J Forensic Sci, Jan. 2005, Vol. 50, No. 1
Paper ID JFS2003387
Available online at: www.astm.org

Michael J. Saks,[1] *Ph.D. and Holly VanderHaar,*[2] *B.A.*

On the "General Acceptance" of Handwriting Identification Principles

ABSTRACT: Informed by the admissibility requirements of the broad and narrow *Frye* tests as well as *Daubert's* general acceptance factor, the present study elicited the views of a homogeneous group of forensic document examiners and a heterogeneous group of handwriting scientists regarding the degree to which a set of principles relating to the nature of handwriting and its identification were generally accepted within the respondents' fields. Among forensic document examiners, the greatest agreement was found concerning the process of examination, and somewhat less agreement on other matters. Forensic document examiners and handwriting scientists appear not to agree on the acceptability of most of the propositions.

KEYWORDS: forensic sciences, admissibility, *Daubert*, document examination, *Frye* test, general acceptance, handwriting identification, *Kumho Tire*

Courts have long struggled with the development of a test to guide their gatekeeping of expert testimony. How is a judge to determine which kinds of opinions from which areas of asserted expertise are dependable enough to be permitted at trial? Before *Frye v. United States* (1923) (1), the courts generally relied on a "marketplace" test, asking whether a proffered expertise had been found to be of value to consumers in the commercial marketplace (2, Chapter 1, Sec. 2.1). For example, if a provider of some type of medical knowledge had proven satisfactory to patients, then the testimony of those practitioners was likely also to be admissible in court.

The marketplace test was impossible to apply to fields for which the courts themselves were the principal, and often the only, consumer. Almost by definition, the forensic sciences presented the courts with just such a predicament. The court in *Frye v. United States* was confronted with one such field, early polygraph examination, for which there was no commercial market. To help it evaluate the admissibility of polygraph expert testimony, the *Frye* court devised a variation of the marketplace test: it substituted an intellectual marketplace for the commercial marketplace. The court asked whether the principles that underlay the proffered testimony had "gained general acceptance in the particular field in which it belongs." The *Frye* test was barely used by other courts for decades following its advent in 1923 (2, Chapter 1, Sec. 2.2). Most of the traditional forensic individualization sciences gained admission without passing the older marketplace test or the *Frye* test (3).

Though the *Frye* test had, by the final quarter of the century, became the dominant touchstone of admission in the United States, it is in the process of being eclipsed by the test prescribed by the United States Supreme Court in *Daubert v. Merrell Dow Pharmaceuticals* (1993) (4). *Daubert* held the test of admission in the federal courts to be essentially that of science: empirical verification of claims of expertise using sound research methods—though it allowed general acceptance to continue to play a part. While the traditional forensic sciences are assumed to face little difficulty in passing a *Frye* test, they confront a more difficult challenge when *Daubert* is conscientiously applied (2, Chapter 1, Sec. 3.4).

The traditional forensic sciences, handwriting examination in particular, almost escaped the scrutiny of *Daubert* entirely. The court in *United States v. Starzecpyzel* (1995) (5) reasoned that because handwriting expertise failed the test of *Daubert*, it was not a science; because it was not a science it was not subject to review under *Daubert*; and therefore it remained admissible. (But, "were the court to apply *Daubert* to the proffered FDE testimony, it would have to be excluded." (p. 1036).) Numerous other courts followed this circular reasoning in order to continue admitting proffered handwriting identification expert testimony. But the "nonscience" loophole was closed when the Supreme Court decided *Kumho Tire v. Carmichael* (1999) (6), in which it held that all fields, regardless of what they were called, had to pass the most appropriate tests of validity applicable to them or be refused admission. Following *Kumho Tire*—when a *Daubert* challenge is raised, when competing presentations of the empirical research bearing on the field's claims are offered, and when the court makes a conscientious attempt to analyze the evidence under the applicable law—it has become common for federal district courts to restrict or exclude handwriting expert testimony (2, Chapter 28, Sec. 1.4.3 and corresponding 2004 Supplement).

Little effort has been made to systematically address the general acceptance element, which remains viable in jurisdictions which continue to follow *Frye* and which remains a part of *Daubert*. At present, the most supportive ground for admission of handwriting identification expert opinion testimony would seem to be its "general acceptance within the particular field." However, even this hurdle is not so easy to overcome and is no longer, by itself, sufficient to gain admission. The Supreme Court has stated that the "general acceptance factor" would do little to "help show that an expert's testimony is reliable where the discipline itself lacks reliability" (p. 151) (6). Thus, what in *Frye* jurisdictions is the principal test of admissibility is, under *Daubert*, neither necessary nor, by itself, sufficient.

Indeed, one of the major criticisms of the general acceptance test had been that once a court decided what constituted "the particular

[1] Arizona State University, College of Law & Department of Psychology, P.O. Box 877906 Tempe, AZ 85287-7906.
[2] Department of Psychology, Arizona State University, AZ.
Received 16 Nov. 2003; and in revised form 21 May and 20 Aug. 2004; accepted 21 Aug. 2004; published 8 Dec. 2004.

field" within which an asserted expertise needed to be generally accepted, the admission decision itself was as good as made (2, Chapter 1, Sec. 2.4). The *narrower* that field was defined to be (e.g., forensic practitioners of spectrographic voice identification) the more likely it was to be found to be "generally accepted." The *broader* the field was defined to be (e.g., acoustical engineers, statisticians, physiologists, and linguists in addition to forensic practitioners of spectrographic voice identification) the less likely it was to be found to be "generally accepted." For example, in a report from the National Academy of Sciences, the broader group of scientists was less convinced of the capabilities of voiceprint identification than the narrower group of forensic practitioners (7). A review of judicial opinions on voiceprint admissibility found that of courts using the narrow test, few if any excluded voiceprint expert testimony; of courts using the broad test, few if any found voiceprints admissible (2, Chapter 31, Sec. 1.1).

The study reported in this article addresses the question of general acceptance by asking, with respect to a set of propositions related to the nature of handwriting and its examination and identification, what the degree of consensus is within a homogeneous group of forensic document examiners (that being the question facing a court applying a narrow general acceptance test). And, second, by comparing the consensus of beliefs among forensic document examiners with those of various kinds of handwriting scientists. Though a court applying a broad test would, in effect, be posing the question of general acceptance to the heterogeneous group of handwriting scientists in addition to the group of forensic document examiners, the most straightforward way to determine whether the narrow and the broad tests would produce the same or different answers to a court's inquiry is to compare the forensic document examiners and the handwriting scientists directly, and that is the approach taken. In addition, it is important to note that we do not ask the respondents about their own personal beliefs about the validity of the propositions, but rather what they understand the consensus of beliefs in their respective fields to be. Again, this aims to approach the problem much as a court would.

Method

Participants

Samples were drawn from two groups: a professional association of forensic document examiners (the American Society of Questioned Document Examiners–ASQDE) and a diverse association of handwriting scientists, researchers, and scholars (the International Graphonomics Society–IGS). For purposes of this article, the ASQDE members are usually referred to as "forensic document examiners" and the IGS members are usually referred to as "handwriting scientists."

The sample of forensic document examiners consisted of all 140 members of the ASQDE who, at the time of data collection, had email addresses included with their listings in the Society's member directory available on its website. When asked about the nature of their work, virtually every ASQDE respondent answered that it involved the examination of questioned documents for judicial purposes. ASQDE is an association of classical Osbornian document examiners, established in 1942 with Albert S. Osborn himself as its first president. The ASQDE is one of four groups which Moenssens (8) names as being among "the most significant group of specialists engaged in the comparisons of handwritings for the purpose of forgery detection and identification of writers" (p. 257). Moenssens describes the ASQDE as being the oldest of the document examiner groups, and having "very strict professional standards" (p. 337). "Membership applicants are required to possess at least a Bachelor's degree or the equivalent, as well as pass a written, practical, and oral examination" (p. 337). Among the group's purposes are to "foster education" and "set standards in QDE" (p. 337). The advantage of using a single organization of forensic document examiners such as this is that one would expect to find a high degree of uniformity of belief owing to the organization's adoption, promulgation, and enforcement of a common approved set of beliefs. Thus, for example, if there are different views expressed by forensic document examiners about the acceptance of a given principle, it cannot be said that this difference of views reflects the putative differential competence of different organizations of examiners or different schools of thought among examiners (8). Those examiners who constituted the ASQDE sample had been professionally involved with handwriting for an average of 25.6 years.

The sample of handwriting scientists was drawn from the membership directory of the IGS, an organization composed of scientists and other professionals concerned with the nature of handwriting from an array of different disciplines. Recall that the purpose for this comparison group was to try to sample the kinds of non-forensic experts whose views courts employing a "broad" general acceptance test would listen to—and have in fact listened to (e.g., 5). Such courts doubtless would receive the testimony of non-forensic experts who conduct research on handwriting from a wide array of scientific and technical disciplines. As suggested by a reviewer of an earlier draft of this paper, it would be desirable to refine the IGS sample by removing from it persons whose background information indicated their work with handwriting to be something other than scientific research, such as handwriting education, computer graphics, or forensic handwriting identification, and persons with these and other nonscientific/nonresearch backgrounds were in fact removed from the sample. Those persons who were retained in the sample described their fields of specialization in the following ways: pattern recognition, kinesiology, experimental psychology, bioengineering, feature extraction research, image processing/pattern recognition, psychomotor handwriting production, motor control, motor control engineering, psychology of handwriting acquisition. Those who were retained in the sample indicated the nature of their current work to be of the following kinds: experiments on dynamic features of genuine and disguised handwriting; experimental research on motor control aspects of handwriting; development of recognizers for handwriting on hand-held devices; medical aspects of handwriting; kinematic handwriting analysis of clinical populations; production and perception of handwriting movement; modeling motor control processes; measurement of writer individuality; handwriting acquisition; software development for financial and forensic signature verification; computer recognition of handwriting; computer forensic analysis; psychomotor research; handwriting models, analysis, and reconstruction; neuromotor processes underlying complex movement production; processes of handwriting acquisition; and analysis of motor control in the generation of handwriting. The advantage of such a diverse group is that it provides a good test of broad general acceptance. The initial sampling frame of handwriting scientists consisted of all 64 members of the IGS whose membership listing at the time of data collection included their email address. Those who formed the final IGS sample had been professionally involved in research on handwriting for an average of 13.4 years.

Of the 140 questionnaires sent to ASQDE members, 32 were ultimately undeliverable, 56 were never responded to, and 39 declined to participate, providing 13 useable replies. The conventional calculation of response rate for the forensic document examiners is therefore 12%. Of the 64 questionnaires sent to IGS members, 12

were ultimately undeliverable, 11 were never responded to, and 17 declined to participate, providing 24 useable replies. The conventional calculation of response rate for the handwriting scientists is 46%. The sample was then refined as described in the preceding paragraph. (Because the "refinement" is tantamount to a change in the universe of "handwriting scientists" (removing certain specialties from it), in order to calculate a more precise response rate, we would need to change the denominator to reflect that smaller universe and then divide it into the number of people from that universe who in fact participated in the study. To find that denominator would require learning background facts about members of the universe who chose not to reply, which is obviously impractical if not impossibility. Fortunately, the best estimate of what the refined response rate would be, however, is the same as the calculation just presented (46%). That is because there is no a priori reason to think that the handwriting scientist subset of IGS members would respond at a significantly different rate than the whole group did.)

Questionnaire

One way to design questions about the matters of concern in this study is to frame general statements of purported principles and ask respondents what they understand their field's degree of consensus to be concerning the validity (or invalidity) of each of those propositions. An alternative approach would be to describe situations to which the field's principles would or would not apply, and asking respondents how they understand that their field would apply its principles to the specified situation. Numerous examples of both approaches—framing general propositions or presenting more specific problems or vignettes—can be found in attitude and opinion research. Both types of queries also are familiar in legal education: one can discuss a general rule of law (the black letter law approach), or specific fact patterns to which the rule might or might not apply (the problem approach). We adopted the former method for several reasons. For the inquiry at hand, it was more straightforward, shorter, simpler, and clearer. Most important, by asking about the degree to which a given *general principle* was accepted (or rejected) by a respondent's field, it was asking exactly the question that the law asks when it inquires about general acceptance (1).

A questionnaire was developed consisting of questions drawn from the forensic document examination literature on handwriting identification (9–13). Drafts of the survey instrument were reviewed and critiqued by experts in handwriting and also in survey research of this kind, and the questionnaire was revised in light of their criticisms. Next, the survey was pilot tested with a small sample of respondents, and refined further. The ten items used in the survey questionnaire are listed in Table 1, with page references to Osborn (9) where he discussed these principles in his landmark book on questioned documents. Some of these items reflect core beliefs about the nature of handwriting or means of identification (e.g., "Inter-writer variation far exceeds intra-writer variation"). Several posed what appear to be core disbeliefs, that is, propositions that the literature led us to expect document examiners would reject (e.g.: "When a person simulates (forges) the writing of another person, a sufficient amount of the forger's own characteristics usually remain in the forgery, so that it generally is possible to identify who the forger is....").

The propositions are phrased as general principles, avoiding absolutes (e.g., no use of "always" or "never," though an occasional "virtually impossible" or "far exceeds"), and allowing for qualifications and exceptions (e.g., "it is generally possible," "more evident from X than from Y") since, of course, no general principle applies to all situations, and even principles that normally apply to a situation may be subject to exceptions. Respondents who added qualifications or explanations along with their basic responses to questionnaire items most often emphasized the assumption that an adequate amount of writing needed to be available for comparison.

Respondents were asked to indicate, "how well accepted each proposition is, among members of your field" by selecting one of the following responses for each proposition:

- Well accepted as true.
- Most people in my field regard this as true, but a respectable minority believe it to be false.
- Divided opinion; not established to be either true or false.
- Most people in my field regard this as false, but a respectable minority believe it to be true.
- Well accepted as false.

Respondents also had the option of refraining from answering an item if they believed a proposition was "not sufficiently well studied to have become established (as true or false)" within their field by so indicating. Not many responses employed this option.

To control for possible order effects, two versions of the questionnaire were created (Form A and Form B). In one version, the order of the first six items was reversed compared to the other version. Because items 7–10 were sub-parts of a single question, they were not reversed, and always came at the end of the questionnaire. Statistical analysis revealed no order effects.

Certain additional information about each respondent's background was requested: the major field or professional group with which the respondent identified, the source of knowledge on which the respondent relied to answer, formal education or training relevant to the subject of handwriting, number of years the person had been professionally involved in the subject of handwriting, and the nature of the respondent's work in regard to handwriting.

Procedure

Prospective respondents in each group were listed alphabetically by name, and each person was alternately assigned survey Form A or Form B. Each potential respondent was assigned an anonymous code number that reflected the person's membership source group and the form of the survey assigned to that person.

The surveys were sent to potential respondents by e-mail. When an e-mail was returned as undeliverable, an attempt was made to determine the correct or an updated e-mail address. As results came in, the name of the respondent was deleted and useable replies were stored for data entry along with their associated code numbers. If a participant replied by declining to participate in the survey, that person's e-mail address was removed from the follow-up list. Those participants not heard from were sent a reminder e-mail approximately every week to ten days up to a total of four such follow-ups.

Data Analysis

The principal data analyses consisted of testing for statistical significance in the comparison of means. Where a number of different items were being compared, the "analysis of variance" was employed. Where that yielded a significant result, it was proper to conduct *post hoc* significance tests to determine which individual items differed from which other items, and for this the "Bonferroni procedure" was used. In some instances, it was of interest to see whether an item's mean differed significantly from a benchmark

TABLE 1—*Propositions, sources, means and standard deviations of responses.*

Principle	Osborn Cites	Examiners	Scientists	*p*
9. By looking at a person's writing on paper (the static trace of the act of writing), it is possible to accurately infer the starts and stops.	109–110, 248–249	1.00 *0.00*	2.47 *1.30*	.001
8. By looking at a person's writing on paper (the static trace of the act of writing), it is possible to accurately infer the direction of strokes.	–	1.10 *0.32*	2.06 *1.44*	.019
7. By looking at a person's writing on paper (the static trace of the act of writing), it is possible to accurately infer the speed of the writing.	110,113,117	1.18 *0.60*	3.29 *1.44*	<.001
1. When writing in a natural way and with no attempt at disguise, no two people write sufficiently alike that one person's writing could be mistaken for that of another.	197	1.25 *0.45*	2.64 *1.21*	.003
2. An individual's writing pattern is sufficiently distinctive that it is virtually impossible to duplicate it (forge it) without detection by an experienced expert. (In other words, handwriting experts are able to determine whether a writing is genuine or forged.)	237–238	1.36 *0.50*	3.07 *1.33*	<.001
5. Inter-writer variation far exceeds intra-writer variation.	196–197	2.00 *1.50*	2.06 *1.18*	ns
3. A person's individual signature pattern is more evident from an examination of the writing as a whole, rather than from an examination of the atomized elements of the writing.	contra, 30, 206, 209, 242–5, 253, 262–3	2.75 *1.75*	2.83 *1.19*	ns
10. By looking at a person's writing on paper (the static trace of the act of writing), it is possible to accurately infer the muscle movements that created the writing.	106	2.83 *1.83*	4.42 *0.90*	.091
4. When a person simulates (forges) the writing of another person, a sufficient amount of the forger's own characteristics usually remain in the forgery, so that it generally is possible to identify who the forger is (if samples of the forger's natural writing are available for comparison).	contra, 13–14	4.18 *0.60*	3.00 *0.95*	.002
6. When trying to determine if a signature is genuine or forged, it is sufficient to have as few as a single exemplar of the known genuine writing to compare the questioned signature to.	contra, 321, 24–25	4.67 *0.89*	3.77 *1.36*	.063

Means are the top numbers in Roman type, *Standard Deviations are the lower numbers in Italics*.

Response Scale: 1 = Well accepted as true; 2 = Most people in my field regard this as true, but a respectable minority believe it to be false; 3 = Divided opinion; not established to be either true or false; 4 = Most people in my field regard this as false, but a respectable minority believe it to be true; 5 = Well accepted as false.

point on the rating scale (for example, were propositions that were rejected in the literature also reliably rejected, in the view of respondents, by their field?) and for this comparison the "one-sample t-test" was employed.

For evaluating the variability of views within a field, it was necessary to conduct similar comparisons between conventional measures of variability (standard deviations) rather than between means (measures of central tendency). To test these differences, absolute deviation scores were calculated and these deviation scores were subjected to the same analyses described in the preceding paragraph. (See Levene (14)).

In making comparisons between the responses to an item by the two groups of respondents (examiners and scientists), "independent sample t-tests" were employed. (Note that these tests did not depend on the assumption of homogeneity of variance, so degrees of freedom typically do not consist of whole numbers.)

Though significance tests are necessary to document the reliability of apparent differences, the average reader need not be unduly concerned about the statistical tests themselves. The means and standard deviations are provided for all of the comparisons, and the article's language is careful to speak of "differences" only when they were found to be different by a statistical test at $p < .05$. Where the statistical test yielded a difference at $.05 < p < .10$, the results are referred to as no more than "marginally" significant. Otherwise, no differences are asserted. Consequently, the reader who wishes to rely on the verbal description of the findings can do so without missing the meaning of the findings.

Results

Differences in Mean Ratings: Different Principles Have Different Degrees of Acceptance Among Forensic Document Examiners

The survey items are provided in Table 1, arranged from those most firmly endorsed by forensic document examiners to those least strongly endorsed, or rejected, by the field. (The item numbers correspond to the item numbers on the basic form of the questionnaire.) Not surprisingly, different propositions received different ratings. (An overall repeated measures analysis of variance of the ten items is highly significant. $F(9,71) = 21.22$, $p < .001$). The patterns of endorsement and rejection of the propositions are of interest.

The top three items—Propositions 7, 8, and 9 (asserting, respectively, that from static writing the speed of its creation, direction of strokes, and starts and stops, can be inferred)—pertain to the process of examination and inference about the writer's motor behavior.

TABLE 2—*Post hoc significance tests of means and absolute deviations (bonferroni post hoc analyses).*

Comparisons of Item Means

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10
Q1			.035	<.001		<.001				
Q2			.096	<.001		<.001				
Q3	.001	.003		.069		.001	.026	.018	.008	
Q4			.003		<.001		<.001	<.001	<.001	.030
Q5						<.001				
Q6			.009				<.001	<.001	<.001	.001
Q7			<.001		.071					
Q8			<.001		.012					.095
Q9			<.001		.001					.049
Q10	.003	.011		.009		.028	.002	<.001	<.001	

Comparisons of Item Variability

Note: When comparisons were not significant at $p < .10$, cells remain blank.

Significance test results of comparisons of item means are above the diagonal; significance test results of comparisons of item absolute deviations are below the diagonal.

The next three items include two foundational principles that are said to make handwriting identification possible (Proposition 1 (asserting that no two people write alike and therefore writing is individualizable) and Proposition 5 (asserting that inter-writer variation exceeds intra-writer variation)).

Though ratings of these five items do not differ significantly from each other, the ratings of Propositions 7, 8 and 9 each do not differ from "well accepted as true" (Item 7: $t(10) = 1.00$, $p = .341$; Item 8: $t(9) = 1.00$, $p = .343$; Item 9: $\bar{x} = 1.00$, variance $= 0$), whereas the ratings of Propositions 1 and 5 differ marginally significantly from "well accepted as true" (Item 1: $t(11) = 1.92$, $p = .082$; Item 5: $t(8) = 2.00$, $p = .081$). These findings suggest that forensic document examiners regard with somewhat greater confidence the first order inferences they draw in the process of performing their examinations and with somewhat less confidence the underlying principles that are needed to make higher order inferences about genuineness or identity.

Proposition 2 (asserting that forgeries are generally detectable as non-genuine) concerns a task that forensic document examiners are called upon to perform relatively often and, based on their theory of the psychological basis for the production of signatures, should be one of their most dependable skills. But the ability to reliably make such determinations is viewed with something less than consensus. (The ratings of Item 2 differ significantly from "well accepted as true" at $t(10) = 2.39$, $p = .038$)

Several of the propositions are generally regarded in the handwriting literature as being false: Proposition 3 (asserting that writing is more effectively evaluated through an examination of the whole rather than its elements), Proposition 4 (asserting that the author of a forgery usually can be identified), and Proposition 6 (asserting that a single known signature is sufficient for determining the genuineness of a questioned signature). These propositions therefore would be expected to, and do, receive ratings close to the regarded-by-the-field-as-false end of the rating scale (see Table 1). Moreover, the ratings of these propositions differ significantly from the ratings of many other propositions (see top half of Table 2). The ratings of Proposition 3, however, differ significantly from the ratings of Proposition 6. Respondents indicate that Proposition 3 is viewed by their field with more uncertainty than the other propositions asserted by the literature to be false. Indeed, the ratings of Item 3 do not differ significantly from "divided opinion; not established to be either true or false" ($t(7) = .403$, $p = .699$). Indeed, only one of these propositions is rated so that it is not significantly different from "well accepted as false," that being Item 6 ($t(11) = 1.30$, $p = .220$). The two other "false" items both stand significantly away from that end of the scale (Item 4: $t(10) = 4.50$, $p = .001$; Item 3: $t(7) = 3.63$, $p = .008$.).

Finally, note that the ratings of one of the assertedly false items (Proposition 3) and one of the assertedly true items (Proposition 5) are both viewed by the field with so little consensus (the one not seen as so clearly false, the other not as so clearly true) that the two items do not differ significantly from each other (see top half of Table 2) (the details from the post hoc analysis are: mean difference $= .75$, SE $= .455$, $p > .999$).

Differences in Variability: Extent of Forensic Document Examiner Agreement about their Field's Acceptance of Certain Principles

In the preceding section, we compared the mean ratings of the various propositions. In this section we look at variability in order to learn about the degree to which views of the field's beliefs were held in common by the respondent examiners. These differences are reflected in the standard deviations contained in Table 1. Overall, ratings of some propositions showed significantly less variability than others ($F(9,70) = 7.07$, $p < .001$), that is, there was more consensus about some propositions and less consensus about others.

Concerning propositions regarded by the field's literature as being true, in decreasing order of consensus, the greatest agreement in perceptions of their field's views obtained for the following propositions: 9 (asserting that from static writing starts and stops can be inferred), 8 (asserting that from static writing the direction of stroke can be inferred), 1 (asserting that no two people write alike and therefore writing is individualizable), 2 (asserting that forgeries are generally detectable as non-genuine), and 7 (asserting that from static writing the speed of its creation can be inferred). (See Table 1.)

Proposition 5 (asserting that inter-writer variation far exceeds intra-writer variation) is not among those on which respondents saw a clear consensus in their field. Indeed, the variability of ratings of Proposition 5 is significantly greater than the variability of Propositions 8 and 9 and marginally significantly greater than the variability of Proposition 7. (See lower half of Table 2.)

Concerning propositions regarded by the field's literature as being false, the greatest agreement in perceptions of their field's views is found for Propositions 4 (asserting that the author of a forgery usually can be identified) and 6 (asserting that a single known signature is sufficient for determining genuineness of a questioned signature). (See Table 1.)

Those propositions about which respondents were in the least agreement on what views their field held were propositions: 10 (asserting that from static writing the muscle movements used to create the writing can be inferred), 3 (asserting, contrary to the literature, that writing is most effectively evaluated by an examination of the whole rather than its elements), and 5 (inter-writer variation exceeds intra-writer variation). (See Table 1.) Views of the field's consensus on Proposition 10 are significantly more varied than views of the field's consensus on Propositions 1, 2, 4, 6, 7, 8, and 9. Views of the field's consensus concerning Proposition 3 are significantly more varied than views of Propositions 4, 6, 7, 8, and 9. And, as noted above, respondents reflected more uncertainty about Proposition 5 than they did about a number of other propositions. (See lower half of Table 2.)

Comparisons Between Examiners and Scientists

Finally, we turn from comparisons *among* forensic document examiners to comparisons *between* forensic document examiners and handwriting scientists.

The first point to note is that scientists were more likely than examiners to believe that insufficient knowledge existed about some propositions for the validity of the proposition to be determined. Scientists responded by saying a proposition was "not sufficiently well studied to have become established as true or false" 12.3% of the time, whereas examiners did so only 3.8% of the time (a difference significant at $t(251.51) = 2.60$, $p = .01$).

Similarly, turning to comparisons of the ratings given the various propositions, examiners generally were more sure of the truth or falsity of a proposition than scientists were. As Table 1 shows, on five of the propositions, examiners reported their field being more sure (than scientists reported their fields were) that a proposition was true. These five propositions were: 1 (asserting that no two people write alike and therefore writing is individualizable) ($t(12.55) = 3.59$, $p = .003$); 2 (asserting that forgeries are generally detectable as non-genuine) ($t(17.45) = 4.42$, $p < .001$); 7 (asserting that from static writing the speed of its creation can be inferred) ($t(18.28) = 4.95$, $p < .001$); 8 (asserting that from static writing the direction of a stroke can be inferred) ($t(17.24) = 2.58$, $p = .019$); 9 (asserting that from static writing starts and stops can be inferred) ($t(14) = 4.36$, $p = .001$).

For a sixth proposition (Item 10), asserting that from static writing the muscle movements used to create the writing can be inferred, examiners thought their field was quite divided over whether or not they had this capability, while scientists believed that their fields were more sure that such inferences could not be drawn (a marginally significant difference: $t(6.24) = 2.00$, $p < .091$).

Examiners reported their field to be significantly more sure than scientists' fields that Proposition 4 (asserting that the author of a forgery usually can be identified) is false ($t(18.76) = 3.58$, $p = .002$). Similarly, examiners were marginally significantly more sure that their field believed that a single known exemplar is an insufficient basis for making an accurate identification (Item 6) than scientists were about their fields (although scientists leaned in the same direction, as indicated by the mean ratings) ($t(20.77) = 1.96$, $p = .063$).

Discussion

The focus of this inquiry was not on the actual validity or lack of validity of any of the propositions employed in the study, but rather on what members of several fields professionally and scientifically concerned with the nature of handwriting understood their field's views to be regarding those propositions and the extent of their field's consensus or dissensus.

From the perspective of a "narrow" general acceptance test, within the field of forensic handwriting identification itself, the propositions endorsed most clearly involve the process of examining writing (inferences that can be drawn from the static trace), while the foundational principles that are said to make handwriting individualization possible were supported somewhat less clearly. The propositions concerning the dependability with which forgeries can be detected and the ratio of inter-writer to intra-writer variability both showed a marginally significant departure from being "well accepted as true." Indeed, ratings of the latter were not significantly different from ratings of one of the propositions regarded by the literature as false.

At the other end of the spectrum, propositions that the field's literature rejects as untrue are rated as less acceptable than principles that the forensic document examination literature endorses as true. At the same time, however, only one of these "false" items was clearly viewed as being "well accepted as false" (that a single known signature is sufficient for determining genuineness). There appears to be less consensus that the other "false" propositions (that the author of a forgery usually can be identified, that it is more effective to examine a writing as a whole than by its constituent elements) are false. Indeed, the latter proposition is viewed as something on which the field holds widely divergent views.

The variability statistics lend confirmation. Respondents were in the greatest agreement about their field's consensus with respect to examination process variables, although they manifested considerable uncertainty about their field's views of two such propositions (whether it is best to examine writing as a whole or by its constituent elements, whether muscle movements can be inferred from the static trace).

From the perspective of a "broad" general acceptance test (comparing the views of forensic document examiners to those of other relevant groups, such as various kinds of scientists who do research on handwriting), we see a number of differences between the beliefs about handwriting held by examiners and those held by scientists. First, forensic document examiners were more likely than handwriting scientists to believe that sufficient knowledge existed on which to base conclusions about most of these propositions. Moreover, respondents indicate that examiners believe in the truth or

falsity of most of the propositions to a greater degree than scientists do. Of the ten propositions, the views of the two groups differed significantly on six and marginally significantly on two more. Interestingly, one of the propositions on which the two groups did not differ was the extent of consensus in the two fields regarding whether inter-writer variation exceeds intra-writer variation; both report "most people in my field regard this as true, but a respectable minority believe it to be false."

Limitations of the Study

A single study rarely, if ever, answers a question definitively. That is as true of the present study as it is of any other. Several weaknesses of this study should be noted.

Two closely related issues are sample size and response rate. Although the study's sample size is reasonable, for a study of this type it would have been desirable for the sample size to be still larger. A larger sample would have increased the power of the statistical tests (thereby reducing the risk of Type II error). Where significant differences were not found, that could be due to the small sample size rather than a true lack of differences. Thus, examiners may draw more distinctions between propositions endorsed and rejected by their field's literature than we were able to detect. Which way that cuts for any given proposition depends on the particular comparison being made. On the other hand, more differences may exist between the beliefs of examiners and scientists than we found in this study. However, where significant differences were obtained in spite of the small sample size and decreased statistical power, one can be all the more confidant that the differences are real.

Parenthetically, we might note that the relatively small sample size and low response rate of forensic document examiners was due to an organized effort that arose within the examiner group. Initially, members of both groups, examiners and scientists, submitted their responses to the questionnaire we had sent them at about the same pace. Before long, however, a few members of the examiner group contacted the researchers, making the extraordinary request that they (the intervening examiners) be allowed to become advisers to or collaborators in the study. When those entreaties were refused, they urged their examiner colleagues to cease further participation in the study. No such organized refusal arose among the group of scientists. These events no doubt account for most of the difference in response rates between the examiners and the scientists.

The real problem with a low response rate, of course, is not size but representativeness. That is, the potential problem is that a sample may fail to be representative of the population from which it was drawn. However, an elementary principle of sampling is that the more homogenous the population being sampled, the smaller the sample needs to be in order to be representative. Because of its procedures, its policies, and its purposes (8), the ASQDE constitutes an unusually homogeneous population with respect to the data being gathered by this survey. Indeed, it would be quite a contradiction to say, on the one hand, that membership in the ASQDE is an assurance of *uniformly* high quality of knowledge and skill (8)—so much so that a court is expected to give great weight to what a single member has to say—and at the same time argue that a dozen of them cannot be depended upon to accurately reflect the views of their association on questions central to their expertise. In short, the use of a single homogeneous organization of this type makes it much less likely that representativeness is a problem.

Moreover, one of the few demographic measures we did collect—the number of years of professional involvement in handwriting examination—showed that our document examiner sample *averaged* 26 years in the field of document examination, suggesting the respondents had no lack of experience in and familiarity with their field.

A related issue is that not all questions are answered by all respondents. One important reason for a non-answer is that when respondents indicated that they did not think that enough knowledge existed for the field to know one way or the other, no rating of their field's views was given. Any given question was answered by as many as 96% of handwriting scientists, as few as 71%, and on average by 88%. Similarly, any given question was answered by as many as 92% of forensic document examiners, as few as 62%, and on average by 81%. Since the significance tests take into account the number of responses, some items were tested with more statistical power than others, and that means that more differences probably exist than were capable of being detected by our study.

Another limitation to note is that the propositions contained in the questionnaire consisted of a sample of propositions from a much larger domain of principles on which the enterprise of handwriting identification relies. Although these were selected to represent a range of propositions—some asserted by the field's literature to be true, others false; some relating to the process of examination, others to the foundational principles on which examination and identification depend—future research sampling different portions of the domain might arrive at different results.

Finally, on reflection, we note that Proposition 6 suffers from a potentially significant and avoidable ambiguity. The proposition should have been further qualified by specifying that it refers to simulation and not to freehand forgery without a model. The latter kind of forgery is likely to show so much disparity between the questioned signature and the known exemplar that any comparison of the two will strongly suggest that the questioned one is not genuine. Only one forensic document examiner commented on the ambiguity, and that examiner declined to respond to this item. Nevertheless, some of the variation in responses of examiners to this item may be the result of this item's ambiguity.

Implications for Admissibility

In jurisdictions where "general acceptance" is important to an admissibility decision, far better evidence can be offered to courts than the statements of a witness in a case asserting that whatever she or he was about to testify to is generally accepted in her or his field, as such proffered witnesses inevitably do. The present study illustrates a methodology for more rigorous assessment of the extent of acceptance (or rejection) of propositions both within and between fields. Though not the first such study of this kind (18, Chapter 7, Sec. 1.1.1), others could usefully be conducted in order to provide a source of systematic information about general acceptance in various fields, something courts rarely have the benefit of.

Turning to the implications for the field of handwriting identification, though it has long been assumed that general acceptance, if nothing else, would support the admissibility of handwriting examiner expert opinion, this first systematic study of the issue raises some doubts, at least for those propositions examined. First, forensic document examiners and handwriting scientists appear not to agree on most of the propositions. Who is correct and who is incorrect on these points of disagreement is beyond the scope of the present study, and ultimately can be resolved only through empirical research on the phenomena in question, the findings of which both experts and scientists heed. More surprising is the finding of a lack of consensus on some important propositions within the handwriting examination field itself. This is all the more

surprising considering that the examiner sample was drawn from a single group of examiners who are said to share a similar background of training, belief, and often occupational setting (8). Indeed, the methodological deck could hardly have been stacked more in favor of finding consensus within the handwriting examination field.

As to courts using *Daubert*, the criteria for admission of forensic handwriting expert testimony are of course more numerous and more demanding than mere general acceptance. But, to the extent that general acceptance can make up for some deficiencies in empirical testing or unknown or unsatisfactory error rates, the present research finds both some strengths and some weaknesses.

References

1. *Frye v. United States*, 293 F. 1013 (D.C. Cir. 1923).
2. Faigman D, Kaye DH, Saks MJ, Sanders J, editors. Modern scientific evidence: the law and science of expert testimony. 2nd ed. St. Paul: West Group, 2002, Supplements 2003, 2004.
3. Saks MJ. Merlin and Solomon: lessons from the law's formative encounters with forensic identification science. Hastings Law J1998;49:1069–141.
4. *Daubert v. Merrell Dow Pharmaceuticals*, 509 U.S. 579 (1993).
5. *United States v. Starzecpyzel*, 880 F. Supp. 1027 (S.D.N.Y.1995).
6. *Kumho Tire Co. v. Carmichael*, 526 U.S. 137 (1999).
7. Committee on Evaluation of Sound Spectrograms, Assembly of Social and Behavioral Sciences. On the theory and practice of voice identification. Washington, DC: National Academy of Sciences, 1979.
8. Moenssens AM. Handwriting identification evidence in the post-*Daubert* world. Univ Missouri Kansas City Law Rev 1998;66:251–343.
9. Osborn AS. Questioned documents. Albany, NY: Boyd, 1910.
10. Hilton O. Scientific examination of questioned documents, revised ed. New York: Elsevier, 1982. (Reprinted: Boca Raton, FL: CRC Press, 1993).
11. Conway JVP. Evidential documents. Springfield, IL: Thomas, 1959.
12. Harrison WR. Suspect documents: their scientific examination. New York: Praeger, 1958.
13. Ellen D. The scientific examination of documents: methods and techniques. New York: Halsted Press, 1989.
14. Levene H. Robust tests for equality of variances. In: Olkin I, editor. Contributions to probability and statistics. Stanford, CA: Stanford U. Pr., 1960;278–92.
15. Hilton O. Can the forger be identified from his handwriting? 43 J Crim L Criminology Police Sci 1952–53;43:547–55.
16. Forensic Sciences Foundation/Collaborative Testing Services, Crime laboratory proficiency testing program – questioned documents analysis report No. 84–7, 1984.
17. Conway JVP. The identification of handprinting. J Crim Law Criminology Police Sci 1955;45:605–12.
18. Diamond SS. Survey research. In: Faigman D, Kaye DH, Saks MJ, Sanders J, editors. Modern scientific evidence: the law and science of expert testimony. 2nd ed. St. Paul: West Group, 2002;Vol. I:286–330.

Additional information and reprint requests:
Michael J. Saks, Ph.D.
Arizona State University
Center for the Study of Law Science & Technology
College of Law & Department of Psychology
P.O. Box 877906
Tempe, AZ 85287-7906