# The shaky criticism of forensic handwriting analysis

Reinoud Stoel, Charles Berger, Elisa van den Heuvel and Wil Fagel[1]

Handwriting analysis is a complicated form of comparative forensic examination with an important subjective component. The underlying basic principle is not that a handwriting examiner woud be able to link handwriting to a unique source. The examination is aimed at an assessment of the evidential value.

In his article "The shaky foundation of forensic handwriting analysis" [2] professor Merckelbach writes extensively on a number of issues that would demonstrate a weak foundation. Merckelbach argues that the analysis of handwriting is flawed and that the judge who decides call upon handwriting analysis anyway would do better to insist on a evidence line-up. His examples come from distant lands and distant pasts, and are therefore not always relevant to current practice at the Netherlands Forensic Institute (NFI). In this article we would like to discuss the problems that do exist, and their solutions.

**Introduction**

Recently Merckelbach criticized the foundations of forensic handwriting analysis, calling into question the reliability and validity, and thus the probative value of the results of comparative handwriting analysis. The adduced examples of mistakes that Merckelbach cited from the United States and France in support of his criticism are however from a century ago. Moreover, these particular failures are often due to other (ethical) problems rather than due to a lack of scientific basis. Also, the examples Merckelbach gave were sometimes not well understood. In the case of the Hitler diaries, for example, handwriting examiners had concluded that the handwriting in the diaries corresponded to the reference material, which - they were told - was written by Hitler. The similarities between the handwriting in the Hitler Diaries and

the reference material in this historic case were not misperceived, but their conclusion turned out to be misleading for another reason. The mistake in that case was that - as it turned out - the forger had also produced part of the reference material. So the experts in a sense were correct when they claimed that the handwriting (diaries and reference material) had a common author, even if it was not Hitler. But it would be too easy and unfair to dismiss the criticism on the basis of forensic handwriting analysis for those reasons.

Merckelbach then focuses on a previously supposed pillar of comparative handwriting analysis, namely that each handwriting is unique and that a handwriting examiner would be able to recognize that uniqueness. This is not the basis of current practice at the NFI. Not because it can't be proven that each piece of handwriting is unique. Indeed, no two objects are identical, because if they were they would not be two but one object. The problem is that several pieces of handwriting by the same author will also never be identical. We shall later return to what is the basis for handwriting analysis. The article by Merckelbach ends somewhat abruptly with the (unfortunately) not worked-out suggestion to perform an evidence line-up, apparently as a solution to the supposed problems.

In this paper we will outline Merckelbach's points of criticism regarding forensic handwriting analysis, and show that solutions exist. It will become clear that a number of issues were confused by Merckelbach. After this, we will discuss the solution proposed by Merckelbach to deal with "the obscuring effects of a priori expectations," namely the evidence line-up. It will become clear that the proposed solution of the evidence line-up is not is not quite as helpful as is pretended.

**Basic principles and logic**

We reported earlier that mentioned pillars of handwriting examination (that each handwriting is unique and that a handwriting examiner would be able to recognize its unique source, also known as 'discernable uniqueness') are not the basis of current practice in the NFI. Since 2010 the handwriting examiners also do not give an opinion on the likelihood of whether or not the same author wrote two pieces of handwriting, because that likelihood also depends on information other than just the pieces of handwriting themselves.

The handwriting expert examines and compares the disputed handwriting with reference material from the suspect and assesses the evidential value of the observed

similarities and differences in the light of the hypotheses. The hypotheses are formulated on the basis of the request and the relevant information in the case. In many cases, the hypotheses are formulated as: 1) the suspect is the writer of the disputed handwriting, versus 2) any other person is the writer of the disputed handwriting. A paper about the principles underlying this method, also known as the 'Likelihood Ratio' method was previously published in this magazine [3]. The core of the Likelihood Ratio method is that the (handwriting) expert is limited to an opinion on the evidential value of his findings in light of the hypotheses. The evidential value, or simply the LR (Likelihood Ratio) is defined as the ratio of two probabilities: the probability of the findings if the suspect was the writer, and the probability of the findings if a random other person was the writer of the disputed handwriting. Extreme cases aside, no opinion is given on the likelihood of the hypotheses. That probability depends also on other information, and assessing it is a task of the court.

In a case with for example an anonymous threatening letter, the handwriting expert can say nothing about the likelihood that the defendant wrote the threatening letter, even if the judge would like him to give an opinion on that. To do this, the handwriting expert would need evaluate information that is clearly outside of his area of expertise. Because tactical information and other evidence contribute to this probability. The handwriting expert can and will therefore only comment on the ratio of the probability of his findings under the relevant hypotheses. For the hypothesis that the accused was the writer, that probability depends heavily on the natural variation of his handwriting, for the hypothesis of any other writer, it's the rarity of the similarities and differences observed [4].

**Subjective observations and objectification**

Although in this paper we will mainly talk about handwriting analysis, the above criticisms apply much broader. An important aim of forensic science is objectifying analysis, comparison and interpretation by reducing the subjective component therein. This is a major challenge for forensic science. First, the relevant features are precisely defined and rendered objectively measurable. Then, a comparison algorithm is defined which yields the degree of similarity (or difference) between two sets of features. To objectify the interpretation of results it must then be determined what degree of similarity (or difference) is expected when comparing material from the same source, and material from different sources.

Forensic scientists from almost all areas of expertise are working hard on that challenge, but for some forensic disciplines results can be expected much earlier than for others. Forensic handwriting analysis is one of the hardest types of examination to objectify. Merckelbach understandably focuses his criticism on comparative handwriting analysis because this type of examination is largely subjective. But subjective is not equivalent to unreliable: a subjective assessment can be very reliable. With subjective conclusions, we refer to conclusions that are not (only) based on hard data, but (also) on an assessment by the expert based on his knowledge and experience.

**Proficiency testing**

Subjective judgments in handwriting analysis are frequently studied in so-called proficiency tests. This means the experts are tested with many sets of manuscripts, each of which may or may not come from the same writer. Unlike in real case work, the actual writer is of course known in such collaborative tests. Therefore, the expert is forced to choose between "same writer," "different writers" and "no opinion" in order to directly compare his results with the known correct answers. Another difference is that researchers will often try hard to involve difficult cases in the collaborative test, which means the examined material is often more complex than in typical real cases.

The goal of a collaborative test is to compare the proficiency of examiners and institutes. The aim is to demonstrate competence and consistency in the judgments. The goal of collaborative testing is not to assign a margin of error to categorical judgments (such as "this signature is a forgery"), which are not given in real cases. In the literature an error rate of around 7% based on such proficiency tests is often mentioned, and so does Merckelbach. Such a percentage is saying very little about the error rate in practice, in which no categorical conclusions are given, and in which the examination is always 'shadowed' by a second examiner. Furthermore, such a percentage says even less about the scientific basis of comparative handwriting analysis.

**Unwanted effects**

Cognitive processes play an important role in human behavior, in human observations and reasoning. These processes allow us to process large amounts of

information, and draw conclusions based on it, even if that information is sometimes ambiguous and incomplete [5]. Precisely these processes, however, also give rise to specific vulnerabilities which may lead to selectivity and distortion of information.

An example in forensics is the ‘context effect’ [6,7,8]. This term indicates that the results of a forensic examination can be influenced by the circumstances in which the examination is conducted, and in particular by the information known to the examiner [9]. In the forensic literature, the term ‘confirmation bias’ is also regularly used. Roughly three types of such information can be distinguished.

First is the information on the base rate which is independent of the specific investigation, but which can have an effect on the expectations of the researcher. The fact that in the past, for example, 95 out of 100 documents were found to be falsified, may have an impact. The examiner potentially sees more value in information that supports the expected conclusion. That influence is undesirable because base rates should have no effect on the probative value of the examination itself.

Then there is the domain-irrelevant case information. Thus, the fact that a suspect for instance was arrested three times previously for a similar offense and in this case has confessed forging a document, is relevant to a court but for the handwriting examiner it is domain-irrelevant information. Even if a professional examiner will not automatically go along with the suggestion, and it can also achieve the opposite effect, he should not be exposed to this information. An illustration of the effect of domain-irrelevant information is provided by an experiment by Dror et al. [10]. Five professional fingerprinting examiners were given a finger mark (trace) and a fingerprint (reference). They had examined them much earlier in a real case, in which they had concluded that they came from the same finger. This time, finger mark and fingerprint were presented as if it were a new case, and in a context that strongly suggested they did not come from the same finger. Four of the five experts subsequently drew a different conclusion than they had previously drawn from the same mark and print.

A third form of influence may lie in the way the examination is carried out. If the disputed material is examined simultaneously with the reference material, post hoc target shifting can take place. Of post hoc target shifting “occurs when deemed relevant aspects of the disputed material be influenced by what the researcher has seen in the reference material. Thompson [11] for example showed how knowledge of the DNA profile of a suspect may affect the interpretation of a disputed (partial and

complex) DNA profile. With that knowledge comes the danger that the expert will 'see' characteristics of the reference profile in the trace profile that otherwise he would not have seen.

It is clear that such forms of influence would jeopardize the accuracy of conclusions, because the conclusion is not purely based on the relevant evidence. The question arises to what extent these effects actually occur in practice. Dror and Cole [12] give an overview of studies of context effects in case work. The few studies carried out by forensic scientists and behavioral scientists partially contradict each other. Hall and Player [13] for example, come to the conclusion that things are not so bad, while other studies have indeed shown effects. Beforehand it is not clear whether or not a context effect would occur in a particular case. It is therefore better to arrange the forensic examination - where feasible - in such a way that the probability of undue influence is minimized. The methods that can prevent or reduce context effects depend on the type of information that underlies these effects.

The effect of domain-irrelevant information can be dealt with by keeping away non-relevant information as much as possible from the expert carrying out the examination. This is partly possible by the creation of a stepwise procedure, in which the expert who coordinates the case is not the expert that carries out the examination. Unfortunately, such a procedure can sometimes not keep away all domain-irrelevant information. Because the context may derive from the content of the disputed material itself, as often occurs in areas such as handwriting and speech analysis. Adapting the examination methodology can keep the researcher blind to the context of the case. Thus, in forensic handwriting examination a selection from the disputed material can sometimes be made such that the handwriting examiner can not understand from the written texts what is the context of the case.

In addition, the expert should have no initial knowledge of the reference material in the case, to avoid *post hoc target shifting*. Böttcher [14] and Froentjes [15] suggested already in the fifties and sixties of last century that the disputed material is to be examined first, and then the reference material. What the expert sees in the disputed material should not be influenced by what he saw in the reference material. The overarching term for keeping this type of information away is 'blind testing', and a well-executed blind test can be expected to reduce these context effects to a minimum.

Blind testing does not prevent the effect of base rate information because it is independent of the context of the case. A possible solution is to add a (large) number of fake cases in which the correct conclusion is not that the hitherto generally correct conclusion. Although effective, it is clear that this solution is not very workable because it is difficult to create many realistic cases where the expert is actually not aware that the case is a fake. Currently such a study is carried out at the NFI in conjunction with the University of Amsterdam in the area of arms and ammunition. Over the next two years - at unknown times - cases will be coming in which will later be revealed to be fake cases. The results of this study will also allow an estimate of how feasible the addition of fake cases is, and whether the study succeeds in creating realistic cases.

**The evidence line-up**

In the literature the evidence line-up is fairly regularly put forward as a remedy for the obscuring effects of the context. This is a procedure in which the disputed material is compared with a line-up of reference material. The expert knows which is the challenged material, and is to compare it with several other items, some of which should come from a suspect, with several other items added which are to a certain extent similar to the disputed material. These added items are called fillers or foil specimens.

The evidence line-up is sometimes confused with blind testing, but there is definitely a difference. Blind testing is a procedure in which the expert assesses the disputed material without knowledge of the domain-irrelevant information or of the reference material. An evidence line-up however, is a procedure in which the expert will be presented with several items, the source of which is unknown to him. This procedure is mentioned regularly in the literature, and bears some resemblance to the famous Oslo confrontation, or eye witness confrontation [16].

In the article by Merckelbach the evidence line-up is also put forward as the remedy against undesirable influences of irrelevant information on the expert opinion. Superficially, the evidence line-up indeed seems to have much to offer, but a critical appraisal shows differently. Many of the recommendations regarding the use of the evidence line-up to prevent the undesirable effect of e.g. prior expectations and context are traced back to the authoritative publication of Risinger et al., (see note 7) but in 1984 and 1987 Miller [17] already wrote on its application in handwriting and

hair examination. After Risinger et al many people stressed the importance of the evidence line-up without actually considering how to set them up and what the outcome would be.

The use of the evidence line-up seems particularly based on the impression of validity or face validity [18], which means that the method seems valid, but not necessarily is. High face validity of a method is a nice but unnecessary property of a method. It is also one the most subjective forms of validity that exist, and offers no guarantee of actual validity. Empirical research into the evidence line-up is virtually absent and the use of the evidence line-up in case work is problematic. Risinger et al seem well aware of this - unlike many others - where they write:

> *'Proper evidence lineups present some nontrivial problems of design, requiring the Evidence and Quality Control Officer both to determine what would constitute appropriately similar foil specimens and to arrange to obtain them. This process would obviously be easier for some types of examinations than for others. Unfortunately, it may often be most difficult precisely where it is most needed, in those areas, such as handwriting identification, with the least instrumentation and greatest subjectivity.'*

**Sham**

It all seems so simple. In addition to the disputed handwriting and the reference handwriting of X (a. and b. in Merckelbach's terms) the examiner also asks for group c. with simulations of handwriting and signatures of X:

> *'The documents in a., b. and c. are presented to the expert as a random series, after which he searches for similarities and differences. Such a procedure provides the beginning of a rigorous test. It allows one to catch the expert making mistakes. If the expert demonstrates he can flawlessly pilot through the documents, this underlines his expertise'*
>
> *(Merckelbach2, p. 416).*

Although this procedure seems watertight, its main shortcoming is that no criteria are available for the selection of the documents in group c. (these are the fillers). And it is this choice which largely determines the results of the examination.

If e.g. you use random handwritings as fillers in the line-up, it will add little to the examination. Its usefulness in court is then mainly based on the aforementioned face validity. If you let someone imitate the disputed handwriting - as proposed by Merckelbach - you make it more difficult for the examiner, but you are especially testing the talent of the imitator for imitating handwriting.

A more interesting option would be to let a second expert look for the most similar handwritings in a large collection. But even then you are testing several things simultaneously: how well does the second expert perform relative to the first, and how large is the collection that he can use. So it is not clear that an evidence line-up offers any guarantee for a better assessment of the probative value of the evidence. The evidence line-up is an attempt to solve several problems simultaneously, but fails because too many variables are varied at the same time. It is better to test evidence and examiner(s) separately.

**Conclusion**

Handwriting analysis is a complex type of forensic comparative examination, with a significant subjective component. The aim is not so much to say who wrote the disputed handwriting, but to describe what the observations of the disputed and reference material are worth to answer that question. The basic underlying principle is not that a handwriting examiner would be able to identify the unique source of the handwriting (‘discernable uniqueness’). The examination aims to assess the evidential value (or ‘likelihood ratio’, ‘diagnostic value’). This is the same sound scientific methodology that is applied in other forensic examinations, such as in DNA testing.

One important difference, however, lies in the difficulty of objectively defining features of handwriting and comparing them. This does not make the scientific basis shaky, but it can lead to susceptibility to undue influence. Depending on the type of influencing information some solutions exist that counteract such influences. The evidence line-up is maybe an appealing, but certainly not the most appropriate method because of the fundamental problems we have discussed in this article. To keep away irrelevant case information, the addition of fake cases and blind testing will therefore lead to better results than an evidence line-up.

## References


[1] Dr. R.D. Stoel, dr. ir. C.E.H. Berger, dr. C.E. van den Heuvel en drs. W.P.F. Fagel all work for the Netherlands Forensic Institute (NFI).

[2] Merckelbach, 'De wankele basis van de forensische handschriftkunde', NJB 2010, 320, afl. 7, p. 413-416.

[3] Berger, 'Criminalistiek is terugredene- ren', NJB 2010, 646, afl. 13, p. 784-789.

[4] When it possibly concerns non-spontaneous handwriting, other hypotheses such as imitation or distorted handwriting need to be considered.

[5] Dror, 'How can Francis Bacon help foren¬sic science? The four idols of human biases', Jurimetrics: The Journal of Law, Science, and Technology 2009, 50, p. 93-110.

[6] Saks, Risinger, Rosenthal & Thompson, 'Context effects in forensic science: a review and application of the science of science to crime laboratory practice in the United States', Science and Justice 2003, 43, p. 77-90.

[7] Risinger, Saks, Thompson & Rosenthal, 'The Daubert/Kumbo implications of observer effects in forensic science: hidden problems of expectation and suggestion', California Law Review 2002, 90, p. 1-56.

[8] Broeders, 'De blinde onderzoeker', Trema 2006, 6, p. 237-243.

[9] Thompson, 'Painting the target around the matching profile: the Texas sharp¬shooter fallacy in forensic DNA interpre¬tation', Law, Probability and Risk 2009, 8, p. 257-276.

[10] Dror, Charlton & Peron, 'Contextual information renders experts vulnerable to making erroneous identifications', Forensic Science International 2006, 156, p. 74-78.

[11] Thompson, 'Painting the target around the matching profile: the Texas sharpshooter fallacy in forensic DNA interpretation', Law, Probability and Risk 2009, 8, p. 257-276; see also: Meulenbroek, Kloosterman & De Blaeij, 'Richtlijnen borgen onbevooroordeeld DNA-onderzoek', Expertise en recht 2009, p. 119-129.

[12] Dror & Cole, ‘The vision in ‘blind’ justice: Expert perception, judgment and visual cognition in forensic pattern recognition’, Psychonomic Bulletin & Review 2010, 17, p. 161-167.

[13] Hall & Player, ‘Will the introduction of an emotional context affect fingerprint analysis and decision-making?’, Forensic Science International 2008, 181, p. 36-39.

[14] Böttcher, ‘Theorie en practijk van de gerechtelijke schriftvergelijking’, Tijdschrift voor Strafrecht 1954, LXIII, p. 77-131.

[15] Froentjes, ‘Schriftonderzoek en statistiek’, NJB 1969, 33, p. 821.

[16] The speech and audio analysis group of the NFI has developed a variant where the disputed material is also included in the line-up. So the expert does not even know which is the disputed material. This variant is like a blind cluster task for the expert, and differs in that the fillers come from the real material from the case itself. The method is applied at the NFI in each case which lends itself for it. The aim is to give the speech expert a subjective sense of whether or not he can properly use the available material. A summary of this variant can be found in: Broeders, ‘Vergelijkend spraakonderzoek’, in: Broeders & Muller (red.), Forensische wetenschap, Deventer: Kluwer 2008, p. 508-514. See also Van Koppen & Malsch, Het praktisch nut van de rechtspsychologie, Deventer: Kluwer 2008.

[17] Miller, ‘Bias among forensic document examiners: a need for procedural changes’, Journal of police science and administration 1984, 12, p. 407-411; Miller, ‘Procedural Bias in Forensic Science Examinations of Human Hair’, Law & Human Behavior 1987, 157, p. 159-162.

[18] Face validity is a term from the psychological literature. Drenth and Sijtsma (2006) describe face validity as a form of apparent evidence for validity. This impression, formed by a layman or an expert, does not need to be supported by empirical research.