Science and Nonscience in the Courts: *Daubert* Meets Handwriting Identification Expertise

*D. Michael Risinger**
*with Michael J. Saks***

INTRODUCTION

Document examiners do many things. They examine typewriting for signs of idiosyncratic typeface wear and alignment that might indicate a common origin for two documents. They analyze ink to reveal its physical and chemical properties. They scrutinize the alignment of printed lines and the overlap of handwritten lines to determine if words or phrases have been after-inserted. And they analyze the composition, method of production, and watermark of paper to ascertain its probable origin and, in some cases, its age. In performing these tasks, the document examiner may often use specialized knowledge of manufacturing processes and manufacturer specifications, not unlike that employed in firearms identification concerning the relative number, spacing, pitch, and direction of twist of grooves and lands in various makes of rifled barrels.

All of these functions generally share the strengths and weaknesses otherwise associated with toolmark evidence, forensic chemistry, and the like. They are not, however, what this Article is about. This Article concerns the asserted skill that historically formed the foundation of the document examiner's trade and that still comprises a surprisingly high percentage of the everyday work of document examiners both in and out of court.[1] This skill is the asserted ability to determine the authorship *vel non* of a piece of handwriting by examining the way in which the letters are inscribed, shaped, and joined[2] and comparing it to exemplars of a

* Prof. of Law, Seton Hall Univ. School of Law; B.A., Yale Univ., 1966; J.D., Harvard Law School, 1969.

** Prof. of Law, Univ. of Iowa; B.A., B.S., Penn. State Univ., 1969; M.A., 1972, Ph.D., 1975, Ohio State Univ.; M.S.L, Yale Univ., 1983. Prof. Saks claims Prof. Risinger wrote 90% of this article. Prof. Risinger claims Prof. Saks's contributions fully entitle him to co-authorship. The "with" was their way of working this out.

1. For instance, in his January 1993 testimony in *U.S. v. Smyth*, an international extradition case involving an alleged IRA terrorist, Special Agent Richard M. Williams, an FBI documents examiner with 17 years of experience, testified that "the bulk of" his work dealt with handwriting. Transcript at 231, United States v. Smyth, 863 F. Supp. 1137 (N.D. Cal. 1994) (No. CR-92-0152-Misc-Bac) [hereinafter *Smyth* transcript].

2. That is, by examining the characteristics of what is sometimes called the "static trace"

putative author's concededly authentic handwriting.

Two events have occurred in recent years which combined to stimulate reevaluation of handwriting identification expertise. The first was the 1989 publication of an article in the *University of Pennsylvania Law Review*, pointing out the lack of empirical validation of the claims of the expertise.[3] The other was the U.S. Supreme Court's 1993 decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*,[4] which rejected previous approaches to the acceptability of scientific expertise under the Federal Rules of Evidence, and put into play the validity of claims of scientific expertise even in areas in which that expertise had long been taken for granted. As a result there has been a flurry of litigation over the validity of handwriting identification expertise, resulting in several recent federal decisions.[5] Before examining these cases, however, a little historical context might be helpful to a full understanding of the problems they present.

HANDWRITING EVIDENCE AT COMMON LAW

The notion that handwriting can be used to identify its author is very old,[6] as is the notion that a person can learn to make such an identification by study. Attempts to develop a system of handwriting expertise appear to have started in Italy and France in the seventeenth century,[7] and by 1737 were well enough accepted in France to have been incorporated into the law. The Code du Faux (Code concerning Forgeries) contained detailed provisions for regulating the collection of exemplars and their presentation to handwriting identification experts, which from the context of the code we may conclude formed a professional cadre of fair number.[8] However, no such claimed expertise then existed in the English speaking

---

left behind by the dynamic act of writing.

3. D. Michael Risinger, Mark P. Denbeaux, & Michael J. Saks, *Exorcism of Ignorance as a Proxy for Rational Knowledge: The Lessons of Handwriting Identification Expertise*, 137 U. Pa. L. Rev. 731 (1989) [hereinafter Risinger et al.].

4. 509 U.S. 579 (1993).

5. Judge Lawrence McKenna analyzed the admissibility of handwriting identification evidence after *Daubert* in his lengthy opinion in *United States v. Starzecpyzel*, 880 F. Supp. 1027 (S.D.N.Y. 1995); the United States Army Criminal Court of Appeals followed McKenna's lead in *United States v. Ruth*, 42 M.J. 730 (1995); and the Third Circuit addressed the issue in *United States v. Velasquez*, 64 F.3d 844 (3d Cir. 1995).

6. Huntington Hartford quotes Aristotle as observing that "[j]ust as all men do not have the same speech-sounds, neither do they have the same writing." Huntington Hartford, You Are What You Write 43 (1973) (quoting Aristotle, On Interpretation, Part 1).

7. The earliest treatise in this line, of a definite graphological cast, appears to be Camillo Baldi, Trattato come da una lettera missiva si conoscano la natura, e qualità dello scrittore [An Essay on the Means of Examining the Character and Qualities of a Writer from His Letters] (Milano, Gio. Batt. Bidelli 1625).

8. *See generally* François Serpillon, Code du Faux, ou Commentaire sur l'Ordonnance du Juillet, 1737, avec une instruction pour les experts en matière de faux (Lyon, Gabriel Regnault 1774).

world. As we shall see, a corps of asserted experts came into existence in Anglo-American courts only after lawyers persuaded the courts to accept the idea that such an expertise might exist.

Experts of any kind did not play a large role in litigation during the common law period. Even in the occasional instances when experts testified, they did so without a clearly defined set of legal principles governing their qualifications or use of their testimony. Not until 1782, however, in the case of *Folkes v. Chadd*, did a reported decision affirm the propriety of the use of such "skilled witnesses," although the practice at trial was of course much older.[9] The earliest trial use of skilled witnesses of which we have a record occurred in the trial of the Earl of Pembroke for the murder of Nathanael Cony in 1678[10] (though it does not appear to have been regarded as a novelty by the judges in that case).

Until the end of the 18th century, every example of expert testimony in the English speaking world involved witnesses whose expertise had primary application to practical affairs outside the courtroom, such as the physicians in The Earl of Pembroke's case or the engineers in *Folkes v. Chadd*. If expertise in general then played a small role in the law, claimed expertise limited to issues that were or might be involved in court cases (what one might call a dominantly forensic expertise), was nonexistent. The first such expertise allowed into the courtroom was that of handwriting identification. Thus, handwriting identification expertise is the oldest "forensic science," although it did not secure a place in the common law courtroom until nearly a century after its introduction into French proceedings.

When handwriting identification expertise finally entered the courtroom, it did so haltingly.[11] In 1792, Lord Kenyon, sitting as a trial judge, allowed two postal inspectors proffered by one of the parties to

---

9. 99 Eng. Rep. 589 (K.B. 1782). In the 18th century, the term "skilled witness" covered everyone we would today refer to as an expert. The term seems to have become recently used to refer to practical, as contrasted with scientific, experts. *See* Fed. R. Evid. 702 advisory committee's note.

10. The Trial of Philip Earl of Pembroke and Montgomery, at Westminster, for the Murder of Nathanael Cony (1678), *in* 6 Cobbett's Complete Collection of State Trials 1310 (1810). It appears that by the 1640s, surgeons were often, though by no means always, called upon to take part in coroner's inquests. Interestingly, the oldest surviving documentation of the practice currently known appears in the Maryland colonial records. *See* Helen Brock & Catherine Crawford, *Forensic Medicine in Early Colonial Maryland, 1633-83, in* M. Clark & C. Crawford, Legal Medicine in History (1994). In addition to the Earl of Pembroke's case, there were two other notorious 17th century trials in which expert testimony was given: The Trial of Robert Green, Henry Berry, and Lawrence Hill, at the Kings-Bench, for the murder of Sir Edmundbury Godfrey (1679), *in* 7 Cobbett's Complete Collection of State Trials 159 (1810), and The Trial of Spencer Cowper, Ellis Stephens, William Rogers, and John Marson, at Hertford Assizes, for the Murder of Mrs. Sarah Stout (1699), *in* 13 A Complete Collection of State Trials 1105 (T.B. Howell ed., 1812). The expertise involved in all three trials was medical expertise. *See generally* T.R. Forbes, Surgeons at the Bailey: English Forensic Medicine to 1878 (1985).

11. This story is told in more detail in Risinger et al., *supra* note 3, at 751-71.

testify concerning authorship by comparing known exemplars of one party's handwriting to a document whose authorship was at issue in the case.[12] However, the very next year Kenyon reversed himself and, in two cases, held such testimony inadmissible.[13] In the 1802 case of *Rex v. Cator*,[14] the court held that a postal inspector might give an opinion as to whether a signature was in a "feigned hand" by examination of the signature alone, but could not compare hands. This position was reaffirmed by a divided court in *Doe d. Mudd v. Suckermore*[15] in 1836, and it was not until an 1854 statute was construed to authorize such testimony that handwriting identification expertise became admissible in English courts.[16]

ADMISSIBILITY IN AMERICAN JURISDICTIONS

In the United States, the story of the admissibility of handwriting identification expertise is even more complex. Until the passage of the English statute, most American jurisdictions followed English practice and rejected such expertise. Some significant exceptions did occur, however. In the 1836 case of *Moody v. Rowell*,[17] Massachusetts became the first common law jurisdiction to authorize the use of such asserted expertise.[18] The rationale of the *Moody* case is telling. Up to that time, in all Anglo-American jurisdictions, handwriting had been formally authenticated as to authorship by the recognition testimony of nonexpert witnesses who were familiar with the putative author's handwriting.[19] This testimony was supplemented on occasion by direct jury comparison between challenged documents and other, authentic, writings of the putative author which had been offered into evidence for other purposes.[20] This was taken to be such weak evidence that, without evaluating the validity of the proffered experts' claims to expertise, the *Moody* court ruled that such asserted expert testimony should be admitted because it couldn't be any worse than what was traditionally relied on.[21] This seems to be the dominant rationale for allowing such testimony in those states which followed

---

12. Goodtitle d. Revett v. Braham, 100 Eng. Rep. 1139 (1792).

13. Stranger v. Searle, 170 Eng. Rep. 661 (1793); Carey v. Pitt, 170 Eng. Rep. 219 (1793).

14. 170 Eng. Rep. 661 (C.P. 1802).

15. 5 A. & E. 703 (K.B. 1836).

16. *See* Risinger et al., *supra* note 3, at 757-58, 757 n.116 (citing Common Law Procedure Act, 1854, 17 & 18 Vict., ch. 125, § 27 (Eng.)).

17. 34 Mass (17 Pick.) 490 (1835).

18. Louisiana's version of the Code Napoleon provided for resort to handwriting experts, but it is not clear that at the time of its statehood there were any such experts in Louisiana. *See* Risinger et al., *supra* note 3, at 761 n.133.

19. *See Moody*, 34 Mass. (17 Pick.) at 495-96.

20. *See id.* at 496.

21. This was the rationale urged by the dissenters in *Doe d. Mudd v. Suckermore*, 111 Eng. Rep. 1331 (K.B. 1836). The *Moody* court concluded that "this species of evidence, though generally very slight, and often wholly immaterial, is competent evidence." 34 Mass. (17 Pick.) at 498.

Massachusetts' lead over the next fifty to seventy-five years. Although a substantial majority of American jurisdictions had accepted such testimony by 1900,[22] the prevailing attitude may be best exemplified by the opinion of the New York Court of Appeals in *Hoag v. Wright*:[23]

> The opinions of experts upon handwriting, who testify from comparison only, are regarded by the courts as of uncertain value, because in so many cases where such evidence is received witnesses of equal honesty, intelligence and experience reach conclusions not only diametrically opposite, but always in favor of the party who called them.[24]

While some courts continued to reject such expertise, and most that allowed it remained skeptical, a group of professional experts was growing up and beginning to seek greater respectability. It is ironic that when expert handwriting identification testimony was first declared admissible in America and England, there were no experts. That is to say, the lawyers seeking to admit such testimony merely had to proffer various witnesses who were willing to assert a kind of *ad hoc* expertise acquired as a side effect of being something else, such as a postal inspector or a bank teller. No practicing forensic document examiner today would concede any expertise to such witnesses.[25]

When the legal system agreed to accept handwriting identification testimony, however, it created a demand which was met by people who increasingly turned their entire attention to filling it. Not surprisingly, these people soon set out to create a standard theory and practice, giving their trade the appearance of "science." Among the first of those people was Charles Chabot,[26] who, despite his name, was English. Originally a lithographer by trade, he developed an interest in handwriting identification about the time such expert testimony was gaining admissibility in English courts. It is unclear how much he was influenced by contemporary French theory and practice, but in 1871, at the urging of his lawyer-disciple Edward Twistleton (who wrote a lengthy theoretical introduction to the book), Chabot published *The Handwriting of Junius Professionally Investigated.* This was the first book in English to assert that there was a science of handwriting identification,[27] and to illustrate its methodology.[28]

---

22. The earliest authority for the admission of such testimony in each U.S. jurisdiction is set out chronologically in Appendix 3 to Risinger et al., *supra* note 3, at 788.

23. 66 N.E. 579 (1903).

24. *Id.* at 581. *See generally* Miles v. Loomis, 75 N.Y. 288 (1878); Mutual Benefit Life Ins. Co. v. Brown, 30 N.J. Eq. 193 (1878); *In re* Fuller's Estate, 70 A. 105 (Pa. 1908).

25. *See* Albert S. Osborn, Questioned Documents 286-87 (2d ed. 1929) [hereinafter Osborn 2d ed.].

26. Chabot and his contemporary Frederick G. Netherclift were the first full-time handwriting identification consultants in England, both beginning their practices in the mid-1850s after earlier careers in engraving and lithography. Two short reports by Netherclift on aspects of the Junius controversy appear in Chabot's book. Chabot, at any rate, was a significant enough character in mid-Victorian London to have rated an entry in the Dictionary of National Biography.

27. Twistleton refers to Chabot's work as a "scientific demonstration." Charles Chabot

Two American books on handwriting identification were published in the 1890s—William E. Hagan's *Disputed Handwriting* (1894)[29] and Daniel T. Ames's *Ames on Forgery* (1899).[30] But the event that was to transform handwriting identification expertise from ugly duckling to swan was the 1910 publication of Albert S. Osborn's *Questioned Documents*, with an introduction by John Henry Wigmore.[31]

Osborn's book, Osborn's personality, and Osborn's friendship with Wigmore were the cornerstones upon which respect was built for handwriting identification expertise in the United States.[32] Osborn set out the theory and practice of the claimed expertise so comprehensively that it is fair to say that all treatments of the subject since have simply been rearrangements or expansions of Osborn's 1910 book.[33] As to his personality, he was clearly a man of exceptional intelligence and critical

---

(and Edward Twistleton), The Handwriting of Junius Professionally Examined 220 (London, John Murray & Sons 1871).

28. *See id.* The subject matter to which Chabot applied his methods was the authorship of the anonymous "Junius" letters, famed in the political controversy of late 18th century England. Interestingly, Albert S. Osborn was a great admirer of the theoretical aspects of both Twistleton's and Chabot's writing in this book. *See* Albert S. Osborn, Questioned Documents 34-35 (1st ed. 1910) [hereinafter Osborn 1st ed.]; Osborn 2d ed., *supra* note 25, at 1000. However, Osborn disagreed with Chabot's conclusion that the Junius letters were written by Sir Philip Francis, asserting that Chabot and Twistleton had been misled by "improper standards" and "planted" documents. *Id.* (Osborn seemed to favor John Horne Tooke as Junius). After the publication of his book, Chabot's testimony in the Tichborne Claimant case (the O.J. Simpson case of Victorian England) brought him to the attention of the general public.

Note that while Chabot's book was the first book in English on the subject, it was not the first written source in English. In 1850, in Massachusetts, under the authority of *Moody v. Rowell*, Nathaniel D. Gould, a teacher of penmanship for 50 years, was called on behalf of the prosecution in the famous trial of Harvard professor Dr. John W. Webster for the murder of Dr. George Parkman. Since this trial was a sensation in its day, a verbatim transcript was made and published. In his preliminary testimony, Mr. Gould sets out the two basic principles of the field, *see infra* text accompanying note 96, which he claims to have derived from his own observation and reflection. *See* Report of the Trial of Prof. John W. Webster Indicted for the Murder of Dr. George Parkman 116 (James W. Stone rep., Boston, Phillips, Sampson & Co. 2d ed., rev. 1850).

29. W.E. Hagan, Disputed Handwriting (Albany, Banks & Bros. 1894). The first edition of Persifor Frazer's *Bibliotics* was published in the same year as A Manual of the Study of Documents (Philadelphia, J.B. Lippincott Co. 1894), but this book was to have no lasting influence since its main original theses were totally rejected by Albert Osborn and his followers. *See* Osborn 2d. ed., *supra* note 25, at 990.

30. Daniel T. Ames, Ames on Forgery (San Francisco, Bankcroft-Whitney 1899).

31. Osborn 1st ed., *supra* note 28.

32. It also did not hurt that the book received a glowing review from Roscoe Pound, another of the legal giants of the era, in the Harvard Law Review. Roscoe Pound, Book Review, 24 Harv. L. Rev. 413 (1910).

33. *See generally* materials collected in note 160 *infra*. This includes Osborn's own 1929 second edition, *supra* note 25, which had surprisingly little new information on handwriting identification theory or practice. Most of its material on those topics is taken verbatim from the 1910 edition, *supra* note 28. Note that the text reference is only to the orthodox nongraphological literature. For more on the graphological literature, see *infra* note 210.

abilities, but with a blind spot. He had a kind of mystical faith in the ability of the human mind to create a system of analytical expertise for the solution of virtually any class of problem. While he could be laudably skeptical regarding the claims of others,[34] he never seemed to notice that most of the generalities upon which he built his system lacked empirical verification.[35] Nevertheless, he had faith in his vision and his ability to sell others on that vision, whether the audience was a jury or a group of students, lawyers, or judges. His most significant convert was Wigmore, the most influential figure in evidence theory in the last century. Together, Osborn and Wigmore conducted a quarter-century public relations campaign on behalf of "scientific" handwriting identification expertise as practiced by Osborn and described in his book.

The ultimate triumph of this vision was finally insured by the Lindbergh baby kidnapping case, *State v. Hauptmann*, in 1935. Osborn was the chief witness called to testify that Bruno Richard Hauptmann had written all of the ransom notes found or sent after the abduction of the son of Charles A. Lindbergh. The public seemed to need to believe Hauptmann was guilty, wanted him convicted, and was grateful to those who supplied the evidence.[36] Osborn became a celebrity. For nearly sixty years after the affirmance of *State v. Hauptmann*,[37] no reported opinion rejected handwriting expertise, nor displayed much skepticism towards it. The testimony, which at the turn of the century was deemed of "uncertain value," became universally regarded as scientific and dependable. In 1977, a New York court noted the change: "Since that rather cynical observation was made by our highest court in *Hoag*, examiners of questioned documents, as handwriting experts prefer to be called, have attained more respectable standing in the courtroom."[38] As a New Jersey court observed in 1957,[39] after the *Hauptmann* case, handwriting identification expertise could no longer be regarded as "the lowest order of evidence, and . . . accorded little evidential weight."[40]

---

34. Notably, graphologists, who claim to be able to determine personality traits from handwriting. *See* Risinger et al., *supra* note 3, at 733 n.13 (quoting Osborn 2d ed., *supra* note 25, at 442-44).

35. *See infra* notes 101-11 and accompanying text (discussing the Forensic Sciences Foundation Proficiency Tests: 1975-1987).

36. *See generally* Ludovic Kennedy, The Airman and the Carpenter (1985); Anthony Scaduto, Scapegoat: The Lonesome Death of Bruno Hauptmann (1976); *but see* Jim Fisher, The Lindbergh Case (1987).

37. 180 A. 809 (N.J. 1935).

38. *In re* Estate of Sylvestri, 390 N.Y.S.2d 598 (N.Y. App. Div. 1977).

39. Morrone v. Morrone, 130 A.2d 396 (N.J. Super. Ct. App. Div. 1957).

40. *Id.* at 400. Osborn recognized the centrality of the *Hauptmann* case. Concerning it, he wrote in 1940, "[i]t can be correctly stated that in that little one hundred year old courtroom at Flemington, N.J., the scientific examination and proof of the facts in document cases was nationally recognized and firmly established as a New Profession." Albert S. Osborn, *A New Profession*, 24 J. Am. Judicature Soc'y 1 (1940), *reprinted in* Albert S. Osborn, Questioned Document Problems 311 (Albert D. Osborn ed., 2d ed. 1946). For a similar evaluation, see James V.P. Conway, Evidential Documents 210 (1959).

FRYE AND *DAUBERT*

For generations judicial thinking concerning the required dependability of expert testimony, especially that which might be labelled "scientific" testimony, was dominated by the so-called *Frye*[41] test.[42] The essence of the test was that testimony concerning scientific expertise was admissible only if the validity of the scientific principle or process upon which it was based had obtained general acceptance in the relevant scientific community.[43] Unfortunately, *Frye* itself contained no guidance on how to determine what constituted the relevant community to be looked to for acceptance. (This is hardly surprising, considering that the *Frye* opinion itself was less than a page and a half long.) In addition, *Frye* suggested that this test was required only for "novel" scientific evidence, without explaining why the same test was not appropriate for older claims and methodologies.

For nearly seventy years, judicial and academic exegetes made virtually whatever they wished out of the *Frye* test.[44] Some found it to be a formidable barrier to admissibility,[45] and others the most illusory of restrictions on the introduction of unvalidated and undependable "expertise." Critics of the *Frye* test attacked it from both sides, some saying its approach should be abandoned because it kept too much out,[46] and yet others saying that it should be abandoned because it let too much in.[47]

When Congress adopted the Federal Rules of Evidence in 1975, it made no reference to *Frye*, either in the language of Federal Rule of Evidence 702 (FRE 702) itself or in the short and particularly unhelpful advisory committee note. In the almost two decades that followed, courts and commentators variously construed FRE 702, the standard it implied, and the judge's role in enforcing it, including whether any of the many versions of "*Frye*" continued to play any proper role at all in federal court trials under the Federal Rules of Evidence.[48] Initially, proponents of the "let it all in" school of thought seemed to dominate under a broad

---

41. Frye v. United States, 293 F. 1013 (D.C. Cir. 1923).

42. Dominated, that is, to the extent that there was any thinking at all. Discussion concerning the issues dealt with by *Frye* began at a trickle and did not reach floodtide until the 1970s. *See* David L. Faigman, Elise Porter, & Michael J. Saks, *Check Your Crystal Ball at the Courthouse Door, Please: Exploring the Past, Understanding the Present and Worrying About the Future of Scientific Evidence*, 15 Cardozo L. Rev. 1799, 1808 (1994).

43. *See Frye*, 293 F. at 1014.

44. *See generally* Mark McCormick, *Scientific Evidence: Defining a New Approach to Admissibility*, 67 Iowa L. Rev. 879 (1982).

45. *See* Risinger et al., *supra* note 3, at 771 n.182 and authorities collected therein.

46. *See, e.g.*, United States v. Stifel, 433 F.2d 431, 438-39 (6th Cir. 1970).

47. *See, e.g.*, Michael J. Saks, *Expert Testimony Before the Bench*, 90 Tech. Rev. 42, 47 (1987). *See generally* Risinger et al., *supra* note 3, at 771-72 n.182, 779-80, and authorities there collected.

48. *See* Daubert v. Merrell Dow Pharms., Inc., 125 L. Ed.2d 469, 479-80, esp. n.5.

construction of what might "assist the trier of fact to understand the evidence."[49] But by the early 1990s, there was increasing sentiment that judges were admitting too much expert testimony of little or no dependability.[50] In 1991, this led the Judicial Conference Advisory Committee on Civil Rules to propose an amendment to FRE 702 which would have required that all expert testimony—not just "scientific" expertise—be subject to a preliminary finding by the judge that it was "reasonably reliable."[51] While this rule was never promulgated, it formed the immediate background for the Supreme Court's decision in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*.[52]

Daubert dealt with the admissibility of expert testimony that Merrell Dow's prescription anti-nausea drug Bendectin could cause birth defects.[53] The Ninth Circuit had held the testimony inadmissible under FRE 702, applying a construction which incorporated one version of *Frye's* "general acceptance" standard.[54] Justice Blackmun, writing for the Court, rejected the Ninth Circuit's absolute *sine qua non* requirement of any version of the "general acceptance" test, since FRE 702 contained no such requirement.[55] In discussing appropriate standards of admissibility under FRE 702, the Court limited itself to a consideration of "scientific" evidence, explicitly declining to make any comments on applicable standards for "technical and other specialized knowledge."[56] Unfortunately, it is less than clear how to tell one from the other. The opinion states that the term "scientific" "implies a grounding in the methods and procedures of science."[57] We are also told that "[s]cience is not an encyclopedic body of knowledge about the universe. Instead, it represents a *process* for proposing and refining theoretical explanations about the world that are subject to further testing and refinement"[58] and further that "in order to qualify as

---

49. Fed. R. Evid. 702; *see* In re Air Crash Disaster at New Orleans, La., 795 F.2d. 1230, 1234 (5th Cir. 1986).

50. *See* David E. Bernstein, *Junk Science in the U.S. and the Commonwealth*, 21 Yale J. Int'l L. 123, 133-34 (1996).

51. 1991 Report of the Advisory to Committee on Civil Rules (discussing change of FRE 702).

52. 509 U.S. 579 (1993).

53. *Id.* at 582-85.

54. *See id.* at 584 (citing Daubert v. Merrell Dow Pharmaceuticals, Inc., 951 F.2d 1128 (9th Cir. 1991)).

55. *Id.* at 588.

56. *Id.* at 590 n.8.

57. *Daubert*, 509 U.S. at 590. However, presumably all expert testimony of any kind is so grounded, at least in the sense that it ought not to generate statements about the factual world inconsistent with those yielded by the "methods and procedures of science." This problem is compounded when the Court immediately quotes a definition of science from Webster's 3rd New International Dictionary 1252 (1986) which, in part, is as broad as the 17th century definition (equally applicable to such things as "political science" and "moral science"): "any body of ideas . . . accepted as truths on good grounds."

58. *Id.* (quoting the Brief for the American Association for the Advancement of Science and the National Academy of Sciences as Amici Curiae at 7-8) (emphasis added).

'scientific' evidence, an inference or assertion must be derived by the scientific method."[59] Later, in a conflation of the definition of science per se with notions of validity, the Court states that a key question concerning a theory or technique is "whether it can be (and has been) tested. 'Scientific methodology today is based on generating hypotheses and testing them to see if they can be falsified; indeed, this methodology is what distinguishes science from other fields of human endeavor.'"[60] Finally, the Court states, "The statements constituting a scientific explanation must be capable of empirical test,"[61] and finally, "[t]he criterion of scientific status of a theory is its falsifiability, or refutability, or testability."[62]

Once a court has determined that proffered testimony deals with "scientific" knowledge and that it is relevant to the case in its scientific aspects (which the *Daubert* opinion refers to as its "fit"[63]), it then must determine if the evidence is sufficiently "reliable" to be admitted.[64] In making that decision, there is no specific litmus test.[65] However, the Court indicated that the chief considerations are the degree to which the testable claims of the evidence have been subjected to attempts at falsification and survived, the publication of such data, the degree of professional evaluation of claims and tests, and, in regard to techniques or processes, whether there is data on the error rate, and what such data reveals concerning likely error rates under varying conditions.[66] Finally, and importantly for this Article, footnote eleven makes clear that the requirements of FRE 702 set out in *Daubert* apply to all scientific evidence, not just to "novel" evidence.

STARZECPYZEL

The first consideration of the admissibility of handwriting identification after *Daubert* occurred in the Southern District of New York in *United States v. Starzecpyzel*.[67] Roberta and Eileen Starzecpyzel were charged with having stolen various works of art from Roberta's elderly (and now senile) aunt. They claimed that the paintings were a gift made

---

59. *Id.* at 590.

60. *Id.* at 593 (quoting Michael D. Green, *Expert Witnesses and Sufficiency of Evidence in Toxic Substances Litigation: The Legacy of* Agent Orange *and Bendectin Litigation*, 86 Nw. U. L. Rev. 643, 645 (1992)).

61. *Id.* (citing Karl R. Popper, Conjectures and Refutations: The Growth of Scientific Knowledge 37 (5th ed. 1989)).

62. *Daubert*, 509 U.S. at 593 (citing Popper, *supra* note 61, at 37).

63. *Id.* at 591.

64. *Id.* at 593-96.

65. *See id.* at 593 ("Many factors bear on the inquiry, and we do not presume to set out a definitive checklist . . . ."); *id* at 594-95 ("The inquiry envisioned by Rule 702 is, we emphasize, a flexible one. Its overarching subject is the scientific validity—and thus the evidentiary relevance and reliability—of the principles that underlie a proposed submission.").

66. *See id.* at 593-94.

67. 880 F. Supp. 1027 (S.D.N.Y. 1995) (McKenna, J.).

prior to the aunt's impairment. Part of the evidence against them was the conclusion of a questioned document examiner that the aunt's signatures on deeds of gift were forgeries. Relying on *Daubert* and on the doubts raised by Risinger et al., their attorneys moved for a hearing to determine the admissibility of the proffered testimony under FRE 702.

Judge McKenna's opinion examines the claims of handwriting identification expertise to scientific status at length and rejects them.[68] Having done this, however, Judge McKenna concludes that since such experts are not practicing a science within the meaning of *Daubert*, *Daubert's* validation requirements therefore do not apply. He then analogizes such a proffered expert to a harbor pilot who learns to do something dependably by experience. As to whether the prosecution's expert would be allowed to testify to his conclusion that the signatures on the documents, which were the subject of the prosecution, were forgeries, based on his examination of the numerous genuine signatures of the putative victim, the court says that the defense had

> presented no evidence, beyond the bald assertions of [its experts], that FDEs [forensic document examiners] cannot reliably perform this task. Defendants have simply challenged the FDE community to prove that this task can be done reliably. Such a demonstration of proof, which may be appropriate for a scientific expert witness, has never been imposed on "skilled" experts.[69]

---

68. The court stated:

> Were the Court to apply *Daubert* to the proffered FDE testimony, it would have to be excluded. This conclusion derives from a straightforward analysis of the suggested *Daubert* factors—testability and known error rate, peer review and publication, and general acceptance—in light of the evidence adduced at the *Daubert* hearing.

Id. at 1036.

69. *Id.* at 1046. The implication that the ultimate risk of nonpersuasion as to reliability is *ever* on the opponent of a proffer of evidence is startling, in light of Fed. R. Evid. 104 and the general notion that the party seeking admission must convince the court affirmatively of admissibility once the question is seriously put in issue. Actually, Judge McKenna seems to have been aware of the problems that would be created by formally placing the burden of persuasion on the opponent of a proffer. The actual position taken by his opinion on that issue is ambiguous and unclear and, one must conclude, intentionally so. In the only explicit discussion of the issue, he concedes that in her chapter in the Federal Judicial Center's Reference Manual on Scientific Evidence (1995), Professor Berger takes the (standard) position that the burden is on the proponent of admissibility. *Starzecpyzel*, 880 F. Supp. at 1031. He then cites an unexamined single line claim to the contrary from the middle of an article by a products liability practitioner whose main position is that *Daubert's* effect should be viewed as allowing more things to be admitted, not fewer. *Id.* (citing Arvin Maskin, *The Impact of* Daubert *on the Admissibility of Scientific Evidence: The Supreme Court Catches Up with a Decade of Jurisprudence*, 15 Cardozo L. Rev. 1929, 1936 (1994)). However, Judge McKenna attempts not to choose between these two positions, characterizing the question before him as "legal" rather than factual, as if that made the problem go away. *Starzecpyzel*, 880 F. Supp. at 1031. His later language from the passage quoted in the text above seems to show his functional adoption of the problematical Maskin position, even though, as the text indicates, he goes on to say that he is affirmatively persuaded that handwriting identification testimony "can be performed" with "sufficient reliability to merit admission." *Id.* at 1046.

Judge McKenna then declared himself persuaded that the inferences as to genuineness of the signature at issue in the case before him "can be performed with sufficient reliability to merit admission."[70]

VELASQUEZ AND *RUTH*

Before the end of 1995, two more published opinions, *United States v. Velasquez*[71] and *United States v. Ruth*,[72] addressed the issue of whether handwriting identification testimony is admissible under FRE 702. *Velasquez* is somewhat more problematic than *Starzecpyzel* because it is more ambiguous. The case presented two issues. The first issue was the propriety of the trial court's admission of the testimony of a questioned document examiner that mailing labels used to ship drugs had been written at least partly by each of two alleged accomplices of defendant Velasquez.[73] This testimony was the *only* evidence establishing the participation of five persons (rather than three) in the charged events and was thus central to the conviction of Velasquez on a charge of engaging in a continuing criminal enterprise of five persons or more in violation of 21 U.S.C. § 848.[74] The second issue was the propriety of the trial court's exclusion of the defendant's expert, Mark P. Denbeaux, whose testimony was offered to educate the jury concerning the lack of published data supporting the accuracy of handwriting identification by questioned document examiners.[75] The Third Circuit held that it was not error to admit the handwriting identification testimony, but that it was error to refuse to admit the counter-testimony and granted a new trial.[76] Only the reasoning for the first ruling concerns us here.

The *Velasquez* court never clearly explained whether it was judging the admissibility of handwriting identification expertise under the *Daubert* criteria for scientific evidence or under some other different standard for "technical or other specialized knowledge."[77] The court cites *Starzecpyzel* as authority for the proposition that the criteria for judging admission of "scientific expertise set out in *Daubert* are inapplicable to non-scientific handwriting identification."[78] However, the court then goes on to say that "in an exercise of caution," it would review the proffered expertise "for

---

70. *Starzecpyzel*, 880 F. Supp. at 1046. Judge McKenna went on to fashion a jury instruction to be given in advance of the expert's testimony to explain that the testimony was not the result of a scientific process, so that the jurors would have no misconceptions in that regard. *Id.* at 1051 (Appendix I).

71. 64 F.3d 844 (3d Cir. 1995).

72. 42 M.J. 730 (1995).

73. *See Velasquez*, 64 F.3d at 846-47.

74. *See id.* at 847 n.5 (discussing Velasquez's conviction under 21 U.S.C. § 848 (1982)).

75. *See id.* at 846.

76. *See id.* at 846-47.

77. *See id.* at 850.

78. *See Velasquez*, 64 F.3d at 850.

qualifications, reliability and fitness as those factors have been explicated in *Daubert*."[79] But the court then proceeds never even to mention the validity criteria for scientific evidence actually set out in *Daubert*. Rather, the court reviews the paper credentials of the government's forensic document examiner and then quotes from the document examiner's trial testimony outlining the standard methodology of orthodox Osbornian handwriting identification.[80] The court then follows that quote with the statement, "we agree with the district court that Ms. Bonjour's proposed testimony concerned 'scientific, technical or other specialized testimony' and was sufficiently reliable to be admissible."[81] This is at best an implied *Starzecpyzel* analysis and not by any means a *Daubert* scientific validity analysis.

Finally, in the *Ruth* case, which involved a different and more difficult handwriting identification task than *Starzecpyzel*, the Court of Military Appeals merely cites *Starzecpyzel* as dispositive, without undertaking any dependability analysis regarding the task there involved.[82]

Thus, the *Starzecpyzel* approach may be the wave of the future, not just in regard to handwriting, but in regard to many areas of forensic "science" that upon examination are found to contain little science. The problem is that once a court deprives an area of the label "science," the validation standards for admissibility may plummet almost to nonexistence. As to any merely "skilled" witness, the burden appears to be on the opponent to prove affirmatively that the skilled witness cannot do what they claim they can do. The shoal upon which this analysis founders in the case of handwriting identification is Judge McKenna's analogy to harbor piloting. Harbor pilots (and many other experts in areas of practical expertise with their main applications in the world outside the courtroom)[83] undergo validation tests every day. They have unmistakable criteria for judging their success and failure in practice, which are apparent not only to themselves, but to those around them. They either show up safely at the right dock or they do not. If the harbor pilot had no feedback of objectively right and wrong results, he would be unlikely to develop any dependable expertise.

Unlike harbor pilots, practitioners of many purely forensic expertises, such as handwriting identification, commonly have no way of knowing whether they have come to the right conclusions in their actual cases. They may know if a jury agreed with them, but neither they, nor anyone else, have a way of checking to see if they were actually right or wrong. It may be true that, in many cases, there is strong independent circumstantial

---

79. *Id.*

80. *See id.* at 850-51.

81. *Id.* at 851.

82. United States v. Ruth, 42 M.J. 730, 732 (1995).

83. Prior to the *Frye* decision, the main practical validity criterion used by judges in deciding the admissibility of proffered expertise seems to have been the existence of a significant market for practical extra-judicial application of the expertise. *See* Faigman et al., *supra* note 42, at 1804.

evidence corroborating the accuracy of their conclusions, but unfortunately this context information is often given to them with the materials submitted for evaluation, which makes it hard to know if their conclusion was based upon their expertise or merely on the suggestion supplied at the outset by the context material. Clearly, as Professor Imwinkelreid has observed in urging courts to develop rational validity standards for nonscientific evidence, "the epistemology of nonscientific expert knowledge is quite different from that of scientific propositions . . . . [T]he development of objective validation standards for nonscientific opinion is likely to prove to be a more difficult task than the formulation of such tests for scientific testimony."[84] While Judge McKenna's approach might be justifiable for a harbor pilot's testimony, in a purely forensic area such as handwriting identification, some higher standard of affirmative proof would seem to be needed to insure that the conclusions proffered can be arrived at dependably.

FUTURE DIRECTIONS IN THE LAW

As of this writing, forensic handwriting identification testimony is admissible in every American jurisdiction, although those federal courts that have examined the question after *Daubert* have found it necessary to shift the grounds for admissibility from its being a science to its being a nonscience. The necessity for either excluding handwriting identification expertise or finding a new basis for admitting it is most clearly analyzed in *U.S. v. Starzecpyzel*, but Judge McKenna's decision to require the opponent of admission to affirmatively prove unreliability seems highly questionable as a matter of sound evidence policy.

Starzecpyzel may be invoked unthinkingly by other courts as a matter of convenience, as in *Ruth* and *Velasquez*, but this would be a mistake even on *Starzecpyzel*'s terms. The *only* opinion that Judge McKenna ruled was sufficiently dependable to be admitted concerned inferring that a challenged signature is not genuine from examination of other genuine signatures.[85] Judge McKenna left different and perhaps more difficult determinations explicitly unaddressed.[86] However, though the underlying issues are not about to go away, significant exclusion of handwriting identification testimony by any court seems unlikely in the near future, perhaps more because of inertia than analysis. It is fair to expect, however,

---

84. Edward J. Imwinkelreid, *The Next Step After* Daubert*: Developing a Similarly Epistemological Approach to Ensuring the Reliability of Nonscientific Expert Testimony*, 15 Cardozo L. Rev. 2271, 2294 (1994).

85. United States v. Starzecpyzel, 880 F. Supp. 1027, 1043 (S.D.N.Y. 1995). This determination is, by the standard theory under which questioned document examiners operate, one of the more generally easy and dependable determinations. *See infra* note 225 and accompanying text.

86. *Starzecpyzel*, 880 F. Supp. at 1043. Judge McKenna specifically included among such more difficult and questionable skills positive identification of an author from limited amounts of writing (like the address labels in *Velasquez*).

that if this testimony continues to be allowed, courts will at least have to develop and enforce standards designed to maximize likely dependability and eliminate the grossly suggestive methods by which handwriting identification problems sometimes are presented to examiners for evaluation.[87] Whatever the future course, if courts are to decide intelligently on the appropriate judicial treatment of handwriting identification testimony, they must learn what forensic document examiners' claims to expertise consist of, what evidence supports or contradicts those claims, and the areas in which evidence is lacking. To that end, we offer the following.

THE NATURE AND DEPENDABILITY OF HANDWRITING IDENTIFICATION

Handwriting identification experts believe they can examine a specimen of adult handwriting and determine whether the author of that specimen is the same person or a different person than the author of any other example of handwriting, if both specimens are of sufficient quantity and not separated by years or the intervention of degenerative disease.[88] The experts further believe that they can accomplish this result with great accuracy and that they are far better than an average literate person attempting the same task.[89] They assert that they can obtain these accurate findings as the result of applying an analytical methodology to the examination of handwriting, according to certain principles which are reflected in the questioned document literature.[90] They believe that this literature explains how to examine handwriting for identifying characteristics, and that by applying these lessons in connection with their experiences in various training exercises and real world problems, they learn to identify handwriting dependably. This section of the Article will examine the justifications for these beliefs to see if any evidence exists to support them.

---

87. *See* Risinger et al., *supra* note 3, at 773-77; *see also infra* note 159 and accompanying text. An attempt at insulation from suggestive context material seems to be routine in all basic research and at least some other forensic disciplines. D. Ubelaker & H. Scammell, Bones: A Forensic Detective's Casebook 36 (1992). In fact, Ubelaker's and Scammell's anecdotes reveal that specimens are often submitted with biasing context information, and identify "being influenced by someone else's expectations" as one of the three greatest dangers "in forensic anthropology—and perhaps in any other forensic science." *Id.* at 279. *See also* Larry Miller, *Procedural Bias in Forensic Science Examinations of Human Hair*, 11 L. & Hum. Behav. 157 (1987).

One final worry: If the *Starzecpyzel* approach placing the burden of proving undependability on the opponent of admission becomes generally accepted, why would any document examiner ever cooperate in any future empirical testing program? They would seem, individually and as a group, to have little to gain and much to lose.

88. *See infra* note 160 and accompanying text.

89. *See id.*

90. *See* Osborn 2d ed., *supra* note 25, at 6 (referring to these principles as "true methods"); David Ellen, The Scientific Examination of Documents: Methods and Techniques 9 (1989) (referring to these principles as "standard methods" and "proper method").

THE POSSIBILITY OF A SCIENCE OF HANDWRITING IDENTIFICATION

The main goal of all forensic identification, including handwriting identification, is individuation. Individuation is the establishment that a given person or object is the same person or object associated with a past event in a particular way, to the exclusion of all other candidates. The major source of individuation evidence is what we may call a tagged residue.[91] A tagged residue exists when a person or thing leaves behind some residue of its presence at a relevant time and place, which contains information that can be used to conclude (with varying certitude) that a particular person or thing produced the residue. In addition, even when individuation is not confidently possible, exclusion may occur based on a decision that a particular source did not produce the tagged residue.

The notion of tagged residues is broader and covers more phenomena than one might at first glance conclude. For instance, eyewitness identification is a special example of the use of tagged residue information. In that case, the residue is not specifically physical, but exists as a memory in the mind of the identifier. Nevertheless, whether we are dealing with mental images, photographs, physical impressions, or traces of bodily fluid, all tagged residue situations present some common characteristics and problems. The hope is that the information in the residue may be processed in such a way that one can properly conclude that one and only one object or person could have caused the residue. This hope is in one sense doomed, since all information about such factual relations is probabilistic. Thus, as science has known at least since Hume, for any identification whatsoever there is some residual probability of error. On the other hand, under some conditions, the probability of particular identification may be so great that it would be nearly deranged to worry about the residual probability of error.

The question, therefore, is what are the main circumstances that affect, or ought to affect either our confidence in particularized identification or exclusion of a person or object as the source of a tagged residue.

The main factors appear to be the following: First, what characteristics of the residue are relevant to specific identification (individuation)? This question inevitably entails, consciously or unconsciously, some notion of separability[92] of characteristics and some notion of base rate incidences of

---

91. The term "tagged residue" to describe this class of phenomena is a neologism. No functionally similar term exists in the forensic science literature.

92. In regard to handwriting, Albert S. Osborn was much more willing to assume total independence of characteristics than some of his successors have been. *Compare* Osborn 2d ed., *supra* note 25, at 229-31 (arguing that handwriting characteristics are totally independent), *with* Wilson R. Harrison, Suspect Documents: Their Scientific Examination 305-07(1958) *and* Ordway Hilton, Scientific Examination of Questioned Documents 9 n.11 (Revised ed., Elsevier Science Publishing Co. 1982, reprinted by CRC Press, Inc, Boca Raton, 1993) (arguing that you must consider all identifying characteristics as a whole). However, all seem confident that in practice there is sufficient individuality to distinguish any two adults'

those characteristics in the population of candidates for the source of the residue. Second, referring to the particular person or object that in fact caused the residue as the source, there is the problem of potential intrasource variation. Residues from the same source may differ each time the source leaves a residue. Further, referring to the people or objects that might have caused the residue as candidates, we have the problem of intercandidate similarity. More than one object may be capable of causing a residue indistinguishable from the residue in question in one or all dimensions. Failure to accurately separate important from unimportant characteristics, accurately assess dependence, and reflect accurate notions of base rate within the candidate population, whether conscious, analytic, and quantified, or unconscious, impressionistic, and unquantified, will obviously lead to error, though the fact that the conclusion is in error may not be obvious. Failure to properly deal with the effects of intrasource variation or intercandidate similarity is a further important potential source of error.

These problems are present in all tagged residue situations, though they may be most troublesome in the area of handwriting identification. It is obvious that each time a person writes, the individual letters are not formed with mechanical similarity to previously made letters. That creates a protean problem of intrasource variation. Less obvious is that some writers may write so much alike that their writing cannot be confidently distinguished, at least with limited samples.[93] This renders intercandidate similarity a significant problem.

It seems fair to say that when intrasource variation, or intercandidate similarity, or both, rise to a substantial level, dependable particularized identification (individuation) becomes rationally impossible.

In principle, there is nothing in the nature of a tagged residue problem which prevents a science of tagged residue individuation from being developed. For a source of asserted factual knowledge to qualify as scientific in the central modern sense, it must be the product of an enterprise displaying certain characteristics. Chief among these are the following:

1. A systematic encouragement for gathering and publishing reproducible sense observations.
2. A taxonomy for organizing such observations that lends itself to dependable reproducibility of observation, and to quantification.
3. A process of hypothesis generation that results in statements about the world and its interrelationships that are consistent with all known observations and that are potentially amenable to falsification through empirical observation.
4. An established regime that attempts to falsify new hypotheses

---

handwriting dependably if the samples presented are large enough. With questioned writings, of course, in contrast to known exemplars, the writing sample often consists of very little.

93. *See* John J. Harris, *How Much Do People Write Alike? A Study of Signatures*, 48 J. Crim. L. & Criminology 647 (1958).

empirically (and that is rewarded for doing so).

Moreover, although it is not a logical *sine qua non* of science, there is virtually universal recognition that academic institutions will play a significant role in the practice of science and the training of scientists in nearly every area.[94]

Some tagged residue individuation processes, such as DNA typing, are sciences, because they are the application of knowledge gained through the process of science. Indeed, there might someday be a science of handwriting identification, and some small first steps have been made in that direction.[95] However, as Judge McKenna concluded in *Starzecpyzel*, forensic document examination is not a science. There is nothing in the enterprise that results in or encourages organized reporting and publication of observations in reproducible form. There is no agreed taxonomy of sufficient refinement to yield dependably quantified data, or dependably comparable observations of any refinement. There has been no theoretical revision of any significance in nearly a century, and there is no professional encouragement or reward for attempts to falsify those theories that exist.

---

94. By contrast, handwriting identification has no academic base. Training is by apprenticeship, and there is no standardization of training enforced either by any licensing agency or by professional tradition. Nor is there a single accepted professional certifying body. The Encyclopedia of Associations lists five different organizations: the National Association of Document Examiners; the National Bureau of Document Examiners; the World Association of Document Examiners; the Independent Association of Questioned Document Examiners; and the American Society of Questioned Document Examiners. At least three of these organizations claim to grant certifications of competency. A glance at the credentials listed by document examiners advertising in various directories aimed at lawyers will quickly reveal at least four or five more such organizations. The American Society of Questioned Document Examiners, which one might characterize as the organization of the Osbornian establishment, is generally most vocal and aggressive in claiming its membership to be of highest quality. However, the certification testing program of its child, the American Board of Forensic Document Examiners (co-sponsored by the American Academy of Forensic Sciences) described by Mary Wenderoth Kelly (a member of that certifying board) in her *Starzecpyzel* testimony, leaves a lot to be desired. While there is a short multiple choice test to measure knowledge of handwriting identification doctrine, the heart of the examination is based on the administration of five of only seven or eight different test problems, only two or three of which involve handwriting identification. The same problems are used year after year on an honor system where they are sent to the candidates for certification through their teaching mentor, and left with them for a month unsupervised before the answers are returned. *Starzecpyzel Daubert* hearing transcript of testimony, 2/28/95 (Kelly direct) at 48-59; Kelly cross, at 175-92; 3/1/95 (Kelly cross), at 249-60 (on file with the *Iowa Law Review*).

95. *See, e.g.*, United States v. Starzecpyzel, 880 F. Supp. 1027, 1027 n.7 (S.D.N.Y. 1995) (noting research currently in progress on handwriting identification); Risinger et al., *supra* note 3, at 739 n.31 (listing sources that discuss several avenues of academic research pertaining to handwriting identification).

THE CLAIMED PRINCIPLES OF HANDWRITING IDENTIFICATION

The foundational principles of the expertise, as they were characterized by Mary Wenderoth Kelly in her testimony in *Starzecpyzel*, are that no two people write exactly alike, and that no one person writes the same word exactly the same way twice.[96] It is most important to note that these two general principles are in their strong form nonscience metaphysical statements (though the statements have an oddly commonsensical appeal). On some level, no two things can be *exactly* alike, and it is this nonempirical intuition that underlies both statements. However, science is not concerned with metaphysical differences, but with *perceivable* differences and similarities that can be used to accurately assess common origin. When so recast, the statement that no two people write so alike that the differences are imperceptible is not intuitively obvious, nor is the claim that no one person ever writes the same word or set of words so similarly that the differences are imperceptible. These forms are subject to potential testing by scientific methods, but they have been subjected to virtually no testing, and the small amount of data available do not provide much support for them. These were among the main reasons why Judge McKenna declared handwriting identification expertise to be nonscience in *Starzecpyzel*.[97]

What occurs in practice is an example of the same kind of clinical empiric that also underlies eyewitness identification, and many other everyday processes. Such a process is at root probabilistically derived, just as is formal science, but with two important differences. First, it is not based on standardized measurements of any precision. Second, the database of examples that defines which characteristics are common and which are unusual is not public, recorded for all who will take the time to see and evaluate. Rather, it is private, based on the experiences of the individual practitioner over a long period of time, and stored internally in such a way that many or most of the individual data may be beyond conscious recall. The problem with such a process is that it is only as good as the unexaminable personal database of the practitioner, and the practitioner's not-fully-explainable method of deriving answers to such problems as, in the case of handwriting identification, what constitutes significant intrawriter variation[98] and interwriter similarity. The practitioner's opinion may be given the appearance of an explanation by pointing out similarities (or differences) between the questioned document

96. Quoted by Judge McKenna in *Starzecpyzel*, 880 F. Supp at 1032. At the time of her testimony, Ms. Kelly was a document examiner with the Cleveland Police Department with 13 years of experience. She was a director of the American Society of Questioned Document Examiners and a Fellow in the Questioned Document Section of the American Academy of Forensic Sciences. She is certified by the American Board of Forensic Document Examiners and at the time of her testimony served on that board as Vice President and Chair of the Committee on Testing. *Id.* at 1027.

97. 880 F. Supp. at 1033-34, 1038.

98. Or "natural variation" as it is usually called in the trade. *See id.* at 1032.

and the exemplars, and by assertions that such characteristics are, in the practitioner's experience, common or uncommon. Arguably, however, any opinion of common authorship will have some similarities to support it, and any differences can be assigned to intrawriter variation.

Handwriting identification theory, as employed by questioned document examiners, does have some process principles and some general rules which are at least not counter-intuitive and which, if followed, restrict somewhat what the document examiner can say. In the Appendix, we have set out to the best of our ability a fair summary of the main outlines of those asserted principles and that process in regard to its two most common general applications, signatures and anonymous writings.[99]

TESTING DOCUMENT EXAMINER DEPENDABILITY

Nothing in the principles of handwriting identification theory, presented in the Appendix, is mystical or visibly implausible but that does not a science make. As sensible as much of it seems, none of the assertions are self-validating, and all are amenable to formal empirical testing which has never been undertaken. Even the assertion that an atomistic analysis leads to better results is not inevitable, as any baseball player whose swing was injured by a batting coach could tell you. And the notion that a learned technique of such analysis, with so many elements of subjective judgement, in an area where clear evidence of accurate conclusions is often not available, will enable all or most people completing such training to give accurate conclusions on both easy and difficult problems, is dubious.

This is not to say that all identification by comparison of hands is necessarily inaccurate. Common experience, once again, confirms that under some circumstances we can recognize a writer by handwriting. There is a tagged signal present, at least sometimes, but how specific or dependably perceived it is in various circumstances is subject to debate. It may be that document examiners, or some of them, pick up a not-fully-analyzable knack of accurate identification of handwriting just by being exposed to a lot of it. It may be that the atomized analysis called for by their method of practice improves the accuracy of some or all of them a little or a lot.

Because such a practical expertise does not have the internal validation of a developed science, it requires the external validation given to an instrument, a black box process, that may or may not lead to dependable results. Here is where science can again come into play, because science can examine the dependability of the results of such a process even when the process is not science. Beyond their own

99. No effort has been made to deal comprehensively with every area of handwriting examination doctrine. For instance, there is no specific treatment of disguise, juvenile handwriting, etc. The purpose of the text is to give the reader unfamiliar with orthodox handwriting examination doctrine a fair sample of its main positions and methodologies.

assertions,[100] however, what evidence, if any, exists to show that document examiners can accurately identify or exclude authorship by comparison of hands, or do so better than the average person?

The answer to that turns out to be almost none. In their 1989 *University of Pennsylvania Law Review* article, Risinger et al. did a full literature search[101] and turned up only one published study bearing on

---

100. We must distinguish among two types of evidence derived from "their own assertions." The first type is simply self faith: "We know we're good because we believe we are," or "we know we're accurate because we follow the revealed true method in reaching our results." The second type is anecdotal: "In such and such a case, I identified the right person, or was right concerning the forged nature of the document." These anecdotes can further be divided into three types: First: "I know I was right because other evidence in the case established the facts independently." In these cases, the examiner almost always, in today's practice, may have known these facts before the examination, so that the handwriting conclusions may be more the result of revelation to the examiner of other facts than the examiner's independent analysis. The opinions in the Lindbergh case are all subject to this suspicion, for example. Second: "The trier of fact agreed with me." Obviously, there may be ego gratification in this, but this factor alone does not establish objective accuracy. Third: "After my opinion, the opponent confessed, or pleaded guilty, or otherwise admitted I was right." The problem with these anecdotes is that guilty pleas do not very clearly prove factual accuracy, and that the confession or admission cases are usually built on other evidence, which tends to make them overlap with the first type of anecdote. All of these validation criteria are weak proxies for the real thing—a gold standard of certain knowledge of authorship against which to compare the examiners' conclusions—such as harbor pilots have in their actual practice. The real risk is that the examiners themselves will make more out of the proxy feedback and self-affirmation than it is rationally worth. That is not to say that there is no anecdotal evidence of value. There are well known cases of miscarriage by document examiners whose existence contradicts at least the more extravagant claims concerning the dependability of the expertise. *See, e.g.*, David Fisher, Hard Evidence 196 (1995) (stating that FBI document section chief Ronald Furgerson has said that all "180" "certified" document examiners in the United States would reach the same conclusions in any given case as he would); *see also* Stephen Fay et al., Hoax: The Inside Story of the Howard Hughes-Clifford Irving Affair 129-33 (1972) (discussing Albert S. Osborn's grandson Russell Osborn's mistaken authentication of Clifford Irving's Howard Hughes forgery); Kenneth W. Rendell, Forging History: The Detection of Fake Letters and Documents 106-23 (1994) (discussing the Hitler diaries hoax and other mistakes of document examiners). The data summarized later in the Article from the Forensic Sciences Foundation's proficiency studies, which show high rates of disagreement among document examiners, raise grave doubts about the validity of such claims. On the other side, there are equally remarkable successes which suggest that under the right conditions and in the best hands, handwriting comparison can do remarkable things. Evidence of these successes is the proof of Michele Sindona's 1981 return to the U.S through the writing on one customs card among thousands in an assumed name (confirmed by a fingerprint on the card), and the 1956 capture of Joseph LaMarca, the Peter Weinberger kidnapper, by identification of his handwriting from among millions of public records documents. In both these cases, there was a very striking and easily recognized peculiarity in the handwriting. *See* Fisher, *supra*, at 196-98.

101. In their recently published article, *The Principle of the "Drunkard's Search" as a Proxy for Scientific Analysis: The Misuse of Handwriting Test Data in a Law Journal Article*, 1 Int'l J. Forensic Document Examiners 7 (1995) [hereinafter Galbraith et al.], Oliver Galbraith III, Craig S. Galbraith, and Nanette G. Galbraith criticize the thoroughness of that search, and their title is even taken from the punch line to an old joke meant to ridicule the search as having looked where it was convenient to look rather than where the answer lay. Ironically, Risinger et al.

the question, plus five unpublished studies carried out by the Forensic Sciences Foundation (FSF). The only published study, Inbau's *Lay Witness Identification of Testimony*,[102] had such a small number of participants and was so methodologically flawed, that it yielded no significant data on any issue.

The results of the FSF studies may or may not have yielded significant data, depending on one's perspective. The studies were designed as proficiency tests for government crime laboratories. The tests were taken only by a voluntarily self-selected group of such laboratories who, for each test, decided to request the test materials and to return their results. The necessity of having comparable test materials administered to each participating laboratory meant that the test materials had to be photocopies rather than original documents, and the test takers all knew that they were taking a test rather than working on a real world project. The FSF itself officially took the position that the results were not necessarily representative of the actual level of performance in the field. However, the FSF could not have regarded the tests as so problematical as to be meaningless, or they would not have continued to administer them.

Beyond these considerations, there is the ever-present problem of test design itself. Designing meaningful studies to test the validity and reliability of a diagnostic process like handwriting identification is not as easy as it might at first appear. As we have seen, a process such as handwriting identification presents a number of potential subtasks dealing with variables such as writing instruments, forgery of various sorts, age, health, and so forth. No single test can map the abilities of any one practitioner, or any group of practitioners. A great many tests (certainly more than have yet been designed and administered) would be necessary to know what, if anything, they can do accurately, and under what conditions. A complete testing regime would have tests which covered the entire spectrum of conditions and difficulties. In addition, the law not only cares about the likely accuracy of results by putative experts, but also about whether the results obtained and testified to by experts lead to more accurate conclusions than would result if the jurors did the comparisons directly without such testimony. "Expertise" exists only if there is a significant accuracy advantage of the putative expert over the average juror. This is especially important in an area of nonscientific subjective judgment. Thus, the tests should ideally be administered to control groups of ordinary people to see if any such accuracy advantage exists.

With these caveats in mind, the following is a summary of the tests

---

took the unusual step of describing in some detail their literature search strategy, and it was a thoroughgoing one. The easiest way to prove that Risinger et al. overlooked important research, of course, would be simply to come forward with it. Despite years of grousing, however, neither the Galbraiths nor anyone else has yet cited a single empirical study or other test data bearing on the issue of handwriting examiner accuracy which existed at the time Risinger et al. was published and which was not addressed in that article.

102. Fred E. Inbau, *Lay Witness Identification of Handwriting*, 34 U. Ill. L. Rev. 433 (1939).

and results available to Risinger et. al. through 1987 as the FSF reported them and as they appeared in that article.[103]

THE FORENSIC SCIENCES FOUNDATION PROFICIENCY TESTS: 1975-1987

The 1975 Test

In 1975, participating laboratories were given a letter made up of both handwriting and typewriting. In addition, they received four examples of handwriting written by four different people. The problem was to determine whether the handwriting on the questioned document was written by any of the four "suspects." In reality, the questioned letter had been written by one of the suspects. The following are the results of the seventy-four laboratories that responded:

66 (89%) correctly identified the writer of the questioned letter.
1 (1%) gave a partially correct and partially incorrect answer (attributing part of the writing in the questioned letter to the right "suspect" but another part to another "wrong" suspect).
4 (5%) asserted that they could not reach any conclusion from the materials supplied them.
3 (4%) identified the wrong person.

The 1984 Test

In 1984, participating labs were supplied with three handwritten letters containing bomb threats, supposedly received by the news media and then followed by terrorist bombings. The labs were also given two pages of known handwriting samples for each of six suspects (a total of twelve pages). They were to determine if the three questioned letters had all been written by the same person, and whether any or all of the questioned letters had been written by any of the suspects. Two of the threat letters were in fact written by one person, who was not among the suspects and whose actual known writing was not given to the labs. The other threat letter was written by one of the suspects whose exemplars were in his normal hand, but who had tried to simulate the writing of the other two threat letters when producing *his* threat letter.

Forty-one labs requested the test materials but only twenty-three submitted answers. The following are the results of those answers:

17 (74%) perceived the difference in authorship of letter 3.
6 (26)% said erroneously that the same person wrote all three threat letters.
23 (100%) failed to recognize that letter 3 was written by one of the suspects for whom they had known writings.[104]

---

103. In Risinger et al., *supra* note 3, there are extensive footnotes to specific pages of the FSF reports for virtually each line of text. These have been omitted and, further, no such notes have been inserted in regard to the more recent FSF studies. The FSF reports are short, and not easily obtainable. If the reader does not obtain them, page references are obviously useless; if the reader manages to obtain them, page references are unnecessary.

104. Judging from comments at a recent annual meeting of the Document Examination

The 1985 Test

Participating laboratories were given twelve checks all having signatures in the same name. They were asked to decide which, if any, of the signatures were made by the same person. In fact, two of the twelve had been signed by the real person whose name appeared on them. Of the remaining ten, one was an attempted freehand forgery by a person without known experience as a forger; another was a tracing. The remaining eight were signed by eight different people in their own normal handwriting. Forty-two labs requested test materials and only thirty-two returned them. The results are as follows:

13 (41%) gave correct results.
2 (6%) wrongly attributed one of the forgeries to the real signatory.
10 (31%) reported that they were unable to reach conclusions.
7 (22%) were substantially wrong, making errors beyond a single misattribution of authorship.

Section of the American Academy of Forensic Sciences (Feb., 1996, Nashville, TN), this question and the finding of 100% error is the single greatest object of complaint by questioned document examiners concerning the FSF tests, a complaint that has some validity, but not as much as they assert. The usual form of their objection appears to be something like, "we don't claim to be able to determine the authorship of a writing made while trying to imitate some other person's writing." (Such an attempt is called "simulation" or "simulated forgery" in normal document examiner parlance.) First, as the authorities cited in the Appendix demonstrate, the orthodox Osbornian position is not that such attribution of authorship is categorically impossible, but merely that it is difficult, because attempting to imitate someone else's writing is an effective way to disguise one's own writing, and may in some or many cases be so effective at suppressing individuality that identification is impossible. *See* Ordway Hilton, *Can the Forger be Identified from his Handwriting?*, 43 J. Crim. L. & Criminology 547 (1952) and *infra* notes 227-29 (Appendix). The real underlying complaint concerning question 2 in the 1984 FSF test seems to be that the skill of the simulator was unusual, so that even though the writing was in a fairly generous amount, the simulator managed to suppress all identifying characteristics. Although this confirms empirically the possibility of false negatives already admitted by Osbornian theory in such cases, this level of skill is rare enough that the meaning of the universal failure of examiners on this question is easy to overstate, especially in aggregating it with the results of other tests or questions. (It should be noted that there is no specific evidence the simulator in fact was unusually skilled. The argument is circular: since we couldn't identify the writer, he or she must have been exceptionally skilled.)

It is true that all aggregation strategies applied to the FSF data are inherently flawed, since no one knows exactly how to measure either the ease or difficulty of a test task, or its statistical incidence in normal practice. However, it appears that there have been more questions in the FSF studies as a whole from what was assumed to be the easy end of the spectrum of their work, so that the inclusion of the occasional hard task in an aggregation seems not as artificial as excluding it completely. Beyond aggregation objections, there can be no objection to asking the question and observing the result, even if it only tends to affirm empirically what orthodox Osbornian theory holds. In addition, it is at least a start to defining empirically the outer limits of whatever expertise, if any, actually exists.

The 1986 Test

The 1986 test involved handprinting. Participating labs were told to assume that police had stopped a car with three known occupants. In the car they found a hand-printed holdup note and other evidence linking the note to a holdup apparently committed by only one person. The labs were given a copy of the holdup note. They were also given a copy of handprinting exemplars from the three occupants of the car. Suspect 1 had actually printed the holdup note. Suspect 2 had not, Suspect 3 had not printed the holdup note either, but he was a document examiner whose handprinted exemplar was an attempt to simulate the printing on the holdup note. Forty-eight labs requested materials and thirty-one returned reports. The results of those reports are as follows:

4 (13%) gave correct answers.
3 (9%) said that none of the authors of the exemplars had written the hold-up note.
10 (32%) were unable to reach any conclusions.
14 (45%) assigned authorship to the forger.

The 1987 Test

Because of document examiner complaints concerning the difficulty of prior tests, the FSF decided to make the 1987 test easy. As its report of results said, "This test was designed to be a relatively easy and straightforward test, because of complaints about previous test design. All the writings in this test were natural and free of disguise." In this test, participating labs were given a copy of a handwritten extortion note. Exemplars of the handwriting of four persons were also supplied, one of whom actually had written the extortion note. The problem was to determine which, if any, of the "suspects" had written the extortion note. Fifty-five laboratories requested materials and thirty-three responded with reports. The results are as follows:

17 (52%) correctly identified the writer of the extortion note.
1 (3%) incorrectly eliminated the correct suspect, asserting that none of the suspects wrote the extortion note.
0 (0%) incorrectly identified an innocent person as the author.
15 (45%) responded that their results were inconclusive.

Risinger et al., in evaluating these results, stated:

> What do all five FSF studies taken together suggest? A rather generous reading of the data would be that in 45% of the reports forensic document examiners reached the correct finding, in 36% they erred partially or completely, and in 19% they were unable to draw a conclusion. If we assume that inconclusive examinations do not wind up as testimony in court, and omit the inconclusive reports, and remain as generous as possible within the bounds of reason, then the most we can conclude is this: Document examiners were correct 57% of the time and incorrect 43% of the time.
>
> But let us turn to more meaningful readings of the aggregate

data. The pilot test in 1975 may have been unrealistically easy, like a line-up with four beefy white policemen and a skinny black person. Did this task present any real difficulty at all? There is no way of knowing whether a group of lay persons would have done any less well, since none was tested. Omitting the 1975 data, the examiners were correct 36% of the time, incorrect 42%, and unable to reach a conclusion[] 22% of the time. Even these results are biased in favor of accuracy because of the intentional ease of the 1987 test. Disguised handwriting fooled them all and forged printing fooled two-thirds of those who hazarded an opinion about it.

Now consider the effect on the aggregate results of the laboratories that requested test materials but did not return them. More likely than not, these non-respondents bias the results further in favor of correct conclusions. Some of the non-responding labs, no doubt, did not even perform the tests due to the press of daily business. But some others very likely performed the tests and then did not return their reports. Assuming that an examiner who has worked on an answer and then decides not to return it has serious doubts about its accuracy, then the sample of respondents is composed of an unrepresentatively large proportion of those who obtained—or at least think they obtained—correct answers.

If a correct answer consists of a report containing correct conclusions returned pursuant to requested and submitted test materials, then of the total submissions to laboratories in the 1984 through 1987 tests, only 18% gave wholly accurate responses (without the 1987 test the figure drops to 13%).

Finally, consider the possible effect on any aggregate conclusions of the fact that, of the more than 250 [crime] laboratories that perform handwriting examination (not to mention a large number of private practitioner document examiners), only a fraction even ordered test materials in the first place. It is at least arguable that, by self-selection, the sample is inherently biased in favor of the more conscientious and capable practitioners to begin with. If this is true, the reported results would overstate the accuracy of the handwriting examination field generally.

The 1984, 1985, and 1986 tests presented examiners with a variety of challenges. The results should provide anyone with cause for concern. The examiners who returned reports on the analysis disagreed among themselves a good deal of the time, suggesting limited reliability, and many of the opinions offered were incorrect, suggesting limited validity.

In addition, the studies failed to reveal that certification or experience enhanced accuracy. The 1987 Proficiency Advisory Committee Comments state that "[a]s usual, there were no correlations between right/wrong answers and certification, experience, amount of time devoted to document examination and length of time spent on this test." Consider what this independence means for a court's likely assumptions about

> whether to admit a proffered expert and for the weight a factfinder is expected to give such testimony. A court is likely to assume that an examiner who is certified, who has been on the job for many years, whose caseload is nothing but document examination, and who has spent a lot of time examining the evidence, is especially likely to have something useful to say to a jury. Yet these data provide no support for these assumptions. Examiners who are uncertified, have little experience, work on document examination only part time, and spend little time on the particular document, are just as likely to be right as someone with more impressive qualifications. Does any of this suggest the existence of expertise?
>
> These are the sorts of findings about the nature and limits of asserted handwriting identification expertise about which both document examiners and the courts need to know but which could not have been known before such studies were undertaken. Though they are not without flaws, these studies represent a step toward systematic and scientific evaluation of the claimed capabilities of this asserted expertise. Perhaps some of the considerably larger number of needed tests yet to be designed and administered would show document examiners faring better, but on the present record we must say that the underpinnings of the "expertise" have degenerated from no data to negative data.
>
> Finally, we cannot emphasize too strongly that from the viewpoint of the law each of these studies suffers from a major omission: the absence of a control or comparison group of lay test-takers. If a jury can compare handwriting no worse than proffered "experts," then the expertise does not exist. For any given task, the level of performance of professional document examiners may be no better than that of laypersons. Indeed, lay persons might perform some tasks consistently better than "experts." While such superiority may seem intuitively improbable, it remains a logical possibility and one not without analogues in other areas. For now, the kindest statement we can make is that no available evidence demonstrates the existence of handwriting identification expertise.[105]

Since the publication of Risinger et al., only four further studies appear to have been undertaken. In 1988 and 1989, the FSF again administered handwriting identification proficiency tests to document examiners at crime laboratories.[106] The results of the 1988 and 1989 FSF

---

105. Risinger et al., *supra* note 3, at 747-51 (citations omitted) (second alteration in original).

106. In 1990, after the existence of Risinger et al., *supra* note 3, became widely known, FSF quit testing on handwriting identification and began testing on such topics as rubber stamp identification (1990) and photocopying machine identification (1991). There was arguably a "handwriting" element in the 1992 test. However, comparison of form had little to do with obtaining the correct results. One could determine that the signature on the top (white) purchase order copy (copy 1) was photocopier generated and not hand signed, by direct

studies have never been published,[107] but summary reports containing the results were issued to participating laboratories by the FSF, and those findings will be published below. In addition, an article by Oliver Galbraith III, Craig S. Galbraith, and Nanette G. Galbraith,[108] reporting the results of administration of the 1987 FSF test materials to nonexperts, appeared in the inaugural issue of a publication called *The International Journal of Forensic Document Examiners.*[109]

Finally, Moshe Kam, Joseph Wetstein, and Robert Conn published an article in 1994, entitled *Proficiency of Professional Document Examiners in Writer Identification,*[110] based on research commissioned by the FBI. We will consider the Galbraith piece first, insofar as it criticized Risinger et al. and will then consider the 1988[111] and 1989 FSF studies, the Galbraiths' original research, and finally the Kam et al. research.

---

examination without reference to exemplars (if appropriately skilled). The "carbonless carbon" on the yellow copy signature could then be determined to be a tracing of the photocopy signature by observing the inkless stylus indentation on the photocopied signature of the white copy, and the exact correspondence between the stylus indentation and the carbonless carbon signature. The determination that a particular one of the exemplars provided had been the signature from which the photocopy on the white purchase order had been generated might be called a "comparison of form" problem, but since perfect superimposition was the key to this determination, it is not a comparison of the kind under consideration in this Article. About three quarters of the ninety respondents identified the right exemplar as the source of the photocopy, but the other quarter simply said nothing, and since they were not explicitly asked, these nonresponses cannot be counted as errors, given the fact that most of them correctly identified the signature on copy one as a photocopy and that on copy two as a tracing over of the photocopy, which resolved the issue of genuineness about which they *were* asked.

107. They were, however, printed and distributed to the heads of participating laboratories and others, and were thus "semi-published."

108. *Supra* note 101. Nanette G. Galbraith is a forensic document examiner in private practice in California. Oliver Galbraith III, Ph.D., was, at the time this piece was published, Professor of Information and Decision Sciences at San Diego State University College of Business. Craig S. Galbraith, Ph.D., was an Associate Professor at the Krannert Graduate School of Management, Purdue University. Their exact relationship to Nanette Galbraith is not given.

109. *Id.* A sort of samizdat xerox form of this article appeared first in 1989, which seems to have had wide hand-to-hand distribution among document examiners at meetings and conventions. In its 1995 published form, the Galbraith article is substantially identical with the 1990 xerox version on file with the editors, except that the number of non-experts to which the 1987 FSF test was administered was expanded from thirty-two in the 1990 version to sixty-five in the 1995 version.

110. Moshe Kam et al., *Proficiency of Professional Document Examiners in Writer Identification,* 39 J. Forensic Sci. 5 (1994).

111. Although the Galbraiths apparently had access to the results of the 1988 FSF study when they originally wrote their piece, since it is reflected in Table 1 of the 1990 xerox version on the validity and reliability problems of the FSF tests, *supra* note 109, they did not deal with it more explicitly either in the text of the 1990 version or in the 1995 published version, nor did they deal with the 1989 results in the 1995 version.

THE GALBRAITHS' CRITIQUE

The Galbraiths' explicit criticisms of Risinger et al. can fairly be summarized as follows:

> 1. All the FSF studies are methodologically so flawed that they cannot be used as a proper basis for drawing substantive conclusions.[112] Risinger et al. failed to recognize this and as a result of that failure they drew inappropriate conclusions.[113]
> 2. Even if the data were valuable, the way Risinger et al. analyzed the data was wrong.[114]
> 3. A proper examination of the data (which the Galbraiths claim to do in reexamination) would result in reclassification of many document examiner responses to the tests from incorrect to correct and make their performance as a whole appear better than Risinger et al. reported.[115]

We will examine each of these criticisms in turn.

The Galbraiths begin their criticisms by invoking "the four types of validity issues identified by Cook and Campbell" in their well respected book *Quasi-Experimentation: Design & Analysis Issues for Field Settings.*[116] Unfortunately, while there are important criticisms to be made of the methodology of the FSF studies, the Cook and Campbell framework adopted by the Galbraiths is largely inapposite. The Galbraiths confused cause-effect studies (experiments and quasi-experiments) with research aimed more simply at measuring some skill or ability. Cook and Campbell make clear that their validity constructs address problems of inferring that a treatment (independent variable) caused the observed effects in a dependent variable.[117] The FSF tests were simply trying to measure accuracy of performance (much like testing how well marksmen can hit targets). They were not testing which of two or more treatment conditions produced better performance (such as testing which of two training methods produced more accurate marksmen). Cook and Campbell's book is concerned with the methodological problems of the latter. The FSF studies are of the former kind. The Galbraiths struggled to apply cause-effect methodological issues to research that aimed instead to measure a

---

112. Galbraith et al., *supra* note 101, at 9-11.

113. *Id.* at 8, 16.

114. *Id.* at 11.

115. *Id.*

116. Thomas D. Cook & Donald T. Campbell, Quasi-Experimentation: Design and Analysis Issues for Field Settings 137-85 (1979) (discussing the four types of validity: statistical conclusion validity, internal validity, construct validity, and external validity).

117. *See generally id.* This is evident throughout the book. But consider the following specifics: The second sentence of the preface: "The designs serve to probe causal hypotheses about a variety of substantive issues in both basic and applied research." *Id.* at ix; the title of the first chapter: "Causal Inference and the Language of Experimentation." *Id.* at 1; the word causation appears in 6 out of 10 section headings within the first chapter. *Id.* at v; the first sentence of the first chapter: "The major purpose of this book is to outline the experimental approach to causal research in field settings." *Id.* at 1. The point is hard to miss.

single variable, test performance, and which made no attempt to draw causal inferences because there were no independent variables. A far more apt research tradition to inform a critique of the FSF studies would have been the literature of psychometrics (*i.e.*, testing). Moreover, virtually all of the methodological objections raised by the Galbraiths were, in fact, dealt with by Risinger et al., either in the text or in footnotes. In the few instances where this is not true, it is generally because the Galbraiths' objections are inapposite. Their fundamental misreading of Cook and Campbell leads the Galbraiths repeatedly into confusion.

For instance, the Galbraiths assert that one ought to discount or disregard the results of question 2 on the 1984 test (which all examiners got wrong) because of "insufficient co-variation."[118] The Galbraiths assert that "in designing or evaluating *any* test or experiment (such as the FSF tests) one must make sure that there is, or will be, sufficient covariation in the data, that is, variation in the test results must be observed in order to relate the results to the issue under investigation."[119] This is simply untrue in the universal form in which it is expressed; it depends entirely upon what issue is under investigation. While this assertion holds true for *cause-effect* studies (which Cook and Campbell were discussing), it does not for more basic skill *measurement* studies (which we, the FSF, and the Galbraiths are discussing). If all we want to know is whether most human beings can hold their breath for ten seconds, the fact that a test administered to one hundred humans results in *all* participants holding their breath for ten seconds in no way undermines the validity of the result, statistically or in any other way. It is true, as Risinger et al. discussed at length,[120] that tests that are too easy or too hard can lead to results which may be misinterpreted. It is also true that a single test designed to discriminate levels of skill has been unsuccessful if everyone does equally well or equally poorly. However, as Risinger et al. point out, handwriting identification is not a single unitary operation, but rather presents "a broad variety of circumstances and tasks. Tests must be designed carefully to present discriminations of meaningful difficulty and variety. Only results from such tests could begin to paint a picture of what both lay people and experts can and cannot do . . . ."[121] Within the context of such a testing regime, it is not a criticism that some particular question or subtest was so hard that none of the experts could do it. By themselves such data are not ipso facto meaningless, as the Galbraiths seem to claim concerning question 2 on the 1984 test. In combination with other data, such results can help determine the limits of the expertise which were not known before the data were developed.[122]

---

118. Galbraith et al., *supra* note 101, at 9, 14.

119. *Id.* at 9 (emphasis added).

120. *See* Risinger et al., *supra* note 3, at 742-43.

121. *Id.* at 742. Ironically, the Galbraiths' own footnote 14 makes the same point in very similar terms. Galbraith et al., *supra* note 101, at 10 n.14.

122. The Galbraiths also dispute the propriety of any cogitation on how non-responses

Methodological misdirections aside, the Galbraiths' main critical thrust is that Risinger et al. erred by accepting the response classifications that had been given to answers in the FSF studies initially *by the responding document examiners themselves* and then *by the FSF in its own summary of results.* Specifically, they claim that many responses which the document examiners labelled "inconclusive" when examining what turned out to be a true author's writings, ought to be treated as correct answers because the explanatory remarks accompanying the answers can be taken to indicate some level of belief that the writer of the exemplar may have written the questioned writing. They assert that these responses should be treated as examples of "qualified opinion" rather than bet-hedging "inconclusives."[123]

Even assuming the validity of their classifications of the data, however, the results do not mean what they go on to claim. The Galbraiths claim to have newly discovered 18 additional truly correct answers out of 192 total responses.[124] However, all but one of these answers newly classified as correct occurred in response to the two clearly easiest tests, 1975 and 1987.[125] Indeed, 13 of the new corrects were on the 1987 test alone, changing the performance on that test from 52% correct by the reckoning of Risinger et al. (based on the FSF classification) to 91% correct (based on the Galbraith reclassification). Second, the Galbraiths then exclude all respondents who actually returned the test, but criticized the test materials and checked off "inconclusive." These, they argue, ought to be treated as nonresponses,[126] and, they argue even more vigorously, no conclusion can ever be fairly made from a nonresponse.[127] They then exclude the bad results of question 2 from 1984, which they disapprove of for the weak "lack of co-variation" reasons discussed above, that is, all responses were wrong.[128] The Galbraiths then aggregate all the results even though the easy 1975 test accounted for nearly two and one-half times the number of responses of the next most responded-to test, and more than three times

---

may have resulted in sample biasing which overstated the skills of document examiners as a whole. They rightly point out that the biasing results of self selection might plausibly have run the other way. The main disagreement between Risinger et al. on the one hand and the Galbraiths on the other seems really to be about who should bear the burden of persuasion and the risk of non-persuasion. The stance of Risinger et al. is clearly that validity is to be treated as unproven until there are sufficient unambiguous data supporting it. The position of the Galbraiths is that one should not doubt validity and "indict the whole field" without affirmative proof of invalidity. The Galbraiths' position appears close to that adopted by Judge McKenna in *United States v. Starzecpyzel.* And it is the opposite of the law's usual view that the burden of proof is on the proponent of the evidence, as well as the view of science that the burden of persuasion is on the one making the claim that some phenomenon exists.

123. Interestingly, the Galbraiths show no examples of moving "inconclusives" to the totally *wrong* column.

124. *See* Galbraith et al., *supra* note 101, at 13.

125. *See id.* at 10.

126. *See id.*

127. *See id.*

128. *See id.* at 11.

the responses of the harder 1984 test. Based on these adjustments, they claim that document examiners were correct 75% of the time on the FSF tests.[129]

The fact remains, however, as even the Galbraiths must concede, that the examiners did well at some tasks and poorly on others. Their own analysis (reproduced as Table 1 below) shows that, in two of six tasks with data sufficient to conduct significance tests, the document examiners could not even exceed *chance* accuracy.

Table 1

Comparison of Expertise Against Chance

(Correct Versus Incorrect)

Exam Year (Question)	Observed Proportion Correct (M1)	Observed Proportion Correct (M2)	Chance Proportion Correct	Exact Probability (M1)	Exact Probability (M2)	Conclusion
1975	0.9429 (66/70)	0.9469 (70/74)	0.2000	0.0001	0.0001	Experts outperform chance
1984 Q1	0.5652 (13/23)	0.5652 (13/23)	0.2500	0.0174	0.0174	Experts outperform chance
1984 Q2	0.0000 (0/23)	0.0000 (0/23)	0.1429	0.0575	0.0575	Experts no different than chance
1985	0.5900 (13/22)	0.5517 (16/29)	0.0002	0.0001	0.0001	Experts outperform chance
1986	0.1905 (4/21)	0.2500 (7/28)	0.2500	0.2652	0.5000	Experts no different than chance
1987	0.9444 (17/18)	0.9375 (30/32)	0.2000	0.0001	0.0001	Experts outperform chance

Source: Adapted from Galbraith et al., *supra* note 101, Table 3

Think about that finding. Is *chance* the criterion of expertise? If a driver manages to stay on the right side of the median stripe more often than chance, if a piano student hits the correct notes more often than chance, if a student scores above chance on an exam—are they to be regarded as "experts"? And the Galbraiths found that document examiners could not even perform at that level on one-third of the test tasks they analyzed. Put simply, beating chance hardly establishes expertise. Even by the law's generous definition, to have expertise one must be able to outperform nonexpert average jurors.[130] On that issue, even the Galbraiths concede that no statistically meaningful data existed in 1989, when Risinger et al. was published.[131] In an attempt to remedy this

129. *See* Galbraith et al., *supra* note 101, at 11.

130. *See* Bernstein, *supra* note 50, at 128 n.21. Also, even if it could be shown that a proffered expert outperforms jurors, if both do terribly, with lay persons being right one time in a thousand and experts one time in a hundred, such expertise still may not be dependable enough for admission.

131. "[T]here certainly has been a shortage of studies comparing handwriting identification expertise with non-expertise . . . ." Galbraith et al., *supra* note 101, at 7 n.7; *see also id.* at 7 (stating that there is an "admittedly sparse history of carefully controlled empirical

absence of data, the Galbraiths administered the 1987 FSF test to two groups of nonexperts to see how their performance compared to the document examiners who had been tested earlier. We will analyze those results below, but first we will complete the review of FSF proficiency test data by reporting the results of the 1988 and 1989 tests.

FORENSIC SCIENCES FOUNDATION PROFICIENCY TESTS: 1988-1989[132]

The 1988 Test

The written instructions of the 1988 FSF proficiency test in handwriting identification told the labs to assume the following case scenario: Four complaints were received from four separate physicians' offices about shipments of narcotics that were not received. The delivery service produced four receipts (Q1-Q4) each containing a signature in the name of a secretary for one each of the four physicians. Each secretary denied writing her own signature or any of the others. (The same driver had apparently made all the deliveries.) Handwriting samples from the four secretaries (labelled K1-K4) and the driver for the delivery service (K5) were requested. The delivery service also sent along two other receipts bearing signatures of "unknown persons."

The labs were supplied with six receipts acknowledging the receipt of goods. Receipt Q1 bore the signature "Sharon D. Clayborne," but in fact was written by Richard D. Osbourn, the driver. Receipt Q2 bore the signature "Lisa D. Bridgeforth" and in fact was signed by Lisa D. Bridgeforth. Receipt Q3 bore the signature "Cynthia Y. Boone," but in fact was written by Richard D. Osbourn, the driver. Receipt Q4 bore the signature "Joanna Neuman" and in fact was signed by Joanna Neuman. Receipt Q5 bore the signature "Linda N. Ninestine" and was not written by any of the five people who provided exemplars. Finally, receipt Q6 was signed "Linda D. Wentworth," but in fact was written by Richard D. Osbourn, the driver.

Sharon D. Clayborne, Lisa Bridgeforth, Cynthia Y. Boone, and Joanna Neuman each provided exemplars in which they signed their own names a number of times and also signed all the other names appearing on the receipts a number of times. Richard D. Osbourn gave exemplars in which he signed all the names appearing on the receipts a number of times. These exemplars were provided to the labs through photographs.

These materials were submitted to seventy-three labs and returned by forty-nine. Before examining the results of these tests, it is necessary to say something about the test design as reflected in the FSF report. The report fails to explicitly set out crucial information concerning the test design

---

studies").

132. We have presented the test reports (and the critique) in detail because the research base is so remarkably small that the whole of it can be presented within the confines of a single article, and also because courts considering what to do with the claims of handwriting expertise are likely to pay minute attention to the few studies that do exist.

necessary to evaluate the difficulty of the task presented to the takers of the test. Fair inference from the report, however, indicates that there were in fact people named Lisa D. Bridgeforth and Joanna Neuman who signed the receipts bearing those names and signed them in their normal signature hands. In addition, it appears that these two women gave their exemplars of their own names in their normal signature hands. (If these things are not as stated in the text, the test is one of the worst that could be designed in terms of the standard internal precepts of the asserted expertise. It is a virtual postulate that signatures are special as to design, speed of execution, uniformity, etc. If supposedly authentic questioned signatures, and especially if assertedly known exemplars, are signatures by people for whom the writing is not a habitual signature, but represent someone else's name, the whole exercise would be, in the internal terms of the discipline, grossly and unfairly misleading to the test takers.) If anything appears to correspond to reality in this field, it is that signatures are usually special manifestations of handwriting. It seems reasonable to believe that as a result of repetition, and psychological factors creating a personal stylistic identification with one's own signature, normal signatures are the most individual and uniform parts of a person's writing. This is not to say that signatures cannot be disguised. Nor is it to say that normal signatures of two people with the exact same name cannot evolve into confusingly similar forms. It is merely to say that it would presumably be rare for the normal signature of one person to look at all like another dissimilarly named person signing that name either in the second person's usual handwriting or in a disguised hand where no real signatures of person number one were available to imitate. Thus, it would appear that perfect scores on Q2 and Q4 were, or ought to have been, giveaways.

In addition, the examiners being tested knew which of the sets of exemplars were from each secretary, and which set of exemplars was from the delivery man Osbourn. If they assumed that the original Clayborne and Boone signatures were natural signatures if genuine (as they had a right to do) and if they further assumed or concluded that the real Clayborne and Boone exemplars bore natural Clayborne and Boone signatures (as they had a right to do), then it would become a simple task to eliminate all the secretarial exemplars as candidates for signing the Clayborne and Boone signatures, since none of the exemplar signatures would be likely to resemble the signatures on the receipts in any arguably significant way. This would reduce the question in regard to Q1 and Q3 to "did Osbourn sign these or not?" The difficulty of that question is presumably somewhat dependent on the nature of the writing presented (which was not reproduced in the report), but the results might also very well be influenced by the fact that Osbourn was the only candidate for those signatures left in the pool, and exterior circumstances already cast suspicion on him. Thus, his identification as the author may have resulted from the design of the test, not from the information derived from his handwriting.

Finally, the test takers might assume (as one respondent explicitly

did) that the secretaries were unlikely to be in a position to sign receipts for deliveries to other doctors, which tends to convert the question in Q5 and Q6 to a question of whether or not the "Linda Ninestine" and "Linda D. Wentworth" signatures were signed by Osbourn or not. The elimination of Osbourn as the signer of the "Linda Ninestine" signature was likely to be made easier by the fact that presumably the signature was signed by a real Linda Ninestine in her real signature hand, and the Richard Osbourn exemplars would not only fail to resemble it, the examiners would already have concluded what Richard Osbourn's attempts at fake signatures looked like from Q1 and Q3. This would also assist them in identifying Osbourn as the author of the "Linda D. Wentworth" signature.

On the other hand, they might decide that since Osbourn signed all the other receipts not signed by known secretaries, he was likely to have signed Q5 also. All this raises the question of bias resulting from the presentation of unnecessary context information. Why not just present the exemplars marked "questioned" and "known," and ask who wrote what, if anything?

For these reasons, the 1988 FSF test must be approached with a substantial grain of salt. Each identification is not an independent event. Instead, the test presented two connected sets of issues, one set relatively easy and one set somewhat more difficult. The easy set of issues effectively asked: Which, if any, of these secretaries signed a receipt bearing her name in her normal signature hand? The harder set effectively asked: Of the four signatures left over after disposing of question one, which, if any, was signed by Richard Boone?

Not surprisingly, out of forty-eight responding labs,[133] all got each response to Q1 substantially right,[134] all but three got each response to Q2 substantially right,[135] and all but one got the answers to Q3 right. (The exception eliminated Osbourn as the author of Q3.) A forty-ninth lab was unable to reach any conclusions on any of the queries, and marked everything "inconclusive."

Apparently, Joanne Neuman has some significant intrawriter, or "natural" variation in her signature, because the responses to Q4 were surprising. Out of forty-eight responses, twenty-two were right, but five were wrong in affirmatively excluding the real Joanne Neuman. Twenty-one gave various versions of "inconclusive," ranging from seven leaning toward

---

133. One lab, 517, counted as responding by FSF, responded "inconclusive" to every part of every question, and in its narrative report protested the structure of the test. While we do not believe that inconclusives are categorically meaningless, under these conditions they are and we have eliminated this lab from consideration.

134. That is, with one exception, all answered either "did not write" or "probably did not write" to each of the secretaries and "did write" or "probably wrote" to the driver Osbourn. The one exception was lab 539, which said Osbourn probably wrote Q1, but responded "inconclusive" to all the secretaries.

135. Labs 522, 539, and 542 failed to identify the true author of the signature, answering "inconclusive" to the question.

Neuman to three leaning toward another writer. Our inclination in dealing with these is to throw out the inconclusive responses and to say that in this particular case, nearly a fifth of examiners with definite opinions were wrong. Clearly, even signatures are no guarantee of absolutely easy problems.

Up to this point the errors had been false negatives, assertions that the true writer did not write the questioned signature. Q5 presents the more troubling problem (for the legal system, especially in a criminal context) of false positives. This time, only seventeen labs were right,[136] but even among this group, the use of "probably did not" instead of "did not" rose in comparison to the answers to the previous questions. Only three labs manifested confidence that Osbourn did not write Q5. On the other hand, three labs affirmatively indicated Osbourn *did* write Q5,[137] and one lab said Sharon D. Clayborne wrote Q5. Twenty-seven labs responded "inconclusive" to the Q5 questions, and in general these were unqualified inconclusives (not leaners in the Galbraith sense). Thus false positives made up 19% of the affirmative responses and 25% of the responses with a confident finding.

Finally, as to Q6, there were forty-two correct responses, two inconclusives that leaned toward Osbourn, three inconclusives, and one wrong exclusion of Osbourn.

Thus, one can say that on the easy problems (Q1-Q3), document examiners gave 144 responses, of which 140 (97.2%) were correct or substantially correct, 1 (1/2%) was an affirmative false exclusion, and 3 (2.5%) were inconclusive. Including Q6 as an easy problem, there were 192 responses, of which 181 (94.3%) were correct, 2 (1%) were affirmative false exclusions, and 9 (4.7%) were inconclusive. On the harder problems (Q4 and Q5), however, document examiners gave 96 responses, of which 39 (40.6%) were correct or substantially correct, 9 (9.4%) were incorrect and 48 (50%) were inconclusive. Of the 48 answers reflecting firm conclusions on the harder problems, 19% were wrong, and of the 96 responses, only 41% were affirmatively right. The examiners did very well (but not perfectly) on the easy questions and substantially less well on the more difficult questions. And, as usual, there was no administration of the test to a control group of nonexperts in an attempt to find out the relative performance advantage, if any, of experts over nonexperts.

---

136. One lab was counted as right even though it had one inconclusive instead of five exclusions.

137. Labs 516, 522, and 545. Lab 507 responded by saying that Osbourn wrote the signature also, but the accompanying narrative comments were inconsistent with an affirmative identification of Osbourn. Although this response was counted as a misidentification by the FSF, it seems to have been a typographical error, and we have not counted it. Lab 545 was counted as indicating Osbourn, even though it amended the answer from "probably" to "possibly."

The 1989 Test

If document examiners find some comfort in the 1988 results, at least as to the easier tasks, the 1989 results should be disconcerting. The test was designed to see how well examiners deal with adolescent handwriting. The participating labs were asked to assume that during the final week of the school year, a high school teacher found the tires on her car slashed in the school parking lot. A handwritten note was found on her windshield. The teacher reported the matter to her principal, who in turn notified the police. On advice of the police, the teacher searched exam papers of her students and identified five students whose writing she thought similar to the threatening note. The principal refused, however, to release these exam papers to the police. The police therefore advised the teacher to ask those students to write the contents of the note to dictation five times on separate sheets of paper. Unfortunately, the instructions were misunderstood, and the writings of each student were made on only one sheet of paper. The students have since dispersed for the summer and are not available to provide additional writings. Along with these facts, the labs were provided with photographs of the note (Q) and the exemplars from the five students (K1-K5), and asked to determine if Q was written by any of the writers of K1-K5.

The questioned document and the five exemplars had been generated in the following way: The questioned document had been written by a fifteen year-old female tenth grader. Sixteen other tenth grade students ages fourteen to sixteen were asked to write exemplar sheets, and the five appearing closest to the questioned document in the opinion of the testers were used as the exemplars. In criticism of this methodology we note that sixteen people compose a small set from which to hope to get five confusingly similar handwritings and the level of expertise of the person who selected the five exemplars is unknown. At any rate, none of the people who wrote the exemplars wrote the threatening note.

The test materials were submitted to 72 labs of which 53 returned the answer sheets. Of those, 13 labs answered "inconclusive" to every exemplar, and 1 answered "other" with no specific explanation. Of the 39 remaining labs which offered an opinion actually including or excluding anybody, 16 misidentified one of the writers of the exemplars as the culprit—41% false positives—and 3 more leaned that way.[138]

---

138. Thus, it would be 49% false positives using the Galbraith methodology of including leaners. An extreme defender of handwriting experts might say that this is the wrong way to look at the data. Each decision as to each exemplar should be treated as a separate judgment. Thus, the 39 responding labs made 195 judgments of inclusion or exclusion, of which only 16 were wrong. However, an extreme critic could counter by saying two things: First, the judgments are clearly not independent. The more you are sure one wrote the note, the more you are sure the other four did not. Second, the important thing in an expertise is not absence of errors, for then a refusal to answer would be counted as establishing the existence of the expertise. The important thing is how often one makes a judgment which is wholly correct when given the opportunity. In this case, 72 labs were given the opportunity, and only

THE GALBRAITH ADMINISTRATION OF THE 1987 FSF TEST TO NONEXPERTS

The Galbraiths obtained copies of the 1987 FSF handwriting test materials and administered them twice, first to a group of thirty-two varied nonexperts who received the materials in photocopy form, and then (much later) to a group of thirty-three nonexperts, who were given photographs of the materials. Some methodological questions exist about the administration of these tests. Nothing is said about the exam conditions, the time allotted, or the instructions given to the nonexperts tested. We do not know if they were administered in a way designed to prevent potential cuing. Further, it is likely that the nonexpert test takers suffered from similar problems of varying motivation that we discuss below in regard to the Kam et al. study. In addition, the nonexperts may have been less likely to hedge their bets with some sort of inconclusive answer, perhaps not regarding such an answer as acceptable in the same way professional document examiners might.[139]

However, taking the data at face value, some interesting details emerge. First, even the lay people did significantly better than chance.[140] Second, if the FSF classification is utilized, the performance of the nonexperts and the experts is virtually identical for true positives (17 out of 33 for the "experts," 16 out of 32 and 17 out of 33 for the two groups of nonexperts). However, the nonexperts were significantly worse when it came to false positives (34% versus none for the document examiners). Finally, if we accept the Galbraith reclassification of answers, document examiners significantly outperformed lay persons as a group (91% correct, no false positives, 3% false negatives, 6% inconclusive; against 58% correct, 34% false positives, 8% false negatives). While these two limited and noncomparable administrations of the test materials cannot establish anything with certainty, they seem to suggest that, at least as to the easiest tasks of handwriting comparison, the experts may have an approach that guards against affirmative errors. They were no more affirmatively accurate than the "nonexpert" population, but they were significantly less inaccurate. Any more confident conclusion as to whether this advantage truly exists, and whether it exists in relation to other more difficult tasks, must await further research.

---

23 unambiguously did the job right. Of these various ways to view the data, we will stick to the position that ours is most meaningful, since it emphasizes the percentage of experts whose responses could lead to court testimony and who would then give mistaken inculpatory testimony: 41%.

139. This may reflect the instructions, or absence of instructions, on this issue.

140. *See* Galbraith et al., *supra* note 101, at 16 (data table 4).

THE KAM, WETSTEIN, AND CONN STUDY

Under a research contract with the FBI, Kam et al. designed a test intended to determine whether some professional document examiners had handwriting identification skills significantly better than those of nonexperts. The test design was as follows:[141] Forty-five Drexel University undergraduates copied five test samples on to five sheets of paper. Thirteen of the students copied one or another of the samples an extra time. All used their normal handwriting. Fourteen of the students wrote with the writing utensils they had with them. Eighteen of the students wrote with medium point Bic pens supplied by the researchers. Thirteen students "randomly swapped pens with each other" between writings, but we are not told what the original source and type of their pens were.[142] This procedure resulted in 238 documents, from which 86 documents were randomly selected as the test materials. These documents represented the work of exactly twenty writers and were then tagged with a "non-trivial" code which would allow the holder of the code to connect each document to a particular writer.

The test consisted of handing each person taking the test the eighty-six documents and telling them to go into a room with a table and sort them into piles each of which represented the work of a single writer. Test subjects were not told how many writers were represented in the eighty-six documents, and no time limits were imposed on them. There were two groups of test subjects. The first consisted of seven FBI document examiners (presumably chosen by the FBI to take the test) who took the test in Washington, D.C. after being handed the materials and given the instructions by one of their own supervisors. The second consisted of ten graduate students in the Drexel graduate engineering and MBA programs who were selected in an unspecified manner to participate and who had the test materials administered to them by their "supervisors" (whatever this may mean in the context of graduate students).[143]

A perfect performance on the test would yield twenty piles, each containing from one to six documents. Each pile would contain all the documents written by a single writer and only the documents written by that writer. Two types of error were possible: First, a participant could put a document in a new pile when there was already a pile for that author's work. This type of error the authors called over-refinement—making

141. *See* Kam et al., *supra* note 110, at 7-8 (outlining all of the test design information).

142. It is totally unclear why all were not simply supplied with medium Bics from the beginning. This variation in ink or pencil types introduced a variable into the study's design having unknown impact and no apparent relevance to the issue of identification from form. For example, it may have created a distraction which the nonexperts were more susceptible to, leading to inflated differences in scores between experts and nonexperts.

143. The form of test administration was selected to impress upon the participants "the importance that their respective institutions attach to these experiments." We suspect that the FBI document examiners were more impressed with this message than the graduate students.

distinctions that were not there.[144] Second, a participant could put a document in a pile containing the work of a different author. This type of error the authors called under-refinement—failing to make a distinction that should have been made.[145] Of the two, the authors observed that under-refinement errors were perhaps "more significant, since they represent a confusion between two writers," that is, they create the risk of false positives, as opposed to the risk of false negatives resulting from over-refinement.[146]

Since the test materials were not reproduced, there is no first hand way of judging whether this sorting test was inherently easy or difficult. However, because the samples were generated with no attempt to apply any standards of pictorial similarity, it would seem to be like matching photos of twenty randomly selected humans of all races and sexes: that is, many of the subtask matches would seem likely to have presented trivial challenges. In addition, as previously noted, the harder tasks are subject to confusion by virtue of the effect of the uncontrolled variable of writing instrument variation. Differences in result between two test takers may actually reflect different assumptions concerning the test structure and the meaning assigned to writing instrument variation. We also do not know what relationship there is between the skills necessary for this test and the skills brought to bear in real-life document examination tasks, since document examiners are not called upon to do this kind of sorting task in their ordinary work. However, the ability to accurately perceive diagnostic patterns of similarity and difference in the writing represented by the test materials would likely be common to both the test and to many kinds of real-life problems. At any rate, the same test was administered to the two test groups with the object of determining whether there were differences in performance between the two, at least as to this test, for whatever reason.

The results obtained by the FBI document examiners were so good that one is tempted to conclude that the task was globally an easy one, especially given the differences in performance between easy and hard tasks which seem to appear in the FSF studies. It may also be that the seven examiners selected to take the tests by the FBI were otherwise reputed to be the best in their employ. Nevertheless, taken at face value, their performance was impressive—five of the seven were perfect and the other two had only two errors each, one making one extra pile and having one incorrect inclusion, and the other making two extra piles.[147]

It is also true that the performance of the nonexpert test takers was, in the aggregate, clearly inferior to that of the FBI document examiners. However, there is another uncontrolled variable which may account for

144. *See* Kam et al., *supra* note 110, at 8.
145. *Id.*
146. *Id.* at 9.
147. *Id.* at 10.

much of the difference—test-taker motivation. Clearly, knowing the growing controversy over the bare existence of any such expertise, the FBI document examiners realized that the foundation of their very careers may have been at stake when they took the test. The graduate students had nothing at stake beyond some varying individual notions of personal pride at doing the best job they could, if that.[148] In line with this, the striking thing about the performance of the graduate student group was its extreme bimodality. Four of the ten made only between nine and fourteen total errors, averaging eleven, which were mainly errors of over-refinement (too many piles, one person's writing counted as two persons' writings).[149] Fully half of the graduate students made only two or three errors of under-refinement, assigning the authorship of a document to the wrong person, and one of the FBI examiners made one such error.[150] If the top 40% of nonexperts represented the properly motivated performance of nonexperts, the differences in performance between the experts and nonexperts, while perhaps statistically significant, is less significant in practical terms.

At the other end of the performance scale, the bottom 40% of the graduate student group made between 31 and 45 errors, averaging 38.5.[151] One person made 21 extra piles (41 total piles) and then assigned 24 of the remaining 45 documents to the wrong piles. Another made a total of 58 piles (38 extra) and assigned 6 of the remaining 28 to wrong piles. (One suspects that there was a subset of 21 or 22 documents in the set of test materials which could have been given as a virtually unfailable test by anyone not blind or massively dyslexic.)

A note on the method of administering the test to the FBI agents: First, we know that the tests were administered "in Washington, D.C." and

---

148. It would be interesting to know how much time was actually spent doing the tests by the members of each group, since there were no imposed time limits, but those data are not provided and, indeed, were not kept. In his phone conversation with one of the editors, Professor Kam indicated that he was present for each of the graduate student runs, and that all expended some hours on the task. However, according to Professor Kam, the test was designed after enquiry to document examiners concerning their notions of how long tasks take, to be a full working day's project for a document examiner, so there is some mild reason to believe that the document examiners spent significantly more time than the graduate students on the task, but there are no actual data to that effect.

One can argue that it is exactly the difference in motivation to expend time examining and comparing details of questioned writings and known exemplars that justifies the admission of document examiner testimony, since the lay jury is not likely to spend the time analyzing the material assertedly necessary for peak performance even if they could perform as well as document examiners if they spent the time. This is an interesting notion, analogous to the rationale for allowing testimony regarding summaries of voluminous material. *Cf.* Fed. R. Evid. 1006. How this fits in with our usual notions of expertise is not clear. At this juncture there is insufficient information on the contours of both lay and document examiner accuracy to justify rethinking the role of or justification for allowing such expertise.

149. *See* Kam et al., *supra* note 110, at 7-8.

150. *See id.* at 10.

151. *See id.*

that they were "administered through the agents' supervisors." We do not know from the published record exactly what this means. However, Professor Kam graciously filled in many of the missing details.[152]

Certain obvious questions occurred to us before the conversation with Professor Kam: Were the test materials sent to Washington and kept there during the period necessary for the seven test runs, without the presence of a representative of Kam et al.? If so, this might severely undermine the validity of the results, raising as it does the distinct possibility of collaboration by substantially interested parties uncommitted to the standards of academic science. Further, while the test subjects were told nothing about the characteristics of the test materials and how they were generated, it is unclear whether their supervisors or others in the FBI were given such information. After all, the test was developed pursuant to an FBI contract with Professor Kam to study the methods of handwriting experts with an eye to developing a computerized scanner which could perform this work as a screening tool. If people in the FBI had information about the characteristics of the test materials, the possibility of intentional or unintentional disclosures could not be ignored. Finally, there is the matter of the "non-trivial code" with which the test materials were inscribed, which allowed them to be matched later for statistical analysis. One could hardly object to randomly generated numbers which matched an identification key kept at all times by the researchers in Philadelphia. However, if to make their computer work easier they embedded the identification information in the code itself, the entire test could have been compromised if the materials were sent to the FBI without researcher proctoring.

In response to these questions, Professor Kam told us that no one in the FBI was given any information on the characteristics of the database from which the test materials were drawn, or on how they were selected. The tests were administered in Washington by sending the materials to an FBI contact after giving him instructions on how to administer the test. He was to administer the test to one agent, then return the results of the sorting to Professor Kam for scoring. Further, as to the code, it did consist of a randomly generated number, and to guard against those who had already taken the test sharing any useful tips with those who had not, the coding to the eighty-six papers was changed each time they were sent to Washington for administration. However, since there was nothing in the actual design of the test which would insure that the test materials were not photocopied the first time they were sent to the FBI, collaboration on the test by the FBI document examiners cannot be procedurally ruled out. On the other hand, the performance of the FBI document examiners was remarkable even if there was collaboration, though the results under those circumstances would be attributable only to the group, or to the best performer in the group.[153]

---

152. Telephone interview by Michael J. Saks of Moshe Kam (Autumn, 1994).

153. Before speaking to Professor Kam, there was some reason to believe that the test

Where does this leave us? If the level of graduate student performance and its variation are the product of motivational differences and the artifactual impact of the writing instruments, the data could no longer be taken to support a marginal performance advantage in the experts, and these artifacts cannot be ruled out as insignificant. If the variations mirror reality, then we potentially have an even stranger situation that is also consistent with the results of the Galbraith study. It may be that average people have a wide range of natural abilities, ranging from good to poor. It may be that, at least as to easy tasks, the expert is a lot better than half the nonexperts but not much better than the other half. (This might simply result from people in the half of the nonexpert population with the natural ability comprising most of those drawn to the work of document examination in the first place.) In the face of harder tasks, there is reason to believe tentatively that the marginal advantage may break down. What should be the law's response to such a situation? The law is clearly not well equipped to put such a situation into its usual paradigm, which assumes a substantial skill differential between all members of the expert group and the overwhelming majority of the randomly selected population.

CONCLUSIONS

In summary, what should we make of the corpus of available empirical data concerning handwriting identification expertise? First, any affirmative use of the data to support any hypothesis must be heavily qualified because of the small number of studies that have been conducted and their methodological problems. Even with this in mind, it seems fair to say that such a skill probably exists in some people for certain tasks, but probably involves a large amount of inherent talent. While training and practice of

---

administration was over-supervised, not under-supervised. Agent Richard Williams, testifying in a proceeding in San Francisco in 1993, testified that he was one of the test subjects, and that during the administration of the tests he was surrounded by people in white coats with clipboards asking him questions about what was going through his mind as he did the test. *See Smyth* transcript, *supra* note 1, at 127 (testimony of Richard Williams). He referred to the test as a "quick and dirty" sorting. *Id.* at 125. If the tests were conducted in this way, they would appear not to have been double blind, and the phenomenon of "Clever Hans" cuing (named after the famous German horse that was thought capable of solving arithmetic problems until he was tested under careful procedures which prevented subtle cues from reaching him) cannot be ruled out. However, at the time of his testimony Williams was trying to diminish the significance of the fact that *any* mistakes had been made by the professionals, since he had earlier indulged in the usual hyperbolic testimony about the virtual perfection of handwriting identification. *Id.* at 17-18, 111. In addition, he may not actually have been one of the seven test subjects (though he insistently testified that he was. *Id.* at 127) but merely one of those document examiners "interviewed at length after the testing" to try to determine the mental processes behind document examiner performance. *See* Kam et al., *supra* note 110, at 13. It seems very likely that these interviews were conducted around an examination of the test materials, and that this is what Williams was involved in, not an actual run of the test itself. Williams' testimony does give rise to questions concerning the way these materials are presented to courts by document examiner witnesses, and the lengths to which they may go to bend the meaning of the tests in a desired direction.

"the true method" of atomized analysis may lead some people to become more dependable, such credentials and experience alone do not necessarily establish the existence of a dependable skill.

In addition, the data do seem to support the notion that there are many context-defined subtasks involved in handwriting identification which vary widely in levels of difficulty. The skill that does exist seems to be undependable even in many of those who have it, when it comes to many difficult subtasks. Not even the most generous reading of currently available data could support a claim that we know how to define all, or even a significant number, of those subtasks. As a practical expertise, handwriting identification differs from many areas (such as Judge McKenna's harbor piloting) in that independent confirmation of the accuracy of one's results does not dependably emerge from everyday practice.

Finally, there is reason to believe that, like eyewitness identification, handwriting identification is strongly influenced by context cuing. In other words, the presentation of extraneous information to the examiner which indicates the answer desired and reasons beyond handwriting for its being the correct conclusion may affect the document examiner's conclusion. Such presentations appear to be the norm in the submission of actual cases to document examiners, as even the materials of the FSF studies reveal.[154] Here the founding fathers of the area were more sensible than their progeny. Hagan, Ames, and Osborn all took strong positions that it was the professional duty of the document examiner to insist that such information not be presented and to take steps to set up modes of consultation to insure against such contamination.[155]

---

154. *See supra* the discussion of the FSF studies (discussing the presentation and use of such extraneous context information by the designers of the FSF studies).

155. Hagan stated:

> [T]he examiner must depend wholly upon what is seen, leaving out of consideration all suggestions or hints from interested parties; and if possible it best subserves the conditions of fair examination that the expert should not know the interest which the party employing him to make the investigation has in the result. Where the expert has no knowledge of the moral evidence or aspects of the case in which signatures are a matter of contest, there is nothing to mislead him, or to influence the forming of an opinion; and while knowing of the case as presented by one side of the contest might or might not shade the opinion formulated, yet it is better that the latter be based entirely on what the writing itself shows, and nothing else.

Hagan, *supra* note 29, at 82.

Ames commented:

> No expert should permit himself to be *retained* in the sense in which an attorney is retained, viz., for the purpose of making the most of and winning a case, right or wrong . . . .
>
> When the services of an expert are sought, he should, so far as is possible, avoid knowing the circumstances or the relations of the party asking his opinion as to the case.

Ames, *supra* note 30, at 89.

Osborn noted:

FUTURE DIRECTIONS

The future of handwriting identification experts, like their past, is closely linked to the law's tolerance for accepting a field's claims of expertise on their face, without requiring empirical, scientific support. Unlike normal sciences, whose value rises or falls on how well their theories and claims are empirically demonstrated, handwriting experts have been admitted on little more than their assertions of expertise. Indeed, as long as the courts did not care what the body of research showed, practitioners of handwriting identification could ignore both the lack of scientific support for their claims and even data that contradicted those claims. Put simply, if courts trust handwriting experts to be experts, little incentive exists to advance the field's knowledge or to test its claims. And so, in the past century virtually no research of that kind has been done.

By contrast, all it took to awaken in at least some document examiners an interest in becoming scientific was Judge McKenna's conclusion that handwriting identification expertise was not a science.[156] In the wake of *United States v. Starzecpyzel*, at least one organization of document examiners has expressed interest in collaborating with empirical researchers in an effort to try to place their field, for the first time, on a scientific foundation.[157] And the American Academy of Forensic Sciences scheduled a set of sessions at its subsequent annual meeting to examine the field's shortcomings as suggested by *Daubert* and as detailed by *Starzecpyzel*. On the other hand, if a majority of the field concludes that remaining unscientific is the only way to remain dependably employed, recent interest in participating in studies may die aborning.[158]

> It should clearly be understood that the chief source of error in these cases is this intense partisan spirit and the spirit of advocacy surrounding the whole proceeding. The scientific examiner deliberately endeavors to keep outside the circle of these influences. Too often the investigation of what is in fact a genuine document, or of what is in fact a crude forgery, is not taken up as a scientific investigation but every argument and every influence is brought to bear in order to get favorable opinions and assistance from those who can assist in any way . . . .
>
> There is, of course, certain legitimate information that the qualified examiner should have as to alleged conditions surrounding a document that is questioned but he does not need to know and should not be told why this or that should have been done or should not have been done by a testator or why, for other outside reasons, it is reasonable to assume that the alleged act was, or was not, performed. One who examines a document should have information as to the condition of an alleged writer or any alleged surrounding conditions that may have affected the result or any facts that are a legitimate part of the technical problem which is submitted to him.

Osborn 2d ed., *supra* note 25, at 2-3.

156. This was so, even though *United States v. Starzecpyzel* still held forensic document examination testimony to be admissible.

157. The Independent Association of Questioned Document Examiners.

158. Consider the observations in *supra* note 106, and the FSF's apparent withdrawal from

Thus, document examiners will dependably do only what judicial opinions make them do. How should the courts respond? Unless and until research is developed establishing the truth of the claims of forensic document examination as a field of scientific expertise, the courts might attempt to require adequate testing of each individual examiner, to ensure that person's ability to do what she or he claims to be able to do. No tests currently exist for this purpose, however. Absent specific proof of the actual skills of a given practitioner through some yet undeveloped testing regime, one would expect the law to assume such skills do not exist.

But experience teaches that the courts will do otherwise. That in itself is an interesting lesson about legal reality. At the very least, judges should admit no testimony concerning handwriting identification unless the proponent presented the documents from which the conclusions were derived to the expert without unnecessary contextual information, and pursuant to a handwriting lineup, or at least in some other manner without context suggestivity which meets appropriate standards to be developed and enforced by the courts.[159]

---

studying handwriting identification proficiency following the publication of Risinger et al., *supra* note 3.

159. That is, the presentation of exemplars from a suspected author mixed in among the other exemplars from other people having reasonably similar system characteristics. *See* Risinger et al., *supra* note 3, at 775-77. Though unheard of in today's practice, such a procedure is not without precedent. A very elaborate handwriting lineup procedure involving all 256 cadets, complete with a changeable random number coding like that in the Kam et al. study, *supra* note 110, was undertaken in the Cadet Whittaker court martial, a 19th century *cause celebre* involving allegations that one of the first black West Point cadets had forged threatening letters to himself to corroborate his charges of harassment by his classmates. For a full description of the procedure used, see Hagan, *supra* note 29, at 160-73 (Hagan was one of the three persons who examined the handwriting for the government); *see also* Ames, *supra* note 30, at 190-92 (Ames was another of the government's witnesses); *Calligraphy and the Whittaker Case*, 2 Crim. L. Mag. 139 (Mar., 1881). The results of the Whittaker case are still controversial. *See generally* John F. Marszalek, Court Martial: A Black Man in America (1972), republished as Assault at West Point: The Court Martial of Johnson Whittaker (1994) (also made into a television movie) (suggesting strongly that Whittaker was innocent).

APPENDIX

A Summary of the Principles of Handwriting Identification Theory

1. There is a signal in a handwriting trace which will allow an observer to establish who wrote it dependably under conditions which are usually (but not always) present.[160]

2. This signal exists because of the development of many personal characteristics in handwriting over time which combine to make a virtually unique individual pattern,[161] which can be determined by observation of sufficient examples of the handwriting to derive the pattern[162] even in the face of inevitable variation around the pattern in any given piece of writing.[163]

3. This individual pattern is almost impossible to duplicate undiscoverably, even by someone of skill trying to do so, because the variables are too numerous and inconspicuous and also because the unconscious conflicting personal habits of the writer will manifest themselves either from the beginning or in a very short time.[164]

4. Even when it cannot be established positively who left the trace, information in the trace may exclude a candidate writer.[165]

5. The best way to dependably extract information from the trace is not by gestalt exam, which is not totally useless but often misleading,[166] but by a system of atomized analysis of elements.[167]

6. Atomized analysis breaks the trace down into components, some of which are measurable, some not.[168] The usual system is a frankly incomplete taxonomy[169] with many categories not suited to objective measurement, or to only imprecise estimate.[170]

---

160. *See* Osborn 1st ed., *supra* note 28, at 200-08. The first edition of Osborn's *Questioned Documents* has been chosen as this Appendix's primary source of authority to emphasize its foundational status. Only on the rare instances in which some later book offered a clear qualification or variation was that other source referenced. However, anyone wondering whether the main outlines of the summary given here still reflect current theory need only compare it to Ordway Hilton's article on the subject in 13 Encyclopedia Americana 765 (1992). Other works thoroughly examined in preparing this summary include: Ames, *supra* note 30; F. Brewester, Contested Documents and Forgeries (1932); Luciano V. Caputo, Questioned Document Case Studies (1982); Chabot (and Twistleton), *supra* note 27; Conway, *supra* note 40; Ellen, *supra* note 90; Hagan, *supra* note 29; Harrison, *supra* note 92; Ordway Hilton, The Scientific Examination of Questioned Documents (1956); Hilton, *supra* note 92; Osborn 2d ed., *supra* note 25; Osborn, *A New Profession*, *supra* note 40.

161. *See* Osborn 1st ed., *supra* note 28, at 196-97.

162. *See id.* at 196, 231.

163. *See id.* at 196-97, 231.

164. *See id.* at 237-38.

165. *See id.* at 18-19.

166. *See* Osborn 1st ed., *supra* note 28, at 206, 244-45, 262-63.

167. *See id.* at 30, 209, 242-43, 253.

168. *See id.* at 109-10.

169. *See id.* at 209.

170. *See id.*

7. Handwriting questioned documents are either signatures alone, or other more or less extended writing, with or without signatures.
8. Signatures generally are the most personal and uniform of writings,[171] though some people may have more than one signature for different uses or contexts.[172]
9. When dealing with a signature, two questions are commonly asked:
   a. Did the person whose name is reflected sign it (is it genuine)?
   b. Did some other particular person sign it if it is not genuine?
   Under most circumstances determining whether or not a signature is genuine is the easier task to perform.[173]
10. On the issue of genuineness, a document examiner should start with the questioned signature,[174] and observe:
   a. the signature as a whole, determining the dominant general underlying handwriting system it represents[175] (if possible; everybody learned some system to start, though some people have been exposed to more than one system even when learning by virtue of moving between schools or cultures);[176]
   b. the size of the writing compared to usual signatures of people in general on comparable documents;[177]
   c. location of words in regard to the (real or imaginary) signature line: above, below, trending up or down;[178]
   d. any oddities of letter alignment relative to the words, above or below the general alignment;[179]
   e. proportion of parts: the ratio of the height to the width of the various lowercase "minimum" letters (like a, o, i, c, etc.) and the base parts of the other lower case letters;[180] the proportion of tall lower case letters above the base line to the height of minimum letters;[181] the proportion of lower case letter extensions below the baseline to minimum letters,[182] and any divergences and oddities from the general pattern for individual letters;[183] proportion of capitals to minimums, and the height-width ratio of capitals;[184]

---

171. *See* Osborn 1st ed., *supra* note 28, at 210; *see also* Hilton, *supra* note 92, at 156 n.4.
172. *See* Osborn 1st ed., *supra* note 28, at 213-14.
173. *See id.* at 13.
174. Actually, Osborn takes the position that it is best to start with an analysis of the known standards before examining the questioned signature, but concedes that this is often not possible in practice. *Id.* at 243-44.
175. *See id.* at 190, 263.
176. *See id.* (generally ch. XI).
177. *See* Osborn 1st ed., *supra* note 28, at 144-45.
178. *See id.* at 142.
179. *See id.* at 123-24, 142.
180. *See id.* at 146.
181. *See id.* at 145-48, 215, 309-10.
182. *See* Osborn 1st ed., *supra* note 28, at 246.
183. *See id.*
184. *See id.* at 145-48, 215, 309-10.

f. slant of writing measured with glass protractor or goniometer,[185] with any divergences from general slant for individual letters noted (some slants may be nearly chaotic);[186]

g. presence or absence of lines connecting words or initials, and their form;[187]

---

185. *See id.* at 150-54, 246. Harrison supplies the word "goniometer." Harrison, *supra* note 92, at 330. Harrison asserts that by his date of publication (1959) most writing submitted to his laboratory in England had such variable slant (called "slope" by him) that it was not worth trying to measure. *Id.*

Here we must consider the role in document examination of that cornerstone of modern science, actual measurement in reproducible quantified standard units of measurement. Osborn had a chapter on measuring instruments in both editions of his book, and the characteristics we are dealing with, such as proportion of parts of various letters, are potentially subject to standard quantified measurement, with mathematical expression of central tendency and variation, etc. In this regard Osborn was on some level aware of the potential value of measurement, and wrote:

> The various parts of an ordinary signature when carefully measured bear a certain proportion to each other that with most writers is found to be surprisingly uniform. . . .
>
> When a considerable amount of writing is in question and an adequate amount of standard writing is supplied for comparison, a system of measurements covering a sufficient number of features and examples may be very forceful evidence. . . . Any system of averages, to be reliable, must be based on an adequate number of separate examples.

Osborn 1st ed., *supra* note 28, at 146. However, Osborn then goes on to say:

> Evidence based on the very great number of minute measurements necessary to show a very slight divergence is not usually of much weight in this or any similar inquiry, because it is practically impossible for court and jury to review and verify the basis of such an opinion. If the difference is apparent by inspection, then the measurements are of value in making definite what is apparently a fact without such proof.

Id. at 146-47. In the rest of the book, Osborn recommended the actual quantified measurement of very few handwriting characteristics (as opposed to typewriting, for instance), slant being primary among them. Though Osborn occasionally refers to the "size," "position," and "distance" of characteristics, *see id.* at 246-47, 309-10, in practice all characteristics but slant seem to have been subject to rough subjective estimation with no precise standard measurement at all. Instead, visual charts are relied upon to illustrate assertions of "size," "position," and "distance." Certainly there is no indication in the *Hauptmann* trial testimony of Osborn, his son, or any of the other experts, of actual quantified measurement of characteristics. Albert S. Osborn testimony, 1/11/35, *Hauptmann* transcript, at 881-1008; Albert S. Osborn testimony, 1/14/35, *id.*, at 1009-51; Elbridge W. Stein testimony, 1/14/35, *id.*, at 1074-153; John F. Tyrell testimony, 1/15/35, *id.*, at 1154-247; Herbert J. Walter testimony, 1/15/35, *id.*, at 1248-71; Harry M. Cassidy testimony, 1/16/35, *id.*, at 1279-301; Wilmer T. Souder testimony, 1/16/35, *id.*, at 1301-36; Albert D. Osborn testimony, 1/16/45, *id.*, at 1337-86; Clark Sellers testimony, 1/16/35, *id.*, at 1386-432. Actual measurement appears to play no greater role in standard practice today than in 1935. As Ordway Hilton said in his 1974 Preface to a facsimile edition of *Bibliotics*, "[w]hile certain workers continued to urge the use of measurements, the method has virtually been discarded as too time consuming for the little value which might be derived from it." Ordway Hilton, *Preface* to Persifor Frazer, Bibliotics ii (New York, AMS Press 1974) (3d ed. 1901).

186. *See* Harrison, *supra* note 92, at 330.

187. *See* Osborn 2d ed., *supra* note 25, at 143-44.

h. presence and placement of punctuation relative to initials;[188]

i. character and construction of connections between letters: smalls to smalls and capitals to smalls;[189]

j. initial strokes, presence or absence, and how formed;[190]

k. presence and placement of any pen lifts within words,[191] and whether they result in discontinuities (spaces) between letters in the middle of words;[192]

l. forms of each individual letter—general pictorial style, existence of loops, retracings, decorations and flourishes, open tops, etc., with observation of variation in form as letters are repeated;[193] whether variation correlates with position in the word, beginning or middle;[194] forms of endings of terminal letters,[195] method of crossing t's,[196] presence and form of i-dots,[197] etc.

i. Generally, some judgement should be made concerning the rarity of those characteristics (like undotted i's) which diverge from the underlying system.[198]

ii. Look also for abbreviated letters (letters only partially made or suggested) which are likely to be relatively personal characteristics;[199]

m. Consider the dynamics that created the static trace,[200] trying to infer pen position and arm and finger movements from line quality[201] (much harder now that nib pen writing is uncommon);[202]

i. direction of stroke—this detail was revealed by shading in the days of nib pens, and was not analyzed separately by Osborn, since stroke direction was then both obvious and apparently more standard. Nonstandard stroke direction is not mentioned

---

188. *See id.*

189. *See* Osborn 1st ed., *supra* note 28, at 225-26.

190. *See id.* at 220.

191. *See id.* at 121-23.

192. *See id.*

193. *See id.* at 246-47.

194. *See* Harrison, *supra* note 92, at 301.

195. *See* Osborn 1st ed., *supra* note 28, at 217-20, 248.

196. *See id.* at 218-20, 248.

197. *See id.* at 248.

198. *See id.* at 228; Osborn 2d ed., *supra* note 25, at 264-66 (expanding upon Osborn 1st ed.).

199. *See* Osborn 1st ed., *supra* note 28, at 247; Osborn 2d ed., *supra* note 25, at 255-57 (elaborating upon Osborn 1st ed.).

200. *See* Osborn 1st ed., *supra* note 28, at 106.

201. *See id.*

202. *See* Harrison, *supra* note 92, at 330 (pointing out that the "wide adoption of the ball-point pen" makes determinations based on shading more difficult); Hilton, *supra* note 92, at 155-56 ("Most modern pens, particularly ball point and soft tip pens . . . may not reveal clearcut evidence as to the angle that the pen makes with the paper and the line of writing . . . .").

in Osborn's first edition and is mentioned only twice in the second edition.[203] With the coming of ball point pens and the degeneration of standard writing discipline, stroke direction appears to have taken on a more important role in attempted identification, especially of block printing, and principals for its inference from various characteristics of the static trace have been proposed.[204]

ii. speed—by looking at smoothness and length of curves, classifying writing from slow to rapid (sometimes has evidence of both in different parts);[205]

iii. blunt or pointed beginning or ending strokes showing drawn starts and stops, or flying starts and stops;[206]

iv. muscle movement[207] (very hard to infer accurately absent nib pen);

v. pen alignment[208] (very hard to infer accurately absent nib pen);[209]

vi. pen pressure—shown by depth of pen indentation into paper, heaviness of ink, width of line, etc.;[210]

---

203. *See* Osborn 2d ed., *supra* note 25, at 360, 408. Both these mentions are in regard to photographic illustrations without further discussion in the text.

204. *See* Ellen, *supra* note 90, at 15-18; *see also* Hilton, *supra* note 92, at 211.

205. *See* Osborn 1st ed., *supra* note 28, at 110, 113, 117.

206. *See id.* at 248.

207. *See id.* at 105-10.

208. *See id.* at 128, 242.

209. Hilton, *supra* note 92, at 155-56.

210. Osborn 1st ed., *supra* note 28, at 132-34. There is a great cleavage in the handwriting identification community, the gulf between the orthodox Osbornians and the graphology-influenced practitioners. The orthodox Osbornians reject any role for theories developed by those seeking to read personality characteristics from handwriting in the identification of authorship, citing, ironically, the absence of validation for those theories. Osborn 2d ed., *supra* note 25, at 442-44. The graphology-oriented practitioners tend to believe in graphology, but they assert that one doesn't have to believe specific character traits are revealed by handwriting to gain useful insights into identification from graphological studies. While both Osbornians and graphological practitioners assert that inferences concerning dynamic aspects of writing from the trace can be important in identification of authorship, graphological practitioners tend to emphasize such dynamic characteristics as speed, pressure, and rhythm. *See* R. Saudek, Experiments with Handwriting 138-48 (1929). A number of further ironies must be reported in this regard. First, while it is true that the Orthodox Osbornians have occupied most prominent positions in the document examiner community in the last ninety years, and have controlled the membership of the American Society of Questioned Document Examiners and certification by the American Board of Questioned Document Examiners, there may be a greater number of graphologically-oriented practitioners doing everyday work in court, being heavily represented in the numerically larger membership of organizations such as the National Association of Document Examiners, the World Association of Document Examiners, and the Independent Association of Questioned Document Examiners. Second, it was only the more scientifically oriented of the graphologists, and none of the Osbornians, who performed any empirical studies on the handwriting phenomenon worthy of the name empirical from the 1920s to the 1980s, even going so far as to develop a specially instrumented pen called a graphodyne to record the dynamic aspects of writing as it was taking place,

vii. impulse—sudden changes of direction, may require magnification to be counted accurately, can be counted for individual upstrokes and downstrokes;[211]

viii. impulse grades into tremor—rapid shaking (important to note on what strokes present, since tremor of old age or nervousness may be less on upstrokes or final stroke than tremor of fraud resulting from trying to draw a facsimile under pressure);[212]

ix. general line impression: flowing, free, rhythmic, halting, slow or drawn;[213]

n. keep alert for retouching, and pen stops or lifts at angles, or on first or last stroke, or other unnatural places.[214]

11. At this point, even without exemplars, a signature may be classified as suspicious if it appears slow and drawn with blunt beginning and ending strokes, much retouching, odd pen lifts, many angular direction changes on what would usually be curves, coupled with tremor on other strokes, especially if all of this occurs in a signature which does not appear erratic and out of control on a gross level.[215]

12. To reach a firmer conclusion, one needs exemplars of genuine signatures, preferably from documents of like kind and like formality from the same time period during which the questioned signature was alleged to have been signed.[216] In addition, the examiner should know any claimed circumstances surrounding the making of the questioned signature that might affect the value of an exemplar, such as age or disease.[217] There should be as many exemplars as possible up to about 50-75,[218] though in some circumstances even one, sufficiently close in time will be enough to expose a crude forgery.[219]

13. The known exemplars should be analyzed just like the questioned

---

the forerunner of today's digitized pressure tablets used in academic motor control studies which utilize handwriting as a convenient motor control phenomenon. *See generally* Hartford, *supra* note 6, at 107-08 (describing Klara Roman's invention of, and C.A. Tripp's refinement of the graphodyne). Third, recent academic studies tend to indicate that the dynamic aspects of handwriting may indeed be more dependably indicative of individual authorship than other characteristics, though those studies have the advantage of analyzing the dynamic characteristics directly as they are performed, not by way of problematical inference back from the static trace. *See Starzecpyzel* Transcript, *Daubert* hearing testimony of Dr. George Stelmach, 3/1/95, at 386-87. Finally, since government laboratories tend to be officially Osbornian, if there are few data on the dependability of identification of authorship by Osbornian practitioners, there are none on the dependability of graphologically oriented practitioners.

211. *See* Osborn 1st ed., *supra* note 28, at 111-14.
212. *See id.* at 116-21.
213. *See id.* at 109-10.
214. *See id.* at 248-49.
215. *See id.* at 18, 245.
216. *See* Osborn 1st ed., *supra* note 28, at 18, 22.
217. *See id.* at 23-24, 216.
218. *See id.* at 19.
219. *See id.*

signature, and a range of variation for each dimension should be determined, though it usually cannot be quantified, but can be observed by juxtaposition of letter examples.[220] If the characteristics of the known exemplars are consistent in every respect with the questioned signature, then it is genuine unless it can be shown to be a tracing,[221] though even a tracing should usually show evidence of being drawn rather than written.[222] If there are significant divergences it is not genuine, though what makes a divergence significant, how many divergences are necessary, and how to weigh them is not quantified.[223] It is a fact-sensitive judgment, and even one inexplicable difference might show that the signature is not genuine.[224]

14. It is much harder to determine accurately who wrote a spurious signature than to determine that it is spurious.[225] This is because a forger is either simulating, tracing from a model, or writing with no model of the true signature. In a tracing, no personal writing of the forger is present, and an attempt at simulation will usually suppress individuality enough in the short space of a signature that no conclusions can be drawn.[226] Only when someone signs another's name in the signer's own usual hand is identification a significant possibility, and even then the small amount of writing and the usual presence of some disguise makes a positive identification difficult.[227] However, there may be enough in a signature to exclude a person as a candidate. For instance, if the signature requires more skill and muscle control than the candidate can muster as determined by a sufficient set of exemplars, exclusion would be justified, because one cannot write with more skill than one has.[228]

15. Moving beyond signatures, affirmative identification of the writer of a questioned document depends on both the quantity of the questioned writing, and the quantity and quality of authentic exemplars of a candidate's writing that can be obtained.[229] Natural exemplars showing unselfconscious writing from about the time of the making of the questioned writing are best,[230] but demand exemplars may do if the questioned document is sufficiently recent and if care is taken to guard against disguise in the demand exemplars by the amount required, variations in speed of production required, etc.[231]

---

220. *See id.* at 203, 243; *see also supra* note 185.

221. *See* Osborn 1st ed., *supra* note 28, at 257-60, 266-301, 308.

222. *See id.* at 267.

223. *See id.* at 212, 214-15.

224. *See id.* at 281.

225. *See id.* at 13-14.

226. *See* Osborn 1st ed., *supra* note 28, at 13-14.

227. *See* Harrison, *supra* note 92, at 387. How uncommon it is to be able to make such a positive identification from the one or two words in such a signature is the subject of some controversy among practitioners.

228. *See* Osborn 1st ed., *supra* note 28, at 110-12.

229. *See id.* at 321.

230. *See id.* at 18-19.

231. *See id.* at 24-25. Different authors have elaborate notions of the best ways to take

16. The analysis of both the questioned and the known writing will be done as above in regard to signatures.[232] In addition, such habits as the layout and margins of the writing on the page will be added.[233] If sufficient peculiarities correspond between the two writings and there are no significant differences, the writer of the exemplars will be established as the writer of the questioned writing.[234] Which peculiarities are rare enough to be significant, which ones are independent of each other in their occurrence, how many are necessary, and how to weight them is not defined, but the following general principles are instructive:

a. System characteristics of writing systems, foreign and domestic, cannot establish identity, even in combination.[235] Thus an examiner must be conversant with a wide range of such systems and their characteristics.[236]

b. The most diagnostic idiosyncrasies are those which diverge furthest from the underlying system, which are inconspicuous,[237] and which are not common products of carelessness or the desire for speed, such as open o's and omitted i-dots.[238]

c. Each case is different and must be judged on its own particular circumstances.

---

demand exemplars. *Compare* Osborn 2d ed., *supra* note 25, at 33-34, *with* Harrison, *supra* note 92, at 442-51 *and* Hilton, *supra* note 92, at 310-22. Note that both Osborn and Harrison refer to these writings as "request writings" and Hilton refers to them as "request standards," but I have used the term "demand" because in practice that is what is involved in the usual case.

232. *See* Osborn 1st ed., *supra* note 28, at 322.

233. *See id.* at 142-43.

234. *See id.* at 211.

235. *See id.* at 169, 206, 210, 214.

236. *See id.* at 214-15.

237. *See* Osborn 1st ed., *supra* note 28, at 210, 308.

238. *See id.* at 261.