Preprint of:

Morrison G.S. (2023). A single test pair does not a method validation make: A response to Kirchhübel et al. (2023) (Letter to the Editor). *Science & Justice*. https://doi.org/10.1016/j.scijus.2023.03.001

Preprint at: https://geoff-morrison.net/#response_to_Kirchhuebel_et_al_2023

A single test pair does not a method validation make: A response to Kirchhübel et al. (2023)

Author and affiliations:

Geoffrey Stewart Morrison *

Forensic Data Science Laboratory, Aston University, Birmingham, UK

Forensic Evaluation Ltd, Birmingham, UK

* e-mail: G.S. Morrison, geoff-morrison@forensic-evaluation.net

ORCID:

Geoffrey Stewart Morrison 0000-0001-8608-8207

Disclaimer:

All opinions expressed in the present response are those of the author, and, unless explicitly stated otherwise, should not be construed as representing the policies or positions of any organizations with which the author is associated.

Declaration of competing interest:

The author is Director and Forensic Consultant for Forensic Evaluation Ltd. Forensic Evaluation Ltd charges clients to conduct forensic voice comparison and provide written reports and oral testimony, including providing critiques of forensic-voice-comparison reports and testimony provided by others.

Acknowledgements:

The writing of this response was supported by Research England’s Expanding Excellence in England Fund as part of funding for the Aston Institute for Forensic Linguistics 2019–2024.

Abstract:

In terms of development of methods that are transparent and reproducible, that are intrinsically resistant to cognitive bias, that use the logically correct framework for interpretation of evidence (the likelihood-ratio framework), and that are empirically validated under casework conditions, forensic voice comparison may be one of the most progressive branches of forensic science. Forensic voice comparison is not, however, a monolithic branch of forensic science. The aforementioned progress has been made in the context of human-supervised-automatic methods. Unfortunately, many or most forensic-voice-comparison practitioners, who use auditory-acoustic-phonetic methods, have not joined in this progress. There are calls going back to the late 1960s for forensic-voice-comparison methods to be meaningfully validated under casework conditions, but many or most forensic-voice-comparison practitioners have still not heeded those calls. A recent example appears in Kirchhübel et al. (2023)

https://doi.org/10.1016/j.scijus.2023.01.004, which proposes a validation protocol in which the number of test pairs is only one. This does not constitute meaningful validation. If the cost and time necessary to conduct meaningful validations of auditory-acoustic-phonetic methods are such that it is not practical to conduct meaningful validations, then forensic voice comparison should not be performed using auditory-acoustic-phonetic methods.

Keywords:

Forensic voice comparison; Validation

Letter to Editor:

A keynote presentation at the European Academy of Forensic Science Conference 2022, published as Morrison [1], described a paradigm shift in evaluation of forensic evidence. According to Morrison [1], in the new paradigm, the forensic-data-science paradigm, the methods used adhere to the following principles: they are transparent and reproducible; they are intrinsically resistant to cognitive bias; they use the logically correct framework for interpretation of evidence (the likelihood-ratio framework); and they are empirically validated under casework conditions. With respect to the last principle, Morrison [1] noted that “practitioners in multiple branches of forensic science often claim that training and experience provide sufficient warrant for their conclusions (see Mnookin et al. [2]; Risinger [3]; President’s Council of Advisors on Science and Technology [4]; Morrison & Thompson [5]), and deny or obfuscate about the need for validation (see Cole [6]; Morrison [7]; President’s Council of Advisors on Science and Technology [4]; Koehler [8]; Morrison et al. [9]), or propose lax validation protocols that do not require demonstration of performance under casework conditions (see Morrison et al. [10], [11]).” Morrison [1] also noted, however, that “protocols for validating systems that output likelihood ratios have been developed, including metrics and graphics appropriate for representing the results of such validations (Meuwly [12]; Brümmer & du Preez [13]; Morrison [14]; Meuwly et al. [15]; Ramos et al. [16]; Morrison et al. [17]). Much of the latter work has been conducted in the context of forensic voice comparison, but the results are applicable across forensic science in general."

Between 2016 and 2019, in the context of a virtual special issue of *Speech Communication*,[1] training and test data reflecting the conditions of a real forensic-voice-comparison case were released, and, using those data, multiple laboratories validated multiple different human-supervised-automatic forensic-voice-comparison systems. The test data consisted of 111 same-speaker pairs of recordings (from 61 unique speakers), and 9720 different-speaker pairs of recordings (from 3660 unique pairs of speakers) (Morrison & Enzinger [18]). Detailed results were published in a series of papers in the virtual special issue, and a summary of the results was published in the virtual special issue's conclusion article (Morrison & Enzinger [19]). The data and answer key have since been made available for others to use. Weber et al. [20] used the data to conduct a benchmark validation of another human-supervised-automatic forensic-voice-comparison system, and Basu et al. [21] used a subset of the data to assess lay listeners' speaker-identification abilities and compare them with the performance of the latter human-supervised-automatic forensic-voice-comparison system. In 2019 and 2020, a group of researchers and practitioners collaborated on developing a *Consensus on validation of forensic voice comparison*. The published version, Morrison et al. [17], had 13 authors and an additional 7 supporters. In order to be able to write a progressive document, the scope of the *Consensus* was restricted to "validation of forensic-voice-comparison systems that are based on relevant data, quantitative measurements, and statistical models, and that output numeric likelihood ratios"; however, with minor wording changes the *Consensus* would be applicable to

[1] https://www.sciencedirect.com/journal/speech-communication/special-issue/10KTJHC7HNM

validating methods for assigning likelihood ratios in other approaches to forensic voice comparison or to validating methods for assigning likelihood ratios which address source-level hypotheses in other branches of forensic science. Since validation is a black-box exercise, the details of the method being validated are not relevant.

The foregoing may seem to suggest that forensic voice comparison is one of the most progressive branches of forensic science, but forensic voice comparison is not a monolithic branch of forensic science. Almost a decade ago, Morrison [7] wrote about and reiterated calls going back to the late 1960s for forensic-voice-comparison methods to be meaningfully validated under casework conditions. Unfortunately many or most forensic-voice-comparison practitioners have still not heeded those calls. A recent example appears in Kirchhübel et al. [22], which proposes a validation protocol in which the number of test pairs is only one.[2]

As previously mentioned, method validation is a black-box exercise. It may be possible to perform the same task using different methods, e.g., it is possible to perform forensic voice comparison using different methods that fall under the human-supervised-automatic approach (the aforementioned virtual special issue of *Speech Communication* provides examples) or different methods that fall under the auditory-acoustic-phonetic approach.[3] What validation does is test how well a particular method performs the task, e.g., the task of assessing the likelihood of obtaining the speech of

[2] Kirchhübel et al. [22] frames the task that the practitioner performed as a "proficiency test", but claims that this serves the purpose of method validation – the title of the article is: *What does method validation look like for forensic voice comparison by a human expert?*

[3] See Morrison & Zhang [23] for a recent description of different approaches to forensic voice comparison, including human-supervised-automatic and auditory-acoustic-phonetic approaches. Kirchhübel et al. [22] claims that the auditory-acoustic-phonetic approach "is the only admissible approach in UK jurisdictions for voice comparison analysis." The implication that the human-supervised-automatic approach would not in-principle be admissible is incorrect. For a review of relevant case law in England & Wales and in Northern Ireland, and a discussion of admissibility of forensic voice comparison in light of England & Wales Criminal Practice Directions 19A ([2015] EWCA Crim 1567 V 19A), see Morrison [24].

interest on the questioned- and known-speaker recordings if they were both produced by the same speaker versus the likelihood of obtaining the speech of interest on the questioned- and known-speaker recordings if they were produced by two different speakers from the relevant population.[4] Black-box testing is not concerned with how a method performs the task, only with how well the method performs the task. It therefore does not matter whether a method for performing forensic voice comparison is a human-supervised-automatic method or an auditory-acoustic-phonetic method; different methods for performing the same task can, and should, be validated using the same validation protocol. The statement in Kirchhübel et al. [22] that “it would not be possible to simply adopt the recommendations made in [the *Consensus*] for the [auditory-acoustic-phonetic] approach” is therefore incorrect unless there is a practical impediment to conducting validation of auditory-acoustic-phonetic methods according to the recommendations of the *Consensus*. The amount of time and amount of human effort that it takes to compare each pair of recordings in the validation set will be much much greater for an auditory-acoustic-phonetic method than for a human-supervised-automatic method. Validating an auditory-acoustic-phonetic method will therefore be practically much more difficult than validating a human-supervised-automatic method. This does not, however, excuse auditory-acoustic-phonetic methods from the requirement applicable to all methods that they be meaningfully validated.

As stated in the *Consensus*, a necessary condition for validation to be meaningful is the

[4] Kirchhübel et al. [22] states that the practitioner and the reviewer “expressed their conclusions with reference to the scale that is recommended by the UK [*sic*] Association of Forensic Science Providers [25] and ENFSI [26] (however, their conclusions were not derived from a numerical likelihood ratio).” Both those scales, however, are intended to provide verbal expressions corresponding to numerical ranges of likelihood ratios, and the ENFSI Guideline states that even if the numerator and denominator of a likelihood ratio are “informed by subjective probabilities using expert knowledge. These probability assignments shall still be expressed by a number between 0 and 1 rather than by an undefined qualifier (such as frequent, rare, etc.).” There is no evidence in Kirchhübel et al. [22] that the practitioner or reviewer actually followed the logic of the likelihood-ratio framework.

following:[5]

> 2.5.2. Validation data [pairs of same-speaker recordings and pairs of different-speaker recordings] should be sufficiently representative of the relevant population for the case, and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case, that the results of validating the system using those data will be informative as to the expected performance of the system when it is applied in the case.

> 2.5.3. One of the criteria for the validation data to be sufficient is that the number of speakers included be sufficient. Because of sampling variability, small validation sets can give results that are not representative of the case conditions.

The number of speakers is a constraint on how many unique same-speaker and unique different-speaker pairs can be constructed. Factorial combinations allow for many more unique different-speaker pairs to be constructed, but the number of unique same-speaker pairs that can be constructed will be limited to the number of speakers from whom pairs of recordings are available. The *Consensus* does not recommend a specific value for what would constitute a sufficient number of speakers, but instead states that:

> 2.6.1. The decision as to whether the calibration data and the validation data are sufficiently representative of the relevant population for the case and sufficiently reflective of the conditions of the questioned-speaker and known-speaker recordings in the case will be the result of a subjective judgment made by the forensic practitioner.

> 2.6.7. The forensic practitioner should communicate to the court a clear description of the calibration data and the validation data used.

> 2.6.8. A description of the calibration and validation data is a prerequisite for a

[5] The following quotations use the original paragraph numbering from Morrison et al. [17].

second forensic practitioner to be able to conduct an independent review so as to be able to opine on whether the data are sufficient.

2.6.9. A description of the calibration and validation data is also a prerequisite for the court to be able to decide to either accept or reject the first forensic practitioner's decision about the sufficiency of the data.

Larger numbers of speakers would be better, but the number of speakers included in the validation set will be constrained by the cost and time required to obtain data that are sufficiently representative of the relevant population and sufficiently reflective of the questioned-speaker and known-speaker recordings' conditions. In may be that a practically achievable validation set consists of pairs of recordings from only upper tens of speakers to lower hundreds of speakers. Whether a validation set of this size would be sufficiently representative of the relevant population and sufficiently reflective of the recording conditions of a case is a matter of judgement, and, ultimately, of acceptance by the court. It is very clear, however, that a validation set consisting of a single test pair, as proposed in Kirchhübel et al. [22], would not be sufficiently representative of the relevant population nor sufficiently reflective of the recording conditions for the validation results to be informative as to the expected performance of the system when it is applied in the case. A validation consisting of a single test pair is not meaningful.[6]

Practitioners of forensic voice comparison should think like forensic scientists not like phoneticians. A practitioner thinking like a forensic scientist will use whatever method they believe will best perform the task of forensic voice comparison. Their choice of method should be informed by prior validation studies. The decision as to whether the performance of the method is sufficiently good in the context of the particular case

[6] In Kirchhübel et al. [22], the single test pair was selected from a set of nine potential pairs. The validation set presented to the practitioner in Kirchhübel et al. [22], however, consisted of that single test pair. If the validation set had consisted of nine test pairs, we would argue that that would also have been too small.

must be based on a validation of the method using data that are representative of the relevant population for the case and reflective of the conditions of the questioned-speaker and known-speaker recordings in the case. The validation may have been conducted ahead of time (anticipatory validation) and a judgement made that the conditions of the case are sufficiently similar to the conditions under which the existing validation was conducted; or, if such a validation does not already exist, a new validation should be conducted using data that are judged to be sufficiently similar to the conditions of the case (case-by-case validation). In contrast, a practitioner thinking like a phonetician will persist in only using auditory-phonetic or acoustic-phonetic methods, even when other methods (such as human-supervised-automatic methods) have been demonstrated to result in superior performance, and even when the performance of the auditory-phonetic or acoustic-phonetic methods have not been meaningfully demonstrated at all. If the cost and time necessary to conduct meaningful validations of auditory-acoustic-phonetic methods are such that it is not practical to conduct meaningful validations, then forensic voice comparison should not be performed using auditory-acoustic-phonetic methods.

References

[1] Morrison G.S. (2022). Advancing a paradigm shift in evaluation of forensic evidence: The rise of forensic data science. *Forensic Science International: Synergy*, 5, 100270. https://doi.org/10.1016/j.fsisyn.2022.100270

[2] Mnookin J.L., Cole S.A., Dror I.E., Fisher B.A.J., Houck M.M., Inman K., Kaye D.H., Koehler J.J., Langenburg G., Risinger D.M., Rudin N., Siegel J., Stoney D.A. (2011). The need for a research culture in the forensic sciences. *UCLA Law Review*, 58, 725–777. https://www.uclalawreview.org/the-need-for-a-research-culture-in-the-forensic-sciences-2/

[3] Risinger D.M. (2013). Reservations about likelihood ratios (and some other aspects of forensic ‘Bayesianism’). *Law, Probability and Risk*, 12, 63–73, http://dx.doi.org/10.1093/lpr/mgs011

[4] President’s Council of Advisors on Science and Technology (2016). *Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Methods*. https://obamawhitehouse.archives.gov/administration/eop/ostp/pcast/docsreports/

[5] Morrison G.S., Thompson W.C. (2017). Assessing the admissibility of a new generation of forensic voice comparison testimony. *Columbia Science and Technology Law Review*, 18, 326–434. https://doi.org/10.7916/stlr.v18i2.4022

[6] Cole S.A. (2006). Is fingerprint identification valid? Rhetorics of reliability in fingerprint proponents’ discourse. *Law & Policy*, 28, 109–135. https://doi.org/10.1111/j.1467-9930.2005.00219.x

[7] Morrison G.S. (2014). Distinguishing between forensic science and forensic pseudoscience: Testing of validity and reliability, and approaches to forensic voice comparison. *Science & Justice*, 54, 245–256. http://dx.doi.org/10.1016/j.scijus.2013.07.004

[8] Koehler J.J. (2017). Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences. *Arizona State Law Journal*, 49(4), 1369–1416. https://arizonastatelawjournal.org/2018/02/07/forensics-or-fauxrensics-ascertaining-accuracy-in-the-forensic-sciences/

[9] Morrison G.S., Ballantyne K., Geoghegan P.H. (2018). A response to Marquis et al (2017) What is the error margin of your signature analysis? *Forensic Science International*, 287, e11–e12. https://doi.org/10.1016/j.forsciint.2018.03.009

[10] Morrison G.S., Neumann C., Geoghegan P.H. (2020). Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 2, 206–209. https://doi.org/10.1016/j.fsisyn.2020.06.005

[11] Morrison G.S., Neumann C., Geoghegan P.H., Edmond G., Grant T., Ostrum R.B., Roberts P., Saks M., Syndercombe Court D., Thompson W.C., Zabell S. (2021). Reply to Response to Vacuous standards – subversion of the OSAC standards-development process. *Forensic Science International: Synergy*, 3, 100149. https://doi.org/10.1016/j.fsisyn.2021.100149

[12] Meuwly D. (2001). *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. Doctoral dissertation, University of Lausanne. https://www.unil.ch/files/live/sites/esc/files/shared/These.Meuwly.pdf

[13] Brümmer N., du Preez J. (2006). Application independent evaluation of speaker detection. *Computer Speech and Language*, 20, 230–275. https://doi.org/10.1016/j.csl.2005.08.001

[14] Morrison G.S. (2011). Measuring the validity and reliability of forensic likelihood-ratio systems. *Science & Justice*, 51, 91–98. http://dx.doi.org/10.1016/j.scijus.2011.03.002

[15] Meuwly D., Ramos D., Haraksim R. (2017). A guideline for the validation of likelihood ratio methods used for forensic evidence evaluation. *Forensic Science International*, 276, 142–153. http://dx.doi.org/10.1016/j.forsciint.2016.03.048

[16] Ramos D., Meuwly D., Haraksim R., Berger C.E.H. (2020). Validation of forensic automatic likelihood ratio methods. In Banks D., Kafadar K., Kaye D.H., Tackett M. (Eds.), *Handbook of Forensic Statistics* (pp. 143–163). Boca

Raton, FL: CRC. https://doi.org/10.1201/9780367527709

[17] Morrison G.S., Enzinger E., Hughes V., Jessen M., Meuwly D., Neumann C., Planting S., Thompson W.C., van der Vloed D., Ypma R.J.F., Zhang C., Anonymous A., Anonymous B. (2021). Consensus on validation of forensic voice comparison. *Science & Justice*, 61, 229–309. https://doi.org/10.1016/j.scijus.2021.02.002

[18] Morrison G.S., Enzinger E. (2016). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) - Introduction. *Speech Communication*, 85, 119–126. https://doi.org/10.1016/j.specom.2016.07.006

[19] Morrison G.S., Enzinger E. (2019). Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) - Conclusion. *Speech Communication*, 112, 37–39. https://doi.org/10.1016/j.specom.2019.06.007

[20] Weber P., Enzinger E., Labrador B., Lozano-Díez A., Ramos D., González-Rodríguez J., Morrison G.S. (2022). Validation of the alpha version of the E^3 Forensic Speech Science System (E^{3}FS^3) core software tools. *Forensic Science International: Synergy*, 4, 100223. https://doi.org/10.1016/j.fsisyn.2022.100223

[21] Basu N., Bali A.S., Weber P., Rosas-Aguilar C., Edmond G., Martire K.A., Morrison G.S. (2022). Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International*, 341, 111499. https://doi.org/10.1016/j.forsciint.2022.111499

[22] Kirchhübel C., Brown G., Foulkes P. (2023). What does method validation look like for forensic voice comparison by a human expert? *Science & Justice*,

63, 251–257. https://doi.org/10.1016/j.scijus.2023.01.004

[23] Morrison G.S., Zhang C. (2023). Forensic voice comparison: Overview. In Houck M., Wilson L., Eldridge H., Lewis S., Lothridge K., Reedy P. (Eds.), *Encyclopedia of Forensic Sciences* (3rd Ed.), vol. 2, pp. 737–750. Elsevier. https://doi.org/10.1016/B978-0-12-823677-2.00130-6

[24] Morrison G.S. (2018). Admissibility of forensic voice comparison testimony in England and Wales. *Criminal Law Review*, 2018(1), 20–33. [Preprint available at http://geoff-morrison.net/#Admissibility_EW_2018]

[25] Association of Forensic Science Providers (2009). Standards for the formulation of evaluative forensic science expert opinion. *Science & Justice*, 49, 161–164. http://dx.doi.org/10.1016/j.scijus.2009.07.004

[26] Willis S.M., McKenna L., McDermott S., O’Donell G., Barrett A., Rasmusson A., Nordgaard A., Berger C.E.H., Sjerps M.J., Lucena-Molina J.J., Zadora G., Aitken C.G.G., Lunt L., Champod C., Biedermann A., Hicks T.N., Taroni F. (2015). *ENFSI Guideline for Evaluative Reporting in Forensic Science*. http://enfsi.eu/wp-content/uploads/2016/09/m1_guideline.pdf