.

Reference List

The Expert Working Group for Human Factors in Handwriting Examination. Forensic Handwriting Examination and Human Factors. Improving the Practice Through a Systems Approach Revision 1. U.S. Department of Commerce, National Institute of Standards and Technology, Internal Report 8282R1; 2021, https://doi.org/10.6028/NIST.IR.8282r1.

Osborne NKP, Bird C. Introduction to the NIST/NIJ Expert Working Group for Human Factors in Handwriting Examination Report Special Edition. Journal of Forensic Document Examination, 2022 [insert edition and DOI].

Harrison, D., Burkes T.M., and Seiger, D.P. Handwriting examination: Meeting the challenges of science and the law. Forensic Science Communications 2009; 11(4): 1-13.

Huber, R.A. and Headrick, A.M. Handwriting Identification: Facts and Fundamentals. Boca Raton: CRC Press LLC. 1999. p. 27.

Caligiuri, M.P. and Mohammed, L.A. The Neuroscience of Handwriting. Boca Raton: CRC Press. 2012. p. 35-56.

Brault, J. and Plamondon, R. A complexity measure of handwritten curves: Modeling of dynamic signature forgery. IEEE Transactions on Systems, Man, and Cybernetics 1993; 23(2): 400–413.

Found, B., Rogers, D. Rowe, V., and Dick, D. Statistical modelling of experts' perceptions of the ease of signature simulation. Journal of Forensic Document Examination 1998; 11: 73–99.

Alewijnse, L.C., van den Heuvel, C.E., and Stoel, R.D. Analysis of signature complexity. Journal of Forensic Document Examination 2011; 21: 37–49.

Merlino, M.L., T.M. Freeman, V. Springer, V. Dahir, D. Hammond, A.D. Dyer, B.J. Found, L. Smith, and I. Duvall. 2015. Final Report for the National Institute of Justice grant titled Validity, Reliability, Accuracy, and Bias in Forensic Signature Identification. https://www.ojp.gov/pdffiles1/nij/grants/248565.pdf

Ulery, B.T., Hicklin, R.A., Buscaglia, J., and Roberts, M.A. Repeatability and reproducibility of decisions by latent fingerprint examiners. PLoSOne 2012; 7(3): 1–12. e32800. https://doi.org/10.1371/journal.pone.0032800.

Borsboom, D., Mellenbergh, G.J., and van Heerden, J. The concept of validity. Psychological Review 2004; 111: 1061–1071.

National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: The National Academies Press. 2009. https://doi.org/10.17226/12589.

ENFSI. Best Practice Manual for the Forensic Examination of Handwriting. ENFSI-BPM-FHX-01, November 2015. Available from: http://enfsi.eu/wp-content/uploads/2016/09/2._forensic_examination_of_handwriting_0.pdf.

Expert Working Group on Human Factors in Latent Print Analysis. Latent Print Examination and Human Factors: Improving the Practice through a Systems Approach. U.S. Department of Commerce. NIST. Washington, DC. 2015

Aitken, C., Roberts, P., and Jackson, G. Fundamentals of Probability and Statistical Evidence in Criminal Proceedings: Guidance for Judges, Lawyers, Forensic Scientists and Expert Witnesses. Royal Statistical Society. 2010. Available from: http://www.rss.org.uk/Images/PDF/influencing-change/rss-fundamentals-probability-statistical-evidence.pdf.

Executive Office of the President, PCAST. Report to the President Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods. 2016.

Evett, I.W., and J.A. Lambert. The interpretation of refractive index measurements. Forensic Science International 1982; 20(3): 237–245

Stoney, D.A. Evaluation of associative evidence: Choosing the relevant question. Journal of the Forensic Science Society 1984; 24(5): 473–482.

SWGDOC, Version 20132; ASTM E1658-08. Standard Terminology for Expressing Conclusions of Forensic Document Examiners. West Conshohocken: ASTM International. 2013. Available from: www.astm.org. (Withdrawn and replaced in 2017)

Balding, D.J. Weight-of-Evidence for Forensic DNA Profiles. Hoboken: John Wiley & Sons. 2005.

Buckleton, J.S., C.M. Triggs, and C. Champod. An extended likelihood ratio framework for interpreting evidence. Science & Justice 2006. 46(2): 69–78

Robertson, B., G.A. Vignaux, and C.E.H. Berger. Interpreting Evidence: Evaluating Forensic Science in the Courtroom. Second Edition. Chichester: Wiley 2016.

RvT 2010. EWCA Crim 2439.

ENFSI. Guideline for Evaluative Reporting in Forensic Science. Approved version 3.0. 2015. Available from: m1_guideline.pdf (enfsi.eu)

American Statistical Association. American Statistical Association Position on Statistical Statements for Forensic Evidence. 2019 https://www.amstat.org/asa/files/pdfs/POL-ForensicScience.pdf

Stevens, S.S. On the theory of scales of measurement. Science 1946; 103(2684): 677–680.

Found, B.J., and Bird, C. The modular forensic handwriting method. Journal of Forensic Document Examination 2016; 26: 7-83. https://doi.org/10.31974/jfde26-7-83