**SUPERIOR COURT OF NEW JERSEY**
**LAW DIVISION – MONMOUTH COUNTY**
**INDICTMENT NO. 19-02-0283**

| | | |
|---|---|---|
| **STATE OF NEW JERSEY,** | **:** | **CRIMINAL ACTION** |
| **Plaintiff** | **:** | |
| **v.** | **:** | |
| **PAUL CANEIRO,** | **:** | **Hon. Marc C. Lemieux, A.J.S.C.** |
| **Defendant** | **:** | |

---

# PROPOSED FINDINGS OF FACT AND CONCLUSIONS OF LAW

---

**Jennifer Sellitti**
**Public Defender**
**Office of the Public Defender**
**Hughes Justice Complex**
**P.O. Box 850, 25 Market Street**
**Trenton, NJ 08625**

**TAMAR Y. LERER**
**Tamar.Lerer@opd.nj.gov**
**Deputy Public Defender**

**DEFENDANT IS CONFINED**

# TABLE OF CONTENTS

**PAGE NOS.**

PRELIMINARY STATEMENT ..... 1

PROPOSED FINDINGS OF FACT ..... 3

A. Two Versions of STRmix Were Used by Two Different Laboratories to Analyze DNA Evidence in this Case ..... 3

B. Witnesses Presented ..... 4

1. State's Witnesses4

a. Monica Ghannam ..... 5

b. Kristen Naughton ..... 5

c. Danielle Reed ..... 6

d. John Buckleton ..... 6

e. Jennifer Thayer ..... 8

f. Christine Schlenker ..... 9

g. Michael Coble ..... 9

2. Defense Witnesses ..... 10

a. Karl Reich ..... 10

b. Mats Heimdahl ..... 11

c. Nathan Adams ..... 13

d. Paul Martin ..... 14

e. Keith Inman ..... 15

C. A basic overview of STRmix ..... 16

1. STRmix is both a mode of DNA interpretation and piece of software. The science of both forensic DNA and software engineering are relevant to assessing its reliability. ..... 16

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

2. STRmix attempts to separate out the DNA profiles of each person who contributed to a mixture through statistical theory, computer algorithms, and probability distributions. ..... 17

3. STRmix produces a likelihood ratio to express the probability of observing the DNA evidence under two different hypotheses ..... 18

a. Different systems, assumptions, and users will lead to different likelihood ratios being produced. ..... 19

b. There is no one true likelihood ratio for any given sample. ..... 21

D. Software engineering has rigorous standards to ensure the reliability of important software such as STRmix ..... 22

1. Software faults are common and difficult to detect. ..... 23

2. Software is determined to be reliable only after it has been verified and validated. ..... 24

3. STRmix is a safety-critical system. ..... 26

4. In order to be considered reliable, safety-critical software must be verified and validated independently. ..... 26

a. Detailed requirements and specifications are a prerequisite for any attempt to verify and validate a software system. ..... 28

b. Verification means testing a software system against its requirements to determine that it has been built as intended. ..... 30

c. Validation means checking that the software fulfills the needs of its stakeholders. ..... 32

d. Validation and verification for safety-critical systems must be undertaken by independent testers. ..... 33

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

5. Verification and validation is a commonplace process that is expected to occur for all software. ..... 34

6. The publication of peer-reviewed articles is not an accepted means of determining the reliability of software. ..... 35

E. STRmix v2.5.11 and v.2.8.0 have not been independently verified and validated. ..... 35

1. STRmix v2.5.11 and STRmix v.2.8.0 have not been verified. ..... 37

2. STRmix v2.5.11 and STRmix v.2.8.0 have not been validated. ..... 39

3. Testing STRmix with ground-truth samples is not a substitute for verification or validation. ..... 42

4. Any verification or validation of STRmix that has occurred has not been undertaken by sufficiently independent testers. ..... 46

5. Coming close to complying with verification and validation standards is not sufficient to assure reliability. ..... 48

6. Source code review in an adversarial setting is not a substitute for IV&V and does not guarantee the software is free of flaws. ..... 49

a. The source code review in this case was limited. ..... 49

b. A source code review by a criminal defendant would never be able to replicate independent validation and verification. ..... 52

7. STRmix's reliability has been insufficiently demonstrated under software engineering standards. ..... 53

F. The ability of any probabilistic genotyping system to reliably analyze complex DNA samples must be empirically demonstrated, not assumed. ..... 53

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

1. All scientific methods have limits. The reliable use of any method requires finding those limits and adhering to them. ..... 53

2. Basics of Forensic DNA Analysis ..... 54

3. Samples that contain more than one person's DNA are more challenging to interpret. ..... 57

4. Some DNA mixtures are very hard to reliably analyze. ..... 58

5. Related contributors create an intractable difficulty for DNA analysis. ..... 62

6. Probabilistic genotyping systems were designed in order to attempt to analyze complex mixtures. ..... 63

7. Probabilistic genotyping systems, including STRmix, will produce false positive and false negative errors. ..... 64

a. In probabilistic genotyping systems, false positives are inclusionary likelihood ratios for non-contributors and false negatives are exclusionary LRs for contributors. ..... 64

b. Probabilistic genotyping systems will frequently give inclusionary likelihood ratios for non-contributors and exclusionary likelihood ratios for contributors. ..... 65

8. Because the error rate of probabilistic genotyping systems will vary across sample types, reliability must be established across those sample types. ..... 68

a. Both developmental and internal validation are necessary to demonstrate the reliability of the use of probabilistic genotyping systems in casework. ..... 68

b. Developmental validation must establish the outermost bounds of the reliable use of probabilistic genotyping systems. ..... 69

c. Internal validation studies are necessary to establish the reliability of a method and are a mandatory prerequisite to their implementation in any laboratory. ..... 70

d. Internal validation studies establish the limits of a laboratory's ability to reliably use a technique, including probabilistic genotyping systems. ..... 71

i. Standards, guidelines, and best practices require the use of internal validations to set the limits of what kinds of samples a laboratory will analyze. ..... 71

ii. Standard operating procedures require the setting of a limit on what kinds of samples a laboratory will analyze. ..... 73

iii. Establishing a limit means establishing a boundary of what kinds of samples a laboratory will and will not analyze. ..... 74

e. It is possible for internal validation studies to study the impact of relatedness on a probabilistic genotyping system's reliability. ..... 76

i. Internal validations reveal that non-donors related to the real donors to a sample are at a significant risk of being falsely included in that sample at a high likelihood ratio. ..... 77

ii. The use of the likelihood ratios produced by STRmix is inappropriate for samples with related contributors. ..... 78

9. Even with the best-calibrated PGS being used appropriately, many non-contributors would yield an inclusionary likelihood ratio if their profile was run against a complex sample. ..... 82

10. Probabilistic genotyping systems, including STRmix, rely on human analysts to make discretionary decisions in operating the software. ..... 83

a. Human analysts make decisions about STRmix inputs ..... 83

b. Probabilistic genotyping systems, including STRmix, rely on human analysts to use their understanding of diagnostics to spot errors. ..... 84

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

c. STRmix diagnostics are new to DNA analysts making these discretionary decisions..... 84

d. Human analysts are subject to cognitive bias and another cognitive limitations..... 85

11. The outputs of probabilistic genotyping systems are likelihood ratios, numbers that should be reported quantitatively..... 87

12. All forms of DNA interpretation, including probabilistic genotyping, are constrained by fundamental principles of genetics, human judgment, and must be appropriately implemented..... 88

G. There is limited evidence to support the foundational reliability of STRmix v2.5.11 and v2.8.0. ..... 89

1. There has been limited testing of STRmix across the range of samples STRmix is used on and almost none of that testing is independent..... 89

a. The published developmental validation of STRmix is not independent and is limited in its scope. ..... 90

b. Internal validation studies cannot compensate for insufficient developmental testing. ..... 92

2. There is almost no information about STRmix's error rate across the range of samples STRmix is used on. ..... 96

a. No false positive or false negative rates have been provided. ..... 96

b. There is no evidence of what the error rates would be across mixtures with various features. ..... 97

c. Even if error rates have been provided, there is no evidence that the tests have been run on large enough samples such that those rates are known to be representative. ..... 100

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

d. There is evidence of significant rates of error. ..... 100

e. Trends do not sufficiently describe the error rate of STRmix across sample types. ..... 102

3. No standards substantively govern STRmix's performance. ..... 104

a. FBI Quality Assurance Standards ..... 104

b. Scientific Working Group on DNA Analysis Methods Guidelines ..... 105

c. ANSI/ASB Standard 18 for the Validation of Probabilistic Genotyping Systems. ..... 105

d. The International Society for Forensic Genetics Recommendations on the Validation of Software Programs Performing Biostatistical Calculations for Forensic Genetics Applications..... 105

e. Forensic Science Regulator of the United Kingdom, Software Validation for DNA Mixture Interpretation ..... 106

f. ISO 17025 ..... 106

g. Audits of accreditation standards do not operate to ensure reliability. ..... 106

h. Analyst intuition is not a replacement for effective standards. ..... 108

4. There are almost no independent, peer-reviewed publications that assess STRmix's reliability. ..... 108

5. The unlimited use of STRmix is not accepted in the field of forensic DNA analysis. ..... 111

6. No limits have been identified for the reliable use of STRmix by the developers. ..... 113

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

7. Humans are a necessary part of PGS which means proficiency and cognitive bias impact the ability to properly use it. 113

H. Even assuming the foundational reliability of STRmix, the use of STRmix in this case has not been demonstrated to be reliable.....113

1. Bode's internal validation. .....114

a. In its validation study, Bode did not attempt to test any sample of similar complexity to the samples it analyzed in this case.....114

i. In its validation study, Bode did not test any sample in which the minor contributor contributed fewer than 25 picograms of DNA. .....115

ii. In its validation study, Bode did not test any sample in which the minor contributor contributed less than 5% of the total mixture.....115

iii. Bode's validation study reveals significant discrepancies between the mixture proportions deliberately created by Bode and STRmix's estimation of those proportions.....115

iv. In its validation study, Bode tested only two samples in which the minor contributor contributed 25 picograms of DNA and 5% of the total mixture. .....117

v. In its validation study, Bode did not test the impacts of relatedness. .....117

vi. In its validation study, Bode did not validate any of the likelihood ratios used to hypothesis the real contributor is a relative of the person of interest. .....117

b. Samples in this case tested by Bode were outside of the limits of the validation study or close to those limits. .....117

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

c. It was inappropriate for Bode to test and report results for samples outside of the boundaries of its validation study. ..... 119

2. Bode's Standard Operating Procedures give very little objective guidance on what kinds of samples analysts should analyze. The little guidance given was not followed in this case. ..... 122

a. Bode's Standard Operating Procedures have very few limits on what samples to run and results to report. ..... 122

b. Bode's Standard Operating Procedures note increased risks of false positives on the kinds of samples analyzed in this case. ..... 122

c. Despite the risk of unreliable results when first-degree relatives may be in a mixture, recognized in its Standard Operating Procedures, Bode ran samples with potential first-degree relatives without any explanation or risk mitigation. ..... 122

d. Despite the risk of unreliable results when profiles with very limited data are analyzed, recognized in its Standard Operating Procedures, Bode analyzed samples where there was very limited profile data about the minor contributor. ..... 124

e. Rather than SOPs, Bode relies significantly on the DNA analyst's intuition. ..... 124

f. Despite analyst exposure to potentially biasing information, the Bode SOPs have no bias mitigation protocols. ..... 125

g. The Bode SOPs do not contain an uninformative range. ..... 125

h. Bode inappropriately conducted visual exclusions in this case instead of running all relevant people through STRmix. ..... 125

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

3. In its validation study, New Jersey State Police did not test the impacts of relatedness. ..... 126

4. New Jersey State Police Standard Operating Procedures give very little objective guidance on what kinds of samples analysts should analyze. ..... 127

a. New Jersey State Police Standard Operating Procedures did not establish limits on suitability for STRmix analysis. ..... 127

b. New Jersey State Police Standard Operating Procedures contain insufficient guidance on analyzing and reporting mixtures comprised of related individuals. ..... 128

c. Rather than Standard Operating Procedures, New Jersey State Police relies significantly on the DNA analyst's intuition. ..... 128

d. Despite analyst exposure to potentially biasing information, the NJSP Standard Operating Procedures have no bias mitigation protocols. ..... 128

5. Both Bode and New Jersey State Police conducted their sensitivity and specificity studies using mixture proportions and amount of DNA in picograms to assess performance. ..... 129

6. Neither internal validation summary provides sufficient information to assess the reliability of STRmix as used by these laboratories across different sample types. ..... 131

7. The laboratories did not stay without the boundaries of the reliable use of STRmix, whatever those undefined boundaries might be. ..... 134

PROPOSED CONCLUSIONS OF LAW ..... 135

THE STATE'S PROFFERED STRMIX EVIDENCE FAILS TO MEET NEW JERSEY'S ADMISSIBILITY STANDARDS AS SET FORTH IN STATE V. OLENOWKSI AND N.J.R.E. 702. ..... 135

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

A. The State has failed to prove that STRmix v2.5.11 and 2.8.0 are foundationally reliable. ..... 136

1. STRmix v.2.5.11 and 2.8.0 have been inadequately tested. ..... 138

2. There is insufficient evidence of STRmix's error rate to demonstrate its reliability. ..... 140

a. The false positive and false negative rates are unknown. ..... 140

b. The risk of error is not limited to false inclusions and exclusions, but incorrect likelihood ratios. ..... 142

c. Testing STRmix by running DNA samples does not capture the risk of software error. ..... 144

d. Without a known rate of error, STRmix is not admissible. ..... 145

3. There are insufficient standards governing the use of STRmix. ..... 146

4. There are almost no independent, peer-reviewed publications that assess STRmix's reliability. ..... 147

5. The State has not demonstrated general acceptance in the relevant fields. ..... 149

6. Very few courts have actually looked at the reliability of STRmix in a thorough and nuanced way. ..... 151

a. Court opinions do not provide guidance when they do not substantively engage with the arguments presented in this case. ..... 151

b. Courts often reflexively accept forensic science as reliable without sufficient examination of the method. New Jersey courts do not. ..... 152

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

c. No judicial opinions persuasively demonstrate the reliability of STRmix. ..... 154

d. United States v. Lewis ..... 156

e. United States v. Gissantaner ..... 158

7. The State has failed to demonstrate the foundational reliability of STRmix v.2.5.11 and v2.8.0. ..... 159

B. The Likelihood Ratios produced by STRmix are not appropriate to analyze mixtures that contain related contributors and are inadmissible. ..... 160

C. The State has failed to demonstrate that STRmix was used reliably in this case. ..... 160

1. Internal validation studies establish the limits of what is admissible in court. ..... 161

2. The State has not established the appropriate limits in this case due to validation summaries that are conclusory and insufficiently detailed. Therefore, no evidence from either lab is admissible. ..... 164

3. In the alternative, testimony about STRmix results is admissible only if the sample tested falls within the range of samples tested in a laboratory' internal validation study. Because in this case the samples tested are more complex than those in the validation summary, the results of those tests are inadmissible. ..... 165

a. All of the results are inadmissible because they all involve analysis of mixtures that involve related people. ..... 166

b. Samples E02b1, E03b1, E04a1, E06a1, and E07a1 are inadmissible because they are samples at or below the limits of what was tested by Bode in its internal validation summary and are comprised of related people. ..... 167

## TABLE OF CONTENTS (CONT'D)

**PAGE NOS.**

4. The State has not established that these analysts can or did reliably use STRmix. ..... 168

5. The results are reported in a manner incompatible with the reliable use of STRmix. ..... 168

a. Bode should be using familial or unified likelihood ratios to consider whether the real contributor could be related to the person of interest. ..... 169

b. The verbal scale is misleading. ..... 169

c. New Jersey State Police's use of STRmix to render a source attribution is inappropriate. ..... 169

6. The STRmix results are inadmissible in this case..... 170

CONCLUSION..... 171

## TRANSCRIPT LEGEND

1T — November 12, 2024 Volume 1 (Ghannam)

2T — November 12 Volume 2 (Ghannam, Naughton)

3T — November 13, 2024 Volume 1 (Naughton, Reed)

4T — November 13, 2024 Volume 2 (Reed)

5T — November 14, 2024 Volume 1 (Reed)

6T — November 14, 2024 Volume 2 (Buckleton)

7T — November 15, 2024 (Buckleton)

8T — November 18, 2024 Volume 1 (Thayer)

9T — November 18, 2024 Volume 2 (Thayer/Schlenker)

10T — November 19, 2024 Volume 1 (Coble)

11T — November 19, 2024 Volume 2 (Coble)

12T — December 2, 2024 (Reich)

13T — December 3, 2024 (Heimdahl)

14T — December 4, 2024 (Adams)

15T — December 6, 2024 (Martin)

16T — December 9, 2024 (Inman)

17T — December 13, 2024 (Closing Arguments)

A list of defense exhibits has filed simultaneously with this brief.

# TABLE OF AUTHORITIES

**PAGE NOS.**

**Cases**

Commonwealth v. Davis, 168 N.E.3d 294 (Mass. 2021) ..... 158

Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579 (1993) ..... 134, 147, 149, 158

General Electric v. Joiner, 522 U.S. 136 (1997) ..... 137

In re Accutane Litig., 234 N.J. 340 (2018) ..... 134, 168

People v. Blash, 2018 WL 4062322 (V.I. Super. 2018) ..... 153

People v. Bullard-Daniel, 54 Misc. 3d 177 N.Y.S.3d 714 (N.Y. Co. Ct. 2016) ..... 153

People v. Kelly, 549 P.2d 1240 (Cal. 1976) ..... 146

People v. Seepersad, 101 N.Y.S.3d 701 (N.Y. Sup. Ct. 2018) ..... 153

People v. Smith, No. 340845, 2018 Mich. App. LEXIS 3274 (Ct. App. Oct. 9, 2018) ..... 153

People v. Williams, 147 N.E.3d 1131 (N.Y. 2020) ..... 150, 151

Romano v. Kimmelman, 96 N.J. 66 (1984) ..... 146

State v. Cassidy, 235 N.J. 482 (2018) ..... 134, 135

State v. Chun, 194 N.J. 54 (2008) ..... 13

State v. Doriguzzi, 334 N.J. Super. 530 (2000) ..... 150

State v. Henderson, 208 N.J. 208 (2011) ..... 145

State v. J.L.G., 234 N.J. 265 (2018) ..... 150

State v. J.R., 227 N.J. 393 (2017) ..... 133

State v. Kelly, 97 N.J. 178 (1984) ..... 133

State v. Lasworth, 42 P.3d 844 (N.M. Ct. App. 2001) ..... 158

State v. Olenowski, 253 N.J. 133 (2023) ..... 133, 134, 144, 149

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Cases (Cont'd)**

State v. Olenowski, 255 N.J. 529 (2023) ..... 4, 87, 133, 138, 144, 158, 159

State v. Pickett, 466 N.J. Super. 270 (App. Div. 2021) ..... passim

State v. Rochat, 470 N.J. Super. 392 (App. Div. 2022) ..... 150, 152

State v. Rochat, 470 N.J. Super. 392 (App. Div. 2022), cert. denied, 252 N.J. 79 (2022) ..... 152

United States v. Christensen, No. 17-CR-20037-JES-JEH, 2019 U.S. Dist. LEXIS 24623, 2019 WL 651500 (C.D. Ill. Feb. 15, 2019) ..... 153

United States v. Francisco Ortiz, Case No.: 21-CR-2503-GPC (S.D. Cal. June 10, 2024) ..... 161

United States v. Jaber, 2023 WL 8254358 (M.D Fla. 2023) ..... 152

United States v. Kevin Johnson, 15-CR-565 (VEC) (S.D.N.Y. July 6, 2016) ..... 150

United States v. Lewis, 442 F. Supp. 3d 1122 (D. Minn. 2020) ..... 154, 155

United States v. Russell, CR-14-2563 MCA, 2018 WL 7286831 (D.N.M. Jan. 10, 2018) ..... 153

United States v. Washington, 2020 WL 3265142 (D. Neb. 2020) ..... 152

United States v. Williams, 2023 WL 5155252 (D. Minn. 2023) ..... 152

Walker v. New York, No. 14-cv-680(NRM)(PK) (E.D.N.Y. September 16, 2024) ..... 161

Windmere Inc. v. Int'l. Ins. Co., 105 N.J. 373 (1987) ..... 145

**Rules of Evidence**

Fed. R. Evid. 702 ..... 133

N.J.R.E. 403 ..... 141, 166

N.J.R.E. 702 ..... 132, 133, 145, 158, 167

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Other Authorities**

ANSI/ASB Standard 18, Standard for Validation of Probabilistic Genotyping Systems (2020) ..... 70

Bess Stiffelman, No Longer the Gold Standard: Probabilistic Genotyping Is Changing the Nature of DNA Evidence in Criminal Trials, 24 Berkeley J. Crim. L. 110 (2019) ..... 19

Buckleton et al, The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity, 64 J. Forensic Sci. 393 (2019) ..... 65

David Murray, Queensland Authorities Confirm "Miscode" Affects DNA evidence in Criminal Cases (March 20, 2015) ..... 12

Department of Justice, Uniform Language For Testimony And Reports For Forensic Autosomal DNA Examinations Using Probabilistic Genotyping Systems (Sept. 2019) ..... 167

Developmental Validation Of STRmix, Expert Software For The Interpretation Of Forensic DNA Profiles, 23 FSI: Genetics 226 (2016) ..... 88

DNA Analysis Methods, Guidelines for the Validation of Probabilistic Genotyping Systems 4 (June 2015) ..... 71, 103

Duke and Myers, Systematic Evaluation Of STRmix Performance On Degraded DNA Profile Data, 44 FSI: Genetics (2020) ..... 108

FBI, Quality Assurance Standards for Forensic DNA Testing Laboratories (2011) ..... 71

Federal Bureau Of Investigation, Quality Assurance Standards for Forensic DNA Testing Laboratories 3 (2011) ..... 69

Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Method 49 (2016) ..... 85

Greenspoon et al., A Tale of Two, supra. Of the four remaining, one is not a peer-reviewed publication, but a master's thesis. Diana Orozco, TrueAllele and STRmix: A Comparison of Two Probabilistic Genotyping Software Programs in Forensic DNA Profile Analysis, University of California, Davis (2023) ..... 108

Hinda Haned, et.al., Validation Of Probabilistic Genotyping Software For Use In Forensic DNA Casework: Definitions And Illustrations, 56 Sci. & Justice 104 (2016) ..... 21

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Other Authorities (Cont'd)**

Ian Sommerville, Software Engineering 21 (2016, 10th Ed.) ..... 22

Jo-Anne Bright and Michael Coble, Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios (2020) ..... 86

Jo-Anne Bright et al., Internal Validation of STRmix – A multi laboratory response to PCAST, 34 FSI: Genetics 11 (2018) ..... 93

John Buckleton et al, A diagnosis of the primary difference between EuroForMix and STRmix™, 69 J. Forensic Sci. 40 (2024) ..... 86

John Buckleton et al., A Series Of Recommended Tests When Validating Probabilistic DNA Profile Interpretation Software. 14 FSI: Genetics 125 (2015) ..... 73

John Buckleton et al., Response to: Commentary on: Bright et al. (2018) Internal validation of STRmix™ – A multi laboratory response to PCAST, 34 Forensic Science International: Genetics, 34: 11–24, 44 FSI: Genetics 1 (2020) ..... 65, 98

John Butler, Fundamentals of DNA Interpretation 230 (2010) ..... 65

John M. Butler et al., National Institute of Standards and Technology, DNA Mixture Interpretation: A NIST Scientific Foundation Review 47 (Dec. 2024) ..... 17

John M. Butler, Advanced Topics in Forensic DNA Typing: Interpretation 136 (2014) ..... 57

John M. Butler, Advanced Topics in Forensic DNA Typing: Methodology 324-26 (2011) ..... 59

John M. Butler, Fundamentals of DNA Typing 19 (2010) ..... 55

John S. Buckleton et al., Are Low Lrs Reliable?, 49 Forensic Sc. Int'l: Genetics ..... 66, 96

Jonathan J. Koehler, Intuitive Error Rate Estimates for the Forensic Sciences, 57 Jurimetrics 153 (2017) ..... 140

Laura Russell et al., A Guide To Results And Diagnostics Within A STRmix™ Report, WIREs Forensic Sci. 2 (2019) ..... 82

Lauren M Brinkac et al., DNAmix 2021: Laboratory Policies, Procedures, And Casework Scenarios Summary And Dataset, 48 Data in Brief 1 (2023) ..... 84

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Other Authorities (Cont'd)**

Lawrence Joseph et al., Bayesian And Mixed Bayesian/Likelihood Criteria For Sample Size Determination, 16 Statistics in Medicine 769 (1997) ..... 99

Lisa Smith et al, Understanding Juror Perceptions of Forensic Evidence: Investigating the Impact of Case Context on Perceptions of Forensic Evidence Strength, 56 J. Forensic Sci. (2011)..... 140

M. McCarthy-et al., Low Lrs Obtained From DNA Mixtures: On Calibration And Discrimination Performance Of Probabilistic Genotyping Software, 73 FSI: Genetics (2024) ..... 108

Medical device software – Software life cycle processes..... 25

Moretti et al., Internal Validation Of STRmix For The Interpretation Of Single Source And Mixed DNA Profiles, 126 FSI: Genetics 126 (2017)..... 99

Nancy Leveson, Medical Devices: The Therac-25 in Safeware: System Safety and Computers at 6-9 (1995) ..... 24

NASA Software Engineering Requirements (eff. March 8, 2022) ..... 25

Nat'l Inst. Standards & Tech., DNA Mixture Interpretation: A NIST Scientific Foundation Review 11 (Dec. 2024) ..... 54

National Institute of Standards and Technology, Forensic DNA Interpretation and Human Factors: Improving Practice Through a Systems Approach 72 (May 2024)..... 82

National Institute of Standards and Technology, Forensic DNA Interpretation and Human factors: Improving Practice Through a Systems Approach 81 (May 2024)..... 19

National Institute of Standards and Technology, Supplemental Document to DNA Mixture Interpretation: A NIST Scientific Foundation Review at 42 (Dec. 2024)..... 17, 106

National Institute of Standards and Technology, Views of the Commission: Validation of Forensic Science Methodology (2016)..... 69

National Research Council, Strengthening Forensic Science in the United States: A Path Forward 112 (2009) ..... 145

NIST Internal Report ..... 32

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Other Authorities (Cont'd)**

NIST Internal Report 8351-Draft DNA Mixture Interpretation: A NIST Foundation Review 3 (2021)..... 27, 40

Noel et al., STRmix™ Put To The Test: 300 000 Non-Contributor Profiles Compared To Four-Contributor DNA Mixtures And The Impact Of Replicates, 41 FSI: Genetics (2019).. 108

People v. Collin T., Notice Of Motion To Preclude Expert Testimony Or For A Limited Remote-Video Hearing under Frye/Wesley at 33 (N.Y. Sup. Ct. June 18, 2020)..... 78

Recommendations On The Validation Of Software Programs Performing Biostatistical Calculations For Forensic Genetics Applications, co-authored by Drs. Buckleton and Coble..... 25

Regulatory Guide: Verification, Validation, Reviews, and Audits for Digital Computer Software Used in Safety Systems of Nuclear Power Plants (July 2013)..... 26

Riman, Iyer, and Vallone, Examining performance and likelihood ratios from two likelihood ratio systems using the PROVEDIt dataset, 16 PLoS One 9 (2021)..... 100

Sarah Riman et al., Examining Performance And Likelihood Ratios From Two Likelihood Ratio Systems Using The Provedit Dataset, 16 PLoS One 9 (2021) ..... 109

Sarah Riman et al., Exploring DNA Interpretation Software Using The Provedit Dataset, 7 FSI: Genetics Supplement Series 724 (2019) ..... 108

Scientific Working Group on DNA Analysis Methods, Guidelines for the Validation of Probabilistic Genotyping Systems 2 (June 2015) ..... 16, 71, 72, 128

Software Considerations in Airborne Systems and Equipment Certification..... 25

Stacy Cowley and Jessica Silver-Greenberg, These Machines Can Put You in Jail. Don't Trust Them. N.Y.Times (Nov 3. 2019) ..... 13

Steven P. Lund and Hari Iyer, Likelihood Ratio as Weight of Forensic Evidence: A Close Look, 122 J. Research of Nat'l Inst. Standards & Tech. (2017) ..... 142

STRmix.com, Incorrect Comments Relating to STRmix in the State of New Jersey v. Corey Pickett (Feb. 16, 2021)..... 12

## TABLE OF AUTHORITIES (CONT'D)

**PAGE NOS.**

**Other Authorities (Cont'd)**

Susan A. Greenspoon et al, A Tale of Two PG Systems: A Comparison of the Two Most Widely Used Continuous Probabilistic Genotyping Systems in the United States, 69 J. Forensic Sci. 1840 (2024) ..... 21

Tim Kalafut, John Buckleton, et al., Investigation Into the Effects of Mixtures Comprising Related People on Non-Donor Likelihood Ratios, and Potential Practices to Mitigate Providing Misleading Opinions, 59 FSI: Genetics 1 (2022) ..... 63

Tim Stelloh and Brenda Breslaur, Previously Unusable DNA Sample Now Evidence in the Quadruple Murder Trial of N.J. Uncle, NBC News (Dec. 27, 2024) .....111

U.S. Dept. Heath of Human Services, General Principles of Software Validation; Final Guidance for Industry and FDA Staff (Jan 11, 2002) ..... 25

William Thompson, Uncertainty In Probabilistic Genotyping Of Low Template DNA: A Case Study Comparing Strmix™ And Trueallele™, J. Forensic Sci. (Jan 2023).....21, 110, 168

## PRELIMINARY STATEMENT

The State is seeking to admit the results generated by a piece of software that has not been proven to be reliable under the standards that all software is held to, for which not a single rate of error has been provided, and which undertakes an analysis too complicated for any human to ever recreate or verify. The bulk of the evidence provided by the State in support of the reliability of this software, STRmix, is that lots of laboratories use it, and the developers think it's really good. That's not enough for STRmix to be admissible.

It is possible that STRmix is good sometimes. But DNA analysis is complicated, and the more complex a sample gets, the higher the rate of error is. The concern that STRmix errs is not hypothetical. According to its creator, even when STRmix is working as expected, it will produce false positives and false negatives. In other words, STRmix at its best is going to reach the wrong answers, and it is going to do it more frequently the more complex the DNA sample is. Without having any idea of its rate of error across sample types, it is impossible for this Court to determine, on its own, when STRmix is good enough to be admissible and when it is not. Only by establishing the boundaries of reliable analysis can it be determined that any given DNA analysis falls within those boundaries or not. Without any boundaries, no results can be admissible.

The versions of STRmix used in this case also do not meet the well-established standards for demonstrating the reliability of software. That is the unanimous opinion of three extremely well-credentialed software engineers. They were also unanimous that trying hard is not good enough when it comes to software. The reliability of software has to be demonstrated through independent verification and validation that the software was built correctly. It was undisputed at this hearing by anyone with relevant expertise that that demonstration has not been made.

Even if STRmix were foundationally reliable across a certain range of samples—a range we do not know—it was used unreliably by the two laboratories in this case when analyzing these samples. These laboratories analyzed samples that they have not demonstrated they can reliably analyze through internal validation, a validation they are required to perform in order to ascertain what kinds of samples STRmix can reliably analyze and what kind of samples it cannot reliably analyze. Every laboratory that implements a new forensic technique must stick to the kind of work that it has demonstrated it can reliably do. Neither laboratory did that in this case. Even the most reliable software in the world would not yield reliable results when used unreliably in the laboratory.

Most concerning is that almost every sample in this case is a mixture in which related people are hypothesized to have contributed. It is an unchangeable law of genetics that people are more likely to be falsely included as contributors to mixtures their relatives actually contributed to. Neither the developers of STRmix nor the laboratories that analyzed the samples in this case have quantified this risk of false inclusion or demonstrated that it is low enough for STRmix to be reliable when applied to mixtures that may contain related people. In this case, the State's reach exceeds its grasp: whatever STRmix may reliably analyze, if anything, the samples in this case are not in that category.

Maybe STRmix is good sometimes. Maybe it's even sometimes good enough to be admissible. Maybe across some range of samples it does a reliably good job. The problem is that we do not know what that range, if it does exist, is. And we don't know that the software will not run into trouble even in that range, even if the mathematical modeling used by STRmix is correct, because of code errors. And we certainly do not know that that range encompasses the samples in this case, analyzed by the specific laboratories in this case.

As the gatekeeper, this court cannot simply take the word of the people who built STRmix—and who have staked their professional reputations and livelihoods on it—and those who paid for and implemented STRmix—because it increases their ability to close cases—that it is reliable. This Court needs proof. The State has failed to provide that proof. The evidence must be excluded.

## PROPOSED FINDINGS OF FACT

### A. Two Versions of STRmix Were Used by Two Different Laboratories to Analyze DNA Evidence in this Case

On November 20, 2018, the Monmouth County Prosecutor's Office responded to two fires, one at 15 Willow Brook Road in Colts Neck, where Jennifer, [redacted] and Keith Caneiro were found dead, and one at defendant Paul Caneiro's home at 27 Tilton Drive in Ocean Township. At each location, police collected items they believed may have DNA evidence on them. They also collected DNA profiles from all five members of the Caneiro family.

These items, along with swabs of other evidence collected, were sent for DNA analysis at the New Jersey State Police Office of Forensic Sciences. (S184-A) For at least some of these items, the DNA profile of the minor contributor was not of sufficient quantity or quality for comparison purposes. (S184-A)

In an attempt to gather DNA evidence from items that the New Jersey State Police was unable to analyze, the Monmouth County Prosecutor's Office sent evidence to Bode Technology, a private company, on June 18, 2020. (Bode Case File (D-60) at 1). Bode processed that evidence using STRmix v2.5.11 and returned statistics for 13 samples. (4T 37-7 to 13) The results returned by Bode supported Paul's inclusion as a minor contributor to 8 samples that were determined to be mixtures, with [redacted] as a major contributor to all of the same samples. (Bode Case File (D-60) at 5). The analyst who ran STRmix at Bode chose not to run some of the Caneiros through

STRmix, instead visually excluding them. (Bode Case File (D-60) at 5). All of the potential contributors analyzed in this case are related to at least two other potential contributors, and most of them are related to three. (Bode Case File (D-60) at 5)

On May 23, 2023, The Prosecutor's Office requested that the New Jersey State Police Laboratory (NJSP) conduct DNA analysis on new pieces of evidence in its own laboratory on a new version of STRmix— v2.8.0. (4T 63-25) As to Item 6-1-4-1, NJSP used STRmix to conclude that Jennifer and Paul were unlikely to be contributors and that Keith was not a contributor. (S-167) The NSJP analyst took STRmix's list of genotype weights it was 99% sure of, made a single-source profile with that genotype, compared it to [redacted] and determined that he was the "source" of the DNA using a Random Match Probability. (9T 82-19 to 84-7)

An Olenowski hearing on the admissibility of both the Bode and the NJSP evidence was held from November 12 through December 13, 2024.

**B. Witnesses Presented**

Below is an overview of each witness's education and experience in their fields. Their opinions are relayed in subsequent subsections. The witnesses' understanding, or lack thereof, of STRmix and of software engineering varied widely. Those who showed a deeper knowledge provided opinions that were more helpful to the Court. The court also notes the professional and personal incentives many of the witnesses have to support the acceptance of STRmix. At bottom this case is resolved based on the underlying scientific principles and their application to this case: "Good scientific research simply does not depend on the credibility of individual witness." State v. Olenowski, 255 N.J. 529, 580 (2023) (internal quotation marks omitted).

**1. State's Witnesses**

Seven witnesses testified for the State. While each had impressive credentials, their careers are inextricably tied to an inherent interest in promoting STRmix. It is the opinion of the

State witnesses that STRmix is reliable. (1T 109-4 to 7; 3T 55-18 to 22; Coble Report, Oct. 26, 2023 (D-22) at 7). Not a single software engineering expert testified on behalf of the State.

**a. Monica Ghannam**

Monica Ghannam was qualified as "an expert in the field of forensic DNA analysis and forensic validation." (1T 45-5 to 12) Ms. Ghannam currently serves as the DNA technical leader for the biology section of the Union County Prosecutor's Office Forensic Laboratory and has been employed in that role since 2005. (1T 12-14 to 25) Ms. Ghannam has been employed as practitioner in forensic science since 1994 after her graduation with a Master's in Science degree from Kings College in London and has worked in both public and private sector settings. (1T 15-1 to 16-24)

In her current role, Ms. Ghannam explained she was responsible for starting the DNA unit of the lab and, as the technical leader, is responsible for implementing new technologies and equipment including validation of those items and training personnel how to use them. (1T 17-6 to 12)

As a forensic science practitioner, Ms. Ghannam testified competently about the procedures she followed but was unable to provide any insight about the underlying functions about STRmix as a program, the diagnostics it utilized, the mathematical algorithms it relied upon, or how and why the software operated the way it did. (2T 4-9 to 8-20)

**b. Kristen Naughton**

Kristen Naughton is the Director of Validation and Quality Control at Bode Technology and was qualified as an expert in the field of forensic DNA analysis and forensic DNA validation. (2T 76-3 to 7) Ms. Naughton has a Bachelor of Science degree in biochemistry and Bachelor of Arts in chemistry with a minor in genetics from North Carolina State University. (2T 2-6) Ms. Naughton has been employed by Bode since 2003 and stated that she a "validation

scientist." (2T 56-1 to 15) Ms. Naughton testified that her agency was accredited by ANSI and is also accredited to the ISO 17025:2017 standards as well as the FBI Quality Assurance Standards. (2T 66-21 to 67-2)

Ms. Naughton indicated the decision to utilize STRmix versus some other probabilistic genotyping software was "mainly client driven" since many of Bode's clients were already implementing STRmix; however, she emphasized that she has been satisfied with that decision. (2T 72-3 to 15). Like Ms. Ghannam, Ms. Naughton had no insight into STRmix software beyond how it was implemented in her laboratory and could not explain many of the concepts underlying STRmix or critical diagnostics used by STRmix. (3T 57-5 to 61-17).

**c. Danielle Reed**

Danielle Reed is the Bode Technology forensic DNA analyst who conducted the forensic analyses the State seeks to introduce as evidence in this case. Ms. Reed has a Bachelor of Arts in criminology and criminal justice and a Bachelor of Science in physiology and neurobiology, both from the University of Maryland. (3T 134-2 to 5) Ms. Reed testified that she has analyzed hundreds of thousands of DNA samples in total over the course of her career and has testified in court 47 times in multiple jurisdictions. (3T 139-19 to 140-6) In the present case, Ms. Reed was qualified as an expert in the field of DNA analysis. (3T 143-4 to 8)

Like Ms. Ghannam and Ms. Naughton, Ms. Reed is a practitioner and was unable to provide significant insight into how STRmix was developed, how the software works, or what the diagnostics mean. (5T 99-6 to 105-2)

**d. John Buckleton**

Dr. John Buckleton was the State's lone expert with knowledge specifically about STRmix's development; the operation of the Institute of Environmental Science and Research (ESR), which developed STRmix; and the relationship between STRmix, ESR, Forensic Science

South Australia (FSSA), and Orbit, all of whom have some role in the development and testing of STRmix. As one of STRmix's creators, Dr. Buckleton has been on the forefront of probabilistic genotyping software since 2011. (6T 8-10 to 18) Despite primary degrees in chemistry, most of Dr. Bucketon's work dating back to 1988 has centered around interpretation of physical evidence. His experience in forensics includes interpretation of glass physical evidence, the beginning of DNA interpretation as a form of evidence in the 90s, and finally probabilistic genotyping. (6T 9-10 to 11-10) Dr. Buckleton does not have any degree in software engineering or coding despite some experience coding in the 80s through the mid-90s. (6T 13-5 to 19). Currently, Dr. Buckleton is the principal scientist at ESR, with whom he's been continuously employed since prior to the creation of STRmix. (5T 5-22 to 23) Dr. Buckleton was qualified as an expert in the fields of DNA analysis, probabilistic genotyping, and software development. (6T 41-24 to 43-12)

The defense objected to qualification of Dr. Buckleton as an expert in software development. (6T 42-8 to 23). Dr. Buckleton is not an expert in software development, having not been involved in any coding in two decades. However, even if he were to be considered an expert in software development, it is important to note that he is not—and the State did not offer him as—as an expert in software engineering. This distinction is critical. Software development is just writing code for software, while software engineering is the scientific discipline concerned with the proper building and testing of an entire software system. (13T 18-3 to 19-1; see also 25T 62-3 to 13 ("[S]oftware engineering looks at the overall design of a system as a whole, a large system. . . . Software development on the other hand is writing the actual code.") Although Dr. Buckleton may have some coding knowledge, he does not have expertise in the field of software engineering, which is the field relevant to assessing the reliability of software.

Dr. Buckleton provided the court with a thorough explanation of STRmix's history and helped the court understand his contributions to STRmix as well as the contributions of his co-founder Dr. Duncan Taylor. Dr. Buckleton credited Dr. Taylor with creating the original source code for STRmix singlehandedly and described himself as primarily responsible for the mathematical algorithms behind STRmix. (6T 14-1 to 24) Dr. Buckleton also asserted that he was responsible for much of STRmix's testing. (6T 21-2 to 17) As demonstrated by his posts about the hearing on his LinkedIn and his watching almost all of the hearing in-person or remotely, Dr. Buckleton is very invested in the outcome of this case.

Dr. Buckleton's knowledge of STRmix and his experience in DNA are impressive. He has held numerous positions on governing bodies that regulate DNA and forensic science, including SWGDAM, OSAC, and the International Society of Forensic Geneticists. (6T 38-13 to 23) He also worked at the National Institute of Standard and Technology (NIST). (7T 110-14 to 15) He testified that he has "a little over 250 publications," of which over half are DNA related. (6T 24-11 to 12)

**e. Jennifer Thayer**

Next to testify was Jennifer Thayer, Lab Director of the New Jersey State Police Office of the Forensic Sciences, DNA Laboratory. Ms. Thayer has a Bachelor of Science in biochemistry and sociology from Muhlenberg College and a Master of Science in forensic science from the University of New Haven. (8T 7-7 to 19) Ms. Thayer shared that she has worked for that agency for about 20 years and stepped into her most recent role as lab director in 2022. (8T 4-16 to 25) In her current role, Ms. Thayer is responsible for all of the day-to-day operations of the laboratory and also serves as the assistance technical leader. (8T 5-13 to 19). Ms. Thayer was accepted as an expert in the fields of forensic DNA analysis and forensic DNA validations.

Ms. Thayer explained that the State Police lab is accredited to ISO 17025 through the ANSI National Accreditation and is also accredited to the FBI Quality Assurance Standards. (8T 18-18 to 19-21). Ms. Thayer also testified that she had been actively involved in validations of new technologies at the State Police lab since 2016 and that she was actively involved in validation of STRmix, which she noted was the largest validation or her career. (8T 11-21 to 25 and 221-21 to 22-2).

**f. Christine Schlenker**

Christine Schlenker is a forensic DNA analyst who has worked with the State Police since 2002 and was the technical reviewer of the original analyses done by the State Police in this case. (8T 76-9 to 19). Ms. Schlenker has Bachelor of Science degrees in biology and a minor in chemistry from the University of New Haven and a master's degree in pharmaceutical sciences with a concentration in forensic DNA and serology from the University of Florida. (9T 69-1 to 7) She was accepted by the court as an expert in forensic DNA analysis. (9T 65-13 to 18) Like the other practitioners in who testified, Schlenker was able to explain how she used STRmix for analysis but is not a software engineer and was unable to provide significant insight into how STRmix was developed, how the software works, or what the diagnostics mean. (9T 101-13 to 19).

**g. Michael Coble**

The State's final witness was Dr. Michael Coble who qualified as an expert "in the field of forensic DNA analysis, probabilistic genotyping and DNA mixture interpretation. (10T 38-19 to 25). Dr. Coble received a master's degree from George Washington University in the field of forensic science with a concentration in molecular biology and a Ph.D. in human genetics also from that university. (10T 9-22 to 10-10). Dr. Coble is on the boards of the Journal of Forensic Science, Forensic Science International, and WIRES Forensic Science. (10T 28-4 to 16). He has

published nearly 100 times and has co-authored a book on likelihood rations in DNA analysis. (10T 28-17 to 29-6).

Dr. Coble is currently employed by the University of North Texas Health Center where he is Executive Director of the Center for Human Identification, a forensic DNA laboratory that analyzes evidence for law enforcement, and is also a professor in the microbiology, immunology, and genetics department. (10T 5-1 to 12; 11T 79-2 to 8) Dr. Coble was a prior chair of SWGDAM and served in thar role when the working group issued their guidance on validating probabilistic genotyping systems. (10T 60-2 to 20)

Dr. Coble was an early adopter of STRmix in the United States after being invited by Dr. Buckleton to New Zealand to learn about STRmix in 2012. (10T 44-14 to 22) Since that time, Dr. Coble has a significant publication history with not only Dr. Buckleton but also with STRMix co-founders Dr. Duncan Taylor and Dr. JoAnne Bright. (10T 52-20 to 53-6) Like Dr. Buckleton, Dr. Coble has also worked at NIST. (10T 9-1 to 5)

**2. Defense Witnesses**

The defense presented five witnesses: two experts in forensic DNA, two experts in software engineering, and one expert in both software engineering and probabilistic genotyping. All five defense witnesses testified about concerns with STRmix and with STRmix's use in this case.

**a. Karl Reich**

Dr. Karl Reich was qualified as an expert in the field of forensic DNA analysis. (12T 15-2 to 8) Dr. Reich is the laboratory director of a small forensic DNA laboratory where he has worked since 2002. (12T 8-4 to 9) Dr. Reich has an undergraduate degree in chemistry from Cornell and received his master's in molecular biology from UCLA, although much of his thesis work was completed at Harvard Medical School. (12T 8-16 to 9-6)

Dr. Reich has significant knowledge in validation and accreditation in the context of a forensic DNA lab. Dr. Reich reported that his own lab is accredited by three independent auditing agencies. (12T 10-2 to 6) In his current role, Dr. Reich has conducted internal validations approximately 25 times. (12T 10-17-20)

Dr. Reich does not use STRmix in his laboratory. (12T 134-21 to 25) As a retained expert in a post-conviction case, has written an affidavit to support a defense motion for someone at his laboratory to swap an item for DNA. (S-188) The swab would be sent to Bode. (S-188) Dr. reich explained that he used to refer clients to other laboratories for testing to avoid any potential conflict of interest, but he has stopped doing so because has come to believe his laboratory does better and more neutral work than most commercial laboratories. (12T 211-10 to 15)

**b. Mats Heimdahl**

Mats Heimdahl is the Head of the Department of Computer Science at the University of Minnesota. Dr Heimdahl has a Ph.D. in computer science from the University of California, Irvine. (13T 5-19 to 6-4) Dr. Heimdahl was qualified as an expert in the field of software engineering. (13T 14 to 18)

Safety-critical systems are the primary focus of Dr. Heimdahl's work. (13T 6-16 to 10-25) Dr. Heimdahl describes a safety critical system as one which "can cause great bodily harm, death, or great environmental danger" if the software is not properly implemented. (13T 7-12 to 18) In his career, Dr. Heimdahl has worked in the field of medical devices, avionics for the FAA, and aviation safety for NASA. (13T 12-9 to 13-10) Specific examples of his work include the Traffic Alert and Collision Avoidance System for the FAA, the display logic for pilots operating Boeing 787s, and a medical infusion pump which provides painkiller medication to patients automatically. (13T 9-13 to 10-25)

The State made some insinuation that Dr. Heimdahl's amicus brief in State v. Pickett, 466 N.J. Super. 270 (App. Div. 2021), somehow rendered him less credible, an insinuation that must be rejected. Now-Justice Fasciale found that brief helpful in reaching the holding that meaningful access to the source code and related development materials is necessary to ascertain the reliability of a piece of software. Id. at 298, 310. One component of this insinuation is that it was somehow incredible for Dr. Heimdahl to cite an article from the Courier Mail about "miscodes" in STRmix. (7T 54-2 to 57-15). Justice Fasciale cited it as well. Id. at 309, n.15. Both the article and the amicus brief stated that the "miscode impacted 60 criminal cases, requiring new likelihood ratios to be issued in 24 cases." (S-191 at 8); David Murray, Queensland Authorities Confirm "Miscode" Affects DNA evidence in Criminal Cases (March 20, 2015), available at https://bit.ly/34DBlZy. The State has pointed out nothing incorrect about the article. The quibble seems to be with the meaning of the word "impact." An author of a response to Pickett on the STRmix website agrees LRs were released in 24 cases; he just does not seem to think that matters: "In Queensland, the 24 statements were amended with a minor change to the LR in some cases. In all instances this change was before the court case was heard. The changes were minor not 'material' or 'outcome-determinative' and affected only a subset of the 60 cases." STRmix.com, Incorrect Comments Relating to STRmix in the State of New Jersey v. Corey Pickett (Feb. 16, 2021), https://www.strmix.com/assets/STRmix/STRmix-PDFs/STRmix_Response_State_of_NJ_v_Pickett_160221.pdf.

The second insinuation is that Dr. Heimdahl is less credible because the amicus brief cites a New York Times article about errors in breathalyzers to highlight the risks of error in forensic software. (13T 93-16 to 99-24). The State takes issue with the brief not emphasizing that thousands of breath tests thrown out in New Jersey "largely due to human errors and lax

governmental oversight." Stacy Cowley and Jessica Silver-Greenberg, These Machines Can Put You in Jail. Don't Trust Them. N.Y.Times (Nov 3, 2019). But Dr. Heimdahl's assertion that "thousands of faults have been discovered in the source code of top breathalyzers system" is both correct and supported by the article: as the article notes, the experts in our Supreme Court's breathalyzer admissibility challenge, State v. Chun, 194 N.J. 54 (2008), said the source code was "littered with thousands of programming errors." Don't Trust Them.

In short, Dr. Heimdahl did not say anything incorrect in his Pickett brief. The implication that the errors in these systems didn't "matter" and that therefore it is misleading to report them misses the point Dr. Heimdahl was making: errors exist in forensic software and, furthermore, can elude developers' notice. Further, as Dr. Heimdahl explained at the hearing, most problems in safety critical systems are not related to bugs in the code but rather in the failure to properly determine the requirements of the software, which can cause issues testing or verifying the software works correctly or using it in appropriate situations. (13T 8-16 to 9-9, 25-6 to 28-6) Those kinds of issues are still considered software failures.

**c. Nathan Adams**

Nathan Adams testified in this case as an expert in software engineering and probabilistic genotyping. (14T 19-21 to 20-2) Mr. Adams has a Bachelor of Science in computer science from Wright State University and is working on a master's in computer science. (14T 10-8 to 22) Mr. Adams is a systems engineer with Forensic Biometrics Services (FBS), a private forensic biology consulting company where he serves as the first point of contact for attorneys seeking his firm's consulting services on a range of issues including forensic DNA analysis. (14T 5-1 to 16). Mr. Adams has been with FBS since 2012, which coincided with the expansion of probabilistic genotyping software around the world; cases involving probabilistic genotyping software are now the majority of those he works on. (14T 7-14 to 25)

Mr. Adams's testified that in his role, he typically reviews the underlying data used in a specific forensic case analysis but also reviews a laboratory's standard operating procedures and a lab's internal validation studies. (14T 8-14 to 9-15) Mr. Adams stated that he had reviewed hundreds of versions of standard operating procedures from dozens of laboratories around the world and has worked on over a thousand forensic DNA cases. (14T 8-14 to 10-6)

Mr. Adams has also reviewed source code for probabilistic genotyping systems as part of his casework. One of the most prominent cases Mr. Adams has worked on is U.S. v. Johnson. In that case, Mr. Adams testified that he found an undisclosed modification that changed the probabilistic genotyping system FST's behavior from the behavior expected based on its original validation study. (14T 12-21 to 15-2) Mr. Adams has also inspected STRmix's source code previously. After testifying about that review in a criminal case, he was sent a cease-and-desist letter by Blake Gerney, a STRmix lawyer, which threatened to sue Mr. Adams for allegedly disclosing confidential information. (14T 58-3 to 24) That claim was resolved without any lawsuit. (14T 58-3 to 24)

**d. Paul Martin**

Dr. Paul Martin testified next for the defense and was qualified as an expert in the field of software engineering. (15T 20-8 to 13) Dr. Martin received a bachelor's degree, a master's degree, and a Ph.D. all from Johns Hopkins University in the field of computer science. (15T 10-1 to 3) Currently, Dr. Martin works as the chief scientist at Harbor Experts where he leads both research-based and litigation-based projects. (15T 5-15 to 20) Although Dr. Martin has done substantial work with many kinds of software, he developed an expansive record working in security analysis and teaches that subject at Johns Hopkins University. (15T 6-24 to 7-10) Dr. Martin also testified that he has multiple patents including a patent on finding vulnerabilities in software. (15T 12-12 to 13-13).

Dr. Martin has extensive experience with conducting source code review of commercial software and does multiple reviews "many times a year." (15T 15-17 to 21) He has conducted source code reviews in the healthcare space for a vendor with Blue Shield of California, in the financial sector for Loan Depot, and in the criminal context looking at True Allele, STRmix's major competitor in the U.S. (15T 15-20 to 18-13)

**e. Keith Inman**

The final witness to testify for the defense was Keith Inman, who qualified as an expert in the fields of probabilistic genotyping and forensic DNA analysis. (16T 70-14 to 20) Mr. Inman is currently employed as an associate professor at California State University East Bay and is a visiting fellow at the University of Dundee in Scotland. (16T 57-13 to 16) Mr. Inman has both bachelor's and master's degrees in criminalistics from the University of California Berkely and is a Ph.D. candidate at the University of Dundee. (16T 57 19 to 22) Mr. Inman has worked in the field of criminalistics for over 40 years, beginning his employment with the Orange County Sheriff's Crime Laboratory, then moving to the California Department of Justice DNA laboratory, then spending time with a private forensic science firm before joining academia. (16T 58-5 to 15)

Mr. Inman has direct experience with both validation studies and DNA case work. Dr. Inman began his time with the California Department of Justice DNA laboratory working exclusively on validation as the state moved to get its DNA laboratory up and running. (16T 58-21 to 59-23) He also testified that he has analyzed thousands to tens of thousands of forensic samples firsthand. (16T 61-1 to 3)

Mr. Inman has served on the Organization of Scientific Advisor Committee on advisory committees for both DNA extraction and quantitation and also the committee for probabilistic genotyping validation, development, and documentation. (16T 61-10 to 63-2) Additionally,

Inman is a member of NIST's DNA Foundational Review Advisory Board and was part of the committee responsible for the 2021 Foundation Review for DNA Mixture Interpretation. (16T 69-1 to 70-13)

**C. A basic overview of STRmix**

Details about many facets of STRmix are discussed in sections D, E, and G, infra. Below is a simplified overview of the purpose of STRmix and its final output, the likelihood ratio, in order to provide context for what follows.

1. **STRmix is both a mode of DNA interpretation and piece of software. The science of both forensic DNA and software engineering are relevant to assessing its reliability.**

STRmix is a probabilistic genotyping system. Probabilistic genotyping was invented to try to address the problems of working with complex DNA mixtures that present significant challenges to reliable interpretation. In the words of our Appellate Division, probabilistic genotyping software is "designed to address intricate interpretational challenges of testing low levels or complex mixtures of DNA." State v. Pickett, 466 N.J. Super. 270, 277 (App. Div. 2021). Probabilistic genotyping attempts to do so through the "use of biological modeling, statistical theory, computer algorithms, and probability distributions to calculate likelihood ratios (LRs) and/or infer genotypes for the DNA typing results of forensic samples." Scientific Working Group on DNA Analysis Methods, Guidelines for the Validation of Probabilistic Genotyping Systems 2 (June 2015) ("SWGDAM Guidelines") (S-6/D-202). In the words of Dr. Buckleton, "STRmix is a software that assists with DNA interpretation." (6T 5-8 to 9) There are many different probabilistic genotyping systems that use different mathematical models and are developed by different people. National Institute of Standards and Technology, Supplemental Document to DNA Mixture Interpretation: A NIST Scientific Foundation Review at 12 (Dec. 2024), https://nvlpubs.nist.gov/nistpubs/ir/2024/NIST.IR.8351sup2.pdf.

Because STRmix is a software that is used to analyze forensic DNA samples, both the science behind software engineering and the science behind forensic DNA analysis are relevant to assessing the reliability of STRmix in general and as applied to this case. Details of each of those fields are discussed at length below.

**2. STRmix attempts to separate out the DNA profiles of each person who contributed to a mixture through statistical theory, computer algorithms, and probability distributions.**

In a nutshell, STRmix is generally used on DNA samples that are presumed to contain more than one person's DNA. STRmix attempts to "deconvolute" or separate out the genetic information in that sample to determine the genetic profiles of each person who contributed to the sample. See 12T 42-2 to 8 (Dr. Reich: "[T]he problem with mixtures is trying to deconvolute, separate the data into different buckets. So, you have several individuals' DNA on a sample and you would like to sort of separate out those profile data into bucket one, person one, bucket two, person two so that you can better understand who might be present. And that can be a very challenging issue.").

Probabilistic genotyping interprets these complex mixtures by using "weighting (based on the probability of) specific genotype contributions through biological and statistical models informed by probabilities of missing alleles. These methods incorporate mathematical modeling that can reflect uncertainty in genotype combinations for the mixture interpretation[.]" John M. Butler et al., National Institute of Standards and Technology, DNA Mixture Interpretation: A NIST Scientific Foundation Review 47 (Dec. 2024) ("Mixture Interpretation").[1] As Dr.

[11] Exhibit D-8 is a draft of this report. A final version of this report was issued by NIST in December after the conclusion of the hearing. Counsel sent a letter to the Court containing the final report. All references in the brief are to the final report.

Buckleton explained, STRmix runs through possible genotyping combinations to determine which ones are the best fit for the evidence by using a sample method known as the Markov Chain Monte Carlo (MCMC). (6T 60-16 to 61-16) Simply put, using the MCMC and "wandering chains," STRmix tests out thousands or millions of different explanations for the DNA observed in a sample—by "explanations," what is meant is different DNA profiles being attributed to different contributors to a sample. (7T 183-19 to 186-13) STRmix reports how likely it believes the observed DNA mixture is given a hypothetical genotype as "genotype weights," expressed as a percentage. (7T 177-15 to 17) The process of trying to sort out what the profiles are of each contributor is referred to as deconvolution. (1T 61-9 to 17)

**3. STRmix produces a likelihood ratio to express the probability of observing the DNA evidence under two different hypotheses**

After the deconvolution, STRmix goes on to interpret the evidence. Its final output is a likelihood ratio, often referred to as an "LR." The LR is expressed as a fraction. "The LR involves a ratio of two conditional probabilities: the probability of the evidence given that one proposition (hypothesis or narrative) is true and the probability of the evidence given an alternative proposition is true." Mixture Interpretation at 48. An LR above 1 provides support for the hypothesis in the numerator, while an LR less than 1 provides support for the hypothesis in the denominator. Ibid. In the example above, an LR above 1 provides support for inclusion of the person of interest in the sample, and an LR below 1 provides support for the exclusion of the person of interest in the sample. (8T 74-21 to 25)

A likelihood ratio compares the evidential support for two competing hypotheses. The hypotheses are chosen by people—they represent subjective choices made by investigators or analysts as to what scenarios they think are likely and should be tested. The LR is the

comparison of these two hypothetical scenarios; they do not report the overall likelihood of a person being present in a DNA mixture independent of these hypotheses.

Likelihood ratios are also meaningless without reference what is referred to as a "prior probability"—a pre-existing belief about how likely a given hypothesis is to be true. National Institute of Standards and Technology, Forensic DNA Interpretation and Human factors: Improving Practice Through a Systems Approach 81 (May 2024) (D-7). Under Bayes's Theorem, the mathematical framework for the use of likelihood ratios, the likelihood ratio has to be multiplied by this prior probability (based on other case-relevant information) to arrive at the posterior probability (which will often be the probability that the suspect was the source of a crime scene sample). Ibid.

The LR "is easily, and most likely taken as the probability of the defendant's guilt. To be clear, this is a grave mistake. LRs are not, in fact, probabilities. They merely express the probabilistic relationship between two hypotheticals." Bess Stiffelman, No Longer the Gold Standard: Probabilistic Genotyping Is Changing the Nature of DNA Evidence in Criminal Trials, 24 Berkeley J. Crim. L. 110, 118-119 (2019). Thus, it is important to keep in mind that "[t]he most supporting proposition is not necessarily the most probable[.]" Human Factors at 80. Only two hypotheses being compared: there may be many more that would be even more supported. Even assuming all relevant hypotheses were tested—which does not always happen because law enforcement selects the hypotheses to test—"the probability of the proposition depends on all the elements in the case and therefore only partly on the DNA results. This is why the so-called posterior probability is not in the domain of the DNA analysis." Ibid. Great care must be taken to accurately convey the meaning of the likelihood ratio.

**a. Different systems, assumptions, and users will lead to different likelihood ratios being produced.**

Different probabilistic genotyping systems analyzing the same evidence will come to different results, as will the same systems being used by different analysts. Mixture Interpretation at 47. A PGS system will produce different results based on the modeling choices made by the designers, data input choices by the analyst, proposition choices and assumptions made by the analyst including the selection of the number of contributors to a sample, and the population database used by a laboratory to provide allele and genotype frequency estimates. Id. at 51.

These differences are not merely theoretical and should not be assumed to be minimal. "There appears to be a general misconception that LR assessments made by different experts will be close enough to one another to not impact the final DNA mixture interpretation conclusions. Although they may be similar in many instances, this is not known for any particular case, and it is not advisable to take this for granted." Mixture Interpretation at 54.

For example, in one paper, different amplifications of the same sample on the same PGS in the same laboratory created LRs from 10 to 100,000. (11T 59-1 to 20) This means that under one analysis, the sample was deemed 10 times more likely if the defendant was a contributor than if it originated solely from unknown contributors, while another analysis, conducted by the same laboratory on the exact same sample, concluded that the sample was 100,000 times more likely if the defendant was a contributor to the sample than if the sample originated solely from unknown contributors. Although both results are positive LRs, the strength of the conclusions is vastly different. (On the SWGDAM verbal scale, discussed further in subsection F.11, 10 provides "limited support" for the prosecutor's hypothesis and 100,000 provides "strong support"). Similarly, another study demonstrated that STRmix and another PGS, TrueAllele, both produced exclusionary LRs when run on the same sample, but the one produced by TrueAllele was five to six orders of magnitude (100,000 to 1,000,000 times) greater. William Thompson,

Uncertainty In Probabilistic Genotyping Of Low Template DNA: A Case Study Comparing Strmix™ And Trueallele™, J. Forensic Sci. at 6 (Jan 2023) (D-1208). Dr. Buckleton writes that changes in STRmix code have impacted likelihood ratios beyond a single order of magnitude (impacting the LR by more than a factor of 10). (D-21 at 3)

Studies have also shown that different probabilistic genotyping systems may provide completely contradictory LRs, with one providing an LR that supports the first hypothesis and one providing an LR that is non-supportive. Susan A. Greenspoon et al, A Tale of Two PG Systems: A Comparison of the Two Most Widely Used Continuous Probabilistic Genotyping Systems in the United States, 69 J. Forensic Sci. 1840, 1856 (2024). In other words, one PGS would provide support for the hypothesis that the defendant had contributed to the sample, while another PGS would provide support for the exact opposite hypothesis—that the defendant did not contribute to the sample.

In sum, the LR produced by PGS will vary, sometimes widely, from system to system and even analyst by analyst using the same system.

**b. There is no one true likelihood ratio for any given sample.**

"For any given sample, there is no single, true likelihood ratio." Mixture Interpretation at 54. Because there is no one, true, correct LR for any given sample, knowing whether the LR produced reflects the correct operation of a specific PGS as designed will often be impossible. See e.g., Hinda Haned, et.al., Validation Of Probabilistic Genotyping Software For Use In Forensic DNA Casework: Definitions And Illustrations, 56 Sci. & Justice 104, 108 (2016) ("Model validation for use in forensic casework is not straightforward because the true weight of the DNA evidence cannot be determined; indeed, the generated [likelihood ratio] always depends on the model's assumptions, no 'gold standard' exists in the form of a true likelihood ratio that can serve as a comparison."). Unlike a Random Match Probability, generally used in manual

DNA analysis, which can be done by hand and verified independently, a "likelihood ratio has no precise, independently ascertainable value with which to compare to ensure that the software is providing an acceptable estimation." Pickett, 466 N.J. Super. at 322.

**D. Software engineering has rigorous standards to ensure the reliability of important software such as STRmix**

Because STRmix is a piece of software, software engineering is the discipline that sets forth the standards for how to correctly build and test it and for when the software has been demonstrated to be sufficiently reliable to meet its intended purpose. "Software engineering is an engineering discipline that is concerned with all aspects of software production from the early stages of system specification through to maintaining the system after it has gone into use." Ian Sommerville, Software Engineering 21 (2016, 10th Ed.) (D-5) See also 13T 42-6 to 8 (Dr. Heimdahl explains that this textbook is one of the most widely used textbooks in the field of software engineering). Software engineering consists of the engineering discipline, which is concerned with the "technical process of software development," and "software product management." Id. at 21-22. In this case, we are concerned with the former.

Software engineering is a scientific field that is concerned with every aspect of software development, not just writing code. Software engineering encompasses, rather, "the big picture, where you really need to understand what is the problem, what are the requirements, what is the solution that can solve this problem for a customer or the users, how to document that, how to validate that you actually got it right." (13T 19-1 to 13) The software engineering field has standards and processes to attempt to prevent errors in software.

As explained below, every single software engineer who testified agreed that these standards exist and must be scrupulously adhered to because software faults are ubiquitous and are very hard to detect, often slipping through the cracks even in incredibly expensive projects

that hundreds of people have worked on, and safety-critical systems such as software used to run planes. In order to determine that important software that has the potential to impact life and limb is sufficiently reliable to be used, software must be independently verified and validated. Verification and validation (V&V) occurs for every kind of software and independent verification and validation (IV&V) occurs for all safety-critical systems. The three software engineering experts and Dr. Buckleton agreed that STRmix is a safety-critical system that must meet the rigorous standards of IV&V. IV&V is the only method to demonstrate the reliability of software. Academic articles, peer-reviewed or otherwise, are not accepted as substitutes for IV&V by software engineers. All of these principles, unanimously agreed upon by the only software engineers who testified at the hearing, are explained in detail below.

1. **Software faults are common and difficult to detect.**

Software faults are common and often difficult to detect. (13T 26-1 to 30-23, 55-12 to 56-8;15T 22-12 to 23-14; Heimdahl Report, June 23, 2024 (D-10) at 6) Faults can range from errors in the code—"bugs"—to user interfaces that are too challenging for the user to understand, to the failure to anticipate conditions under which software might fail. All of these kinds of faults are considered software faults that should be prevented by rigorous adherence to software engineering standards when building and testing software. These faults, and resulting catastrophic errors, occur even in software that is incredibly expensive, important, and extensively tested.

For instance, a small error in calculations for the software for the Mars Lunar Orbiter, designed by NASA, was undetected until the satellite hit Mars and crashed. (13T 56-1 to 8) The reason the Therac-25 radiation device overdosed six people receiving cancer treatment was in large part because of choices made about how the software would look to the user; a modification to make it easier to enter a patient's treatment plan combined with error messages

issued by the software that were too hard for an operator to understand were two of the core issues that led to the malfunction. Nancy Leveson, Medical Devices: The Therac-25 in Safeware: System Safety and Computers at 6-9 (1995) (in reviewing the overdoses, a FDA memorandum noted that "[t]his software package does not appear to contain a safety system to prevent parameters being entered and intermixed that would result in excessive radiation being delivered to the patient under treatment," which combined with "cryptic" and overly frequent error messages resulted in the overdoses). See also 15T 21-16 to 22-8; Heimdahl Report, June 23, 2024 (D-10) at 20. The Y2K "bug" occurred because software designed in the 1990s did not have the capacity to store four numbers. (15T 23-17 to 23) The people who designed software did not realize that when the year 2000 came, the software would perceive the year as 00 or 1900, which could create significant problems with the use of the software. (15T 23-18 to 24-12)

In sum, software faults run the gamut and are not solely caused by errors in the code itself. As discussed below, the field of software engineering has developed standards for the building and testing of software to ensure it is as error-free as possible. Those standards are the minimum that must be met in order for software to be considered reliable, not the maximum. Software that does not meet those standards is not considered reliable.

**2. Software is determined to be reliable only after it has been verified and validated.**

All three software engineering experts agreed that verification and validation is the only way that the reliability of software is ensured according to software engineering standards. (13T 37-2 to 13; 14T 32-12 to 18; 15T 31-19 to 32, 100-1 to 12) All types of software are verified and validated. (13T 38-5; 14T 35-19 to 25; 15T 32-6 to 13) Software that has not been verified and validated is not considered reliable. (Heimdahl Report, June 23, 2024 (D-10) at 12-14, 18-20; Adams Report, May 13, 2022 (D-16) at 5-7)

There are many standards for verification and validation. There are no software engineering standards specific to probabilistic genotyping. (14T 29-24 to 25) Standard 1012 of the Institute of Electrical and Electronics Engineers ("IEEE 1012") is a leading global standard for software verification and validation. (D-200; 13T 31-3 to 6) The IEEE is very well-respected in the software engineering field. (15T 36-3 to 4) IEEE 1012 "offers a common framework that can be used to demonstrate that which is a consensus base standard by the largest professional software engineering organization in the world." (14T 30-10 to 15) The International Society of Forensic Genetics claims to incorporate, to at least some degree, IEEE 1012 in its Recommendations on the Validation of Software Programs Performing Biostatistical Calculations for Forensic Genetics Applications, co-authored by Drs. Buckleton and Coble. (D-203 at 192 ("International industry standards apply to software validation, verification [citation to IEEE-1012] and test documentation. These standards can be simplified and extrapolated to forensic genetics") (some citations omitted)).

Many industries that produce important software have their own standards. (13T 31-3 to 14) The aviation industry requires software verification and validation, including the Federal Aviation Administration, DO-178C, Software Considerations in Airborne Systems and Equipment Certification (D-216), the National Air and Space Administration, NASA Procedural Requirement 7150.2D, NASA Software Engineering Requirements (eff. March 8, 2022) (D-207). The medical device industry requires software verification and validation. IEC 62304, Medical device software – Software life cycle processes, (D-217). The U.S. Food and Drug Administration requires software verification and validation. 21 U.S.C.A. § 360j; 21 C.F.R. § 820.1 et seq; U.S. Dept. Heath of Human Services, General Principles of Software Validation; Final Guidance for Industry and FDA Staff (Jan 11, 2002) (D-208), as does the Department of

Defense. 48 C.F.R. § 209.57-1(3)(i)(G). The U.S. Nuclear Regulatory Commission requires verification and validation. U.S. Nuclear Regulatory Commission, Regulatory Guide: Verification, Validation, Reviews, and Audits for Digital Computer Software Used in Safety Systems of Nuclear Power Plants (July 2013) (D-206). Software, including STRmix, could be demonstrated to be reliable without following IEEE-1012, but in that case the standards would have to be scrutinized to ensure that all essential V&V activities are included. (14T 30-15 to 19).

**3. STRmix is a safety-critical system.**

All three software engineering experts, as well as Dr. Buckleton, agreed that STRmix should be considered as a piece of software that has severe consequences for failure, including loss of life or liberty; these types of systems are known as safety-critical systems. (6T 116-16 to 117-12; 13T 8-1 to 12, 32-12 to 14; 14T 37-1 to 7; 15T 37-17 to 23; D-13 at 34; D-21 at 9) In IEEE-1012 parlance, all software engineers agreed that STRmix should be considered an integrity level 3 or 4 system. (14T 37-5 to 7; 97-15 to 98-22) Dr. Buckleton claims that STRmix is an integrity level 4 system and meets IEEE-1012 requirements for that system. (6T 116-11 to 117-1) IEEE agrees that probabilistic genotyping systems are integrity level 4 system. (D-251). Because probabilistic genotyping systems are a safety-critical system that must meet integrity level 4 standards, IEEE explains, "DNA mixture interpretation using DNA software should only be deemed reliable based on objective information gathered through independent verification and validation." (D-251 at 3)

**4. In order to be considered reliable, safety-critical software must be verified and validated independently.**

Safety-critical software must meet the highest standards of independent verification and validation. The standards are highest for this kind of system because of the severe consequences of failure for this kind of software: "loss of life, loss of mission, significant social loss, or

financial loss." IEEE 1012-2016 at 199 (D-200). Integrity level 4 products go through the most rigorous level of independent verification and validation (IV&V). As IEEE explains, "DNA software should only be deemed reliable based on objective information gathered through independent verification and validation as determined by IEEE Standard 1012." IEEE, Response, NIST Internal Report 8351-Draft DNA Mixture Interpretation: A NIST Foundation Review 3 (2021) (D-251). Under this standard, STRmix can be considered reliable only if there is demonstrated compliance with software engineering norms for independent verification and validation.

Verification and validation are discussed at length below. Before delving into the details, a general overview may be helpful. "The goal of the verification and validation processes is to establish confidence that the software system is 'fit for purpose.'" Sommerville at 288. That means the question is not just whether a system is abstractly "good," but whether it is "good enough for its intended use." Ibid. The intended use is important to define clearly. As IEEE explains, "No software or hardware is 'generally' reliable—any technology is only fit for certain purposes." (D-251 at 7) A "core premise of labeling a product or process as 'well-engineered'" is that the operating conditions under which the product will work reliably "are specifically defined, tested against pre-defined standards, and accompanied with estimated rates of failure." (D-251 at 7) A common way to understand the difference between verification and validation is that verification asks, "Are we building the product right?" and validation asks, "Are we building the right product?" Sommerville at 228; 13T 16-21 to 35-20; 14T 34-10 to 12.

A simple analogy may help. Let us imagine a bridge needs to be built. A very detailed blueprint is created for the bridge. Verification is checking the physical bridge against the blueprint to ensure that the bridge that was built was the bridge that was designed. But if a

perfect suspension bridge has been built exactly to specifications in the blueprints, but, due to boat traffic in the body of water that the bridge spans, it actually needed to be a drawbridge, that bridge cannot be validated. The bridge might have been perfectly built according to its architects and civil engineers, but it is not fit for purpose: spanning this particular waterway through which tall boats pass.

Implementing the most rigorous V&V standards is not a guarantee that a piece of software is unflawed. As IEEE writes in its disclaimer, meeting these standards is necessary but not sufficient to ensure the reliability of a system and to avoid adverse consequences: "This standard establishes minimum criteria for V&V processes, activities, and tasks." IEEE 1012-2016 at 21; IEEE 1012-2012 at 7. Meeting these minimum standards cannot guarantee that there will be no software failures and adverse consequences. Ibid. But it is the bare minimum in the software engineering field to appropriately minimize those risks.

**a. Detailed requirements and specifications are a prerequisite for any attempt to verify and validate a software system.**

Before software can be built well, the purposes it is supposed to serve and how it is supposed to serve those purposes must be defined. That is done by writing requirements and specifications. A requirement is "[a] description of the behaviors that the software must fulfill. So this is the expectations of the user or acquirer, the purchaser of the software program, what needs they're seeking to have fulfilled by the system." (14T 23-17 to 21) Requirements are a first step in building a piece of software, before anything is coded. (13T 38-10 to 14) That is because it is extremely difficult to first build software and then ensure its quality. (13T 40-1 to 4) To build a system before objective requirements are laid out is "almost guaranteed failure"—there's no way to know "you're building the software correctly" if requirements that lay out what "correctly" is do not exist. (13T 40-4 to 11).

"A general principle of good requirements engineering practice is that requirements should be testable." Sommerville at 245. In other words, requirements have to be objective demands that a system can pass or fail. To return to the hypothetical bridge, a poor requirement would be "must be able to hold the weight of traffic." Such a requirement is too ambiguous. What does "traffic" mean? How much weight is that? A proper requirement would say "must be able to hold the weight of a two-axle truck weighing 20,000 pounds."

After requirements are created, software specifications must be created. Specification is "the process of understanding and defining what services are required from the system and identifying constraints on the system's operation and development." Id. at 54. While stemming from requirements, "[a] specification is a more technical description that may not be very accessible to even an informed user of the system[.]" (14T 14-5 to 8) Developing specifications is "necessary and ultimately the basis of the software testing process." (14T 14-5 to 8)

All of the requirements and specifications must be clearly and objectively written. "The software requirements document (sometimes called the software requirements specification or SRS) is an official statement of what the system developers should implement." Sommerville at 126. "Critical systems need detailed requirements because safety and security have to be analyzed in detail to find possible requirements errors." Id. at 127. "The complexity of even what might seem like trivial software means that the more objectively defined your specifications are the less ambiguous it's going to be when your claims of testing activities are completed, whether that can be universally agreed upon or not. So objectivity reigns supreme." (14T 26-24 to 24) Moreover, because software is tested by its meeting of its requirements, testing a system that does not have detailed requirements is like grading a test without an answer key—with nothing to compare the answers to, there is no objective way to determine if those answers are right or

wrong. Without the blueprint to judge the bridge against, it is impossible to know if the bridge was built correctly.

**b. Verification means testing a software system against its requirements to determine that it has been built as intended.**

Verification is the process of testing a system against it is requirements and specifications. Heimdahl 13T 36-15 to 19 (verification is checking that the software "faithfully implements what the requirements said that it was going to do"); Adams 14T 33-20 to 25 (verification is "an evaluation of those specifications and a demonstration through the provision of objective evidence that those specifications have been fulfilled"); Martin 15T 32-17 to 19 ("[V]erification is a process where you're trying to make sure that your software performs as expected based on the engineering requirements that you have."); Sommerville at 228 ("Software verification is the process of checking that the software meets its stated functional and non-functional requirements[.]") (emphasis added).

During the verification process, software is tested against its requirements. "Testing is intended to show that a program does what it is intended to do and to discover program defects before it is put into use." Sommerville at 227. This involves both demonstrating that the software meets its requirements and finding inputs "where the behavior of the software is incorrect, undesirable, or does not conform to its specification." Ibid.

It is critical that a sufficient number of tests are run. Not all possible inputs can be tested, because there are "close to infinite combinations of those variables[.]" (13T 28-1 to 3) Therefore, to adequately test a system, the testers have to demonstrate that they "sample[d] that input space and pick[ed] test cases that are exercising that software well, and that's a meticulous process that needs to be done well." (13T 28-3 to 6) Testing has to cover each individual part of the system, known as unit tests; has to test many units together, known as component tests; and must also

test the system as a whole. (13T 47-14 to 48-15; 14T 73-16 to 74-7; 15T 32-23 to 33-8) A significant amount of code should be tested; "in principle every code segment that you write should have at least one associated test." Sommerville at 243. The amount of code that is tested is referred to as "code coverage." Ibid. However, "testing a requirement does not mean just writing a single test. You normally have to write several tests to ensure that you have coverage of the requirement." Id. at 246.

A crucial purpose of testing is finding the limits of the system's functionality. When testing software, you should try to 'break' the software by using experience and guidelines to choose types of test cases that have been effective in discovering defects in other systems. Sommerville at 252.

In addition, traceability is a requirement of well-designed software. Traceability is the ability to connect the requirements, the code, and the tests. (D-13 at 27) Software must be traceable to be considered reliable. 13T 47-1 to 14 (Heimdahl: a software engineer needs to trace a test "either back to the requirements or back to the architecture or design"); 14T 67-2 to 5 (Adams: a software engineer "need[s] to be able to connect our tests through the implementation to the specification. So that's the traceability of those requirement and specifications all the way through the development process."). Traceability also helps developers keep track of how changes to the code or to the requirements impacts other parts of the code. Sommerville at 134 ("You need to keep track of the relationships between requirements, their sources, and the system design so that you can analyze the reasons for proposed changes and the impact that these changes are likely to have on other parts of the system. You need to be able to trace how a change ripples its way through the system.").

Testing is a critical and mandatory part of assuring software reliability. However, it is impossible to determine that a software is free from faults through testing. "Testing cannot demonstrate that the software is free of defects or that it will behave as specified in every circumstance." Sommerville at 227. In other words, testing is necessary to assure reliability, but even the best testing cannot guarantee a system free of error.

The process of creating requirements, specifications, running tests, and making the system traceable must generate a lot of documentation. The documentation generated is necessary to the demonstration that a piece of software is reliable, not an optional component. 14T 31- 20 to 32-4 (as a consequence of adhering to software engineering principles, "the generation of specific types of documentation" must occur, so that "it's not something held in an individual's head of what has actually been accomplished but it's something that can be ideally objectively described in black and white. This is the prescribed process. This is our plan to accomplish that process. This is the demonstration that we did."). It is especially important to generate and allow for the inspection of such documentation for safety-critical systems, "so you can demonstrate to a third party that your software is actually going to work and that it's always going to work and what its limitations might be." (13T 17-24 to 18-2) It is not acceptable in the field of software engineering to simply "trust" that a developer has built the system and tested appropriately. (13T 40-14 to 41-7)

**c. Validation means checking that the software fulfills the needs of its stakeholders.**

Validation is a more general process than verification. Sommerville at 228. The aim of validation is to ensure that the software does what the customer or acquirer intends it to do. Ibid. (See also 15T 32-12 to 18) That means that the system has to meet "stakeholder needs." (14T 33-2) See also IEEE 1012-2016 at 64-65 (explaining that requirements must be defined through conversations with "all stakeholders"); IEEE, Response, NIST Internal Report (Da- 251)

(describing users for DNA mixture interpretation as including the "public upon whom these systems are used, litigants, academics, journalists, and other researchers"). "The aims of the requirements elicitation process are to understand the work that stakeholders do and how they might use a new system to help support that work." Somerville at 112. Validation comes after these requirements have been discovered, converted, and then the system built. Id. at 111. In the case of probabilistic genotyping systems used to generate evidence provided in court, those stakeholders include criminal defendants, prosecutors, the criminal justice system, and the public. (14T 33-6 to 15)

Moreover, in order for a system to be validated, the end user needs to be able to interpret the data. (15T 60-20 to 61-3) In the case of Therac-25, discussed above, the system may have been verified but it was not validated. The choices of how to build the alerts resulted in the users—those who were dispensing radiation—being unable to understand key diagnostics. The results were catastrophic.

**d. Validation and verification for safety-critical systems must be undertaken by independent testers.**

Verification and validation of safety-critical systems requires that V&V is undertaken by people independent of the developers of the system. IEEE 1012 requires rigorous technical, management, and financial independence for an integrity level 4 system such as STRmix. IEEE 1012-2016 at 199. Level 3 systems require rigorous technical and financial independence and can have conditional managerial independence. Ibid. The verification and validation standards for medical devices and airplanes have similar independence requirements. (Dr. Heimdahl Report, D-10 at 17-18)

Independence is incredibly important for the V&V effort of a safety-critical system because of the biases and motivations, conscious and unconscious, of the developers of a system.

(14T 39-20 to 40-12) The developers of a system do not make good testers because they need to "test the software to demonstrate that it actually does" what is supposed to do instead of finding flaws in the software. (Heimdahl 13T 58-21 to 22) For commercial products, developers "often have an incentive to push their devices to market quickly." (15T 37-1 to 2) Although it is laudable for internal testers to communicate with the developers about issues they find, that communication is "why you also need independent validation and verification," because those communications ultimately undermine any independence testers may have. (15T 38-7 to 12) It is "common in the industry" of software development for developers "to write their own unit tests to test the components they design," but "[t]hat's why you need an outside person because when you're so close to it you might actually be myopic or blinded by how close to the product development you are. Someone outside the system needs to test it." (15T 38-15 to 23) In other words, it is important and praiseworthy for a system to be tested as its developed by the people who are developing it, but that is not a substitute for IV&V.

In sum, independence is required for all aspects of the IV&V process for safety critical software. Such independence "requires the exclusion of parties with a stake in the outcome" in all stages of IV&V. IEEE Response at 4 (Da 251) In the case of forensic software, interested parties "include forensic labs that, while not financially dependent on developers, have a shared interest in the software's acceptance." Ibid.

**5. Verification and validation is a commonplace process that is expected to occur for all software.**

Verification and validation is a commonplace process that should and usually does occur for all software. (13T 38-5; 14T 180-9 to 13; 15T 31-19 to 32-8) Independent verification and validation is a commonplace process that should and usually does occur for all safety-critical software. (13T 38-6 to 4107; 14T 180-9 to 13; 15T 31-15 to 33-8; 39- 2 to 9, 100-11 to 13)

STRmix could have IV&V performed by an outside tester if it paid for it; finding an entity to perform IV&V is "not difficult." (15T 40-16 to 22) Harbor Experts is an example of an organization that would be willing to contract with STRmix to conduct an appropriate IV&V effort. (15T 39-16 to 40-20)

**6. The publication of peer-reviewed articles is not an accepted means of determining the reliability of software.**

It is well-established in the software engineering community that peer review and scientific articles are not an accepted means of determining the reliability of software. Dr. Heimdahl Report of June 2024 at 15 (D-10); Dr. Martin at 15T 33-14 to 35-13; IEEE Response to NIST at 5 (Da 251) ("[P]eer-reviewed publications, while a priceless tool for scientific inquiry, are not a substitute, nor a valid approximation of IV&V when determining reliability or trustworthiness of a deployed system."). In other words, such review is "not a substitute for a formal IV&V process." (15T 33-14-15)

The publication of a peer-reviewed article and IV&V serve different processes. The articles assess if the mathematical model is yielding correct outcomes under certain circumstances, but they do not assess if the code is free of bugs and the system is operating as required by the stakeholders. (15T 33-16 to 35-1) For reasons discussed further in subsection E.3, the peer-reviewed studies that examine STRmix's performance are a particularly inapt substitute for IV&V. In short, the software engineering community is unanimous that peer-reviewde publications are not a substitute of IV&V and do not demonstrate the reliability of software.

**E. STRmix v2.5.11 and v.2.8.0 have not been independently verified and validated.**

Only three software engineering experts testified in this case. They all agreed that neither STRmix v2.5.11 nor v2.8.0 has been independently verified and validated. Therefore, their

reliability has not been established. Dr. Heimdahl concluded that the documentation he was provided does not "even come close" to "establishing that the STRmix software has been sufficiently validated and verified." (13T 43-25 to 44-5) Dr. Heimdahl concluded that "[t]he documentation provided in support of STRmix™'s quality is completely unacceptable for a forensic tool that is being used to potentially deprive a defendant of their freedom or, potentially, their life. The documentation gives no indication that a suitable software development process has been followed, nor does it inform us about the validation and, in particular, verification efforts undertaken to ensure that STRmix™ operates correctly." (D-10 at 34-35) See also 13T 51-6 to 23. Dr. Heimdahl concluded that STRmix does not meet IEEE 1012 or any other standard governing safety-critical systems. (13T 69-2 to 12)

Similarly, Mr. Adams concluded that "the evidence of the V&V activities claimed to have occurred by the STRmix team and that I've reviewed, do not comply with the expectations that the software engineering community has for a software program with these identified consequences, potential consequences." (14T 86-2 to 7) See also D-17 at 16 ("STRmix v2.5.11 development materials provided for review in this case exhibit deficiencies and deviations from software engineering norms in terms of both construction and documentation. STRmix has not been demonstrated to meet compliance at any integrity level of IEEE Std 1012[.]").

Dr. Martin likewise concluded that STRmix v2.8.0 does not comply with IEEE 1012. (15T 67-23 to 24).

The bases for these conclusions are explored at length below. The unanimous opinion of three very knowledgeable experts in the field of software engineering leads to the inescapable conclusion that they are correct: neither STRmix v2.5.11 nor v2.8.0 have met the standards of independent verification and validation required for safety-critical systems.

**1. STRmix v2.5.11 and STRmix v.2.8.0 have not been verified.**

The first failure to demonstrate that STRmix has been verified (or validated) comes from the failure to create appropriate requirements and specifications. STRmix was created without a requirements document. (7T 12-2 to 4, 195-1 to 5) In fact, the first time Adams reviewed v2.5.11 he was not given a requirements document. (14T 166-1 to 14) This requirements document was created between Mr. Adams's first review and his review in this case, specifically due to his criticism following his first review. (7T 12-2 to 4) But as discussed above, requirements must be written before anything is built. (13T 38-10 to 14) It is "very, very difficult" to build first and then "test quality into a product." (13T 40-3 to 4) That is why developers are supposed to build quality in "from the beginning" by adhering to software engineering standards. (13T 17-10 to 17)

Additionally, considering these belatedly created requirements documents on their merits, the requirements documents are inadequate. Dr. Heimdahl, a preeminent expert in safety-critical systems, reviewed both requirements documents. He concluded that they did not actually describe the requirements for STRmix and should not be considered requirements documents. (13T 44-9 to 45-1) Dr. Martin agreed that the requirements document was "light in key areas." (15T 44-23) Mr. Adams agreed that the requirements document did not sufficiently lay out the system requirements specifications and there "seems to be a misunderstanding of what" a requirements document is "supposed to be." (14T 77-10 to 16)

Perhaps one reason for that misunderstanding is that according to both Dr. Buckleton and Mr. Adams, the requirements document conflated specifications—what a program is supposed to do—with a separate component of software development, design, which is how something will be accomplished by the developers. (14T 79-5 to 8; 7T 13-1 to 10) See Adams Report, July 2023 (D-17) at 4-5; Buckleton Report, Aug. 2023 (S-152) at 19).

Without a real requirements document, you can't test the software. "You need to know what the system is supposed to be doing before you can really test it." (13T 46-8 to 10) Therefore, the failure to create a sufficient requirements document is fatal to any claim that STRmix has been verified.

Additionally, the testing that was done is lacking in scope and in traceability. Dr. Heimdahl explained that "there was no real documentation of any actual tests that had been run, nor where those tests might have come from." (13T 47-1 to 3) This lack of documentation is not acceptable in a safety-critical system. (13T 47-4 to 6)

Mr. Adams agreed that there was insufficient evidence of necessary testing. He explained that "[t]here were certainly documents that pertain to testing activities. They largely lacked traceability to specifications. There's also a concerning lack of evaluations of test adequacy or sufficiency. So how many tests of which components of the system do we need to execute for there to be a sufficient set of tests." (14T 69-11 to 13) Particularly concerning to Mr. Adams was that the tests did not appear to attempt to challenge STRmix in order to find faults and limitations: "the tests that were documented that we did provide or did receive information about seem to largely be evaluating the intended functionality of the system, rather than trying to find gaps or holes in the system if it was challenged with potentially breaking sets of data." (14T 70-6 to 11)

Both Mr. Adams and Dr. Martin found no basis for the claim that 85% of the code has been tested, found it implausible, and explained that it would be quite easy to provide evidence of how much of the code had been tested. (14T 76-9 to 12; 15T 63-16 to 64-3)

Overall, Mr. Adams concluded that the testing that was done was not acceptable for a safety-critical system and that STRmix did not "even come close to" demonstrating sufficient

testing of a safety-critical system: "These are fundamental software engineering concepts that they're failing to adhere to." (14T 69-21 to 22; 14T 74-13 to 17) "[N]ot having that component level or integration level testing is concerning, because it appears that there's been a focus on small units and then the whole system, but not the building up of the hierarchy through the system architecture." (14T 74-13 to 17)

Dr. Martin also concluded that both versions of STRmix were inadequately tested. He explained that v2.5.11 did not have unit tests and v2.8.0 had some but not "an adequate amount of them to cover most of the functions of the code." (15T 42-3 to 5) The tests that did exist were not traceable because he "did not see testing for code that implemented requirements. If software is not traceable to its requirements, you don't have tests to make sure that you implemented each requirement properly." (15T 63-13 to 15) Dr. Martin explained that there were insufficient tests and documentation of those tests for a safety critical system. (15T 65-6 to 2)

Dr. Buckleton agreed that traceability is a valid concern about the software. He also agreed that STRmix was "probably deficient on putting sufficient documentation in front of people like Mr. Adams," by which he presumably meant software engineers. (7T 207-21 to 23) "I think our software documentation needs a step up," he opined. (7T 207-23 to 24)

**2. STRmix v2.5.11 and STRmix v.2.8.0 have not been validated.**

Without the necessary requirements, it is impossible to validate STRmix. The lack of sufficient requirements documentation ends any argument that STRmix has been or can be validated. Two specific concerns arise even in that vacuum of information.

First, there is no testing that demonstrates that the actual person operating the software in the laboratory will properly deploy it. If an ordinary operator cannot understand the software and use it properly, then the software is not fit for purpose. (15T 60-20 to 61-3) The lack of any documentation that STRmix can be used appropriately by its users—analysts, prosecutors,

judges, juries—is fatal to a claim of validation. Further, as discussed further in subsection F.10.c, there were red flags that suggested that one important group of users, DNA analysts, do not have the necessary tools to understand and use STRmix. Dr. Buckleton agreed with the criticism in Mr. Caneiro's prehearing brief that STRmix "need[s] to do more to help guide the user to use it properly." (7T 209-4 to 8) From a lack of understanding about the diagnostics to inappropriate visual exclusions before selecting inputs into the problem, there is deep cause for concern about validation.

Second, STRmix was designed without all of the stakeholders impacted by STRmix being involved in the process of determining what the requirements should be. Before requirements are converted into specifications, the requirements should be elicited through conversation with all of the stakeholders. Somerville at 111. See also IEEE 1012-2016 at 64-65 (explaining that requirements must be defined through conversations with "all stakeholders"); IEEE 1012-2012 at 39, 41, 21 (same); IEEE, Response, NIST Internal Report (describing users for DNA mixture interpretation as including the "public upon whom these systems are used, litigants, academics, journalists, and other researchers"). "The aims of the requirements elicitation process are to understand the work that stakeholders do and how they might use a new system to help support that work." Somerville at 112. Validation comes after these requirements have been discovered, converted, and then the system built. Id. at 111.

ESR, which sells STRmix, developed the software perhaps with only one user in mind: laboratories. They did not consider the views of judges, laypeople, or criminal defense attorneys. Yet the inclusion of all stakeholders is especially essential due to the incentive structure in the criminal justice system, which is very different than in other fields. Every single person involved in the development of autopilot software—the pilots, the software engineers, the passengers, the

profit-seeking companies—want planes to stay in the sky and not fall out of the sky. In contrast, here, there is a separation of who buys the software and who bears the cost of its failures. As Dr. Heimdahl explains, "[t]he customers are prosecutors and law enforcement, but those harmed by software flaws are criminal defendants." (D-10 at 25) This creates a lack of incentive to discover, and report, flaws. This lack of incentive could have been lessened if all interested stakeholders participated in setting the requirements for STRmix. Then, validation could take into account the needs of all stakeholders by having their representatives involved in the process of developing requirements. But it did not. This undermines any ability to appropriately validate STRmix.

The views of all relevant stakeholders—judges, lay people, criminal defense attorneys, and others—are necessary to make sure that STRmix is the right product that makes the right tradeoffs between accuracy, efficiency, and hardware needs. These different stakeholders would have different preferences and priorities regarding many of the decisions made in the design and implementation of STRmix. These decisions are issues of validation. (14T 59-1 to 13) For instance, when modeling dropout, STRmix does not consider all of the possible combinations of alleles that could have dropped out. Instead, it uses a catch-all "Q" allele. This catch-all allows the model to work more quickly, but it introduces a new problem: because Q is not a real allele, it does not have a real molecular weight. Without a molecular weight, STRmix cannot model degradation. To fix this, STRmix assigns the hypothetical Q allele a hypothetical molecular weight. This lowers the overall runtime for the system, at the expense of accuracy. User Manual, STRmix v2.5.11 at 115 (D-301); 7T 178-15 to 17, 181-6 to 12; 186-3 to 25. This is the kind of trade-off that must be articulated and tested, involving the input of all stakeholders in order to determine whether it is the right trade-off given the use of the system. See also IEEE 1012-2016 at 70-71 (explaining that tradeoff considerations must be articulated and verified).

The designers have also decided to decrease reproducibility in order to decrease runtime. User Manual, STRmix v.2.5.11 at 149 ("Using a stochastic system like MCMC, the results of no two analyses will be completely the same. This is an issue that is relatively new to forensic science, which up until this point has always had the luxury of (at least theoretically) completely reproducible results. Increasing the number of accepts or the random walk standard deviation (step size) ameliorates but does not remove the variation. There is, however, an associated runtime cost. Hence a trade-off between reproducibility and runtime must be struck.") (emphasis added).

The developers have also chosen to recommend that any runs that produce a Gelman-Rubin score of 1.2 or below should be considered acceptable in order to save run time, even though the original recommendation by the developers of that score was actually a 1.1. (7T 196-7 to25)

It is not that we know that any of these tradeoffs are right or wrong. Rather, in order for software to be validated, all of the tradeoffs must discussed by all of the stakeholders so the right product can be made for the purposes it will serve. For instance, many of the modeling compromises made for the sake of runtime are only necessary because STRmix is designed to run on a personal laptop computer. Whether this self-imposed limitation is necessary or appropriate should include input from all stakeholders, not just ESR and its customers.

**3. Testing STRmix with ground-truth samples is not a substitute for verification or validation.**

Many studies have been conducted by running samples whose composition is known—samples where "ground truth" is known—through STRmix and checking the results against what is known about the samples. The software engineering experts were unanimous that these

studies, including "validation studies," are not adequate substitutes for verification or for validation as those terms are meant by the field of software engineering. (15T 85-9 to 20)

These studies cannot ever be enough to see if the code is operating currently because they do not test the code. At most, these studies are testing the DNA science behind the software, but that is not the same thing as testing the software itself. The science might be sound conceptually, and might work in many instances, but the software might fail in many circumstances due to any number of issues with the software itself, which can range from specification errors to errors that result from unanticipated interactions between components to component interface problems and beyond. Sommerville at 59-63, 270, 356. Thus, as Dr. Heimdahl explained, these ground-truth studies are "testing the science behind STRmix," but that is "a completely different activity than developing a highly trusted production pieces of software." (13T 63-14 to 17) Thus, as Dr. Heimdahl explained, these studies "say absolutely nothing about the quality of the software or the development techniques that's been used or how the software itself has been meticulously tested." (13T 68-17 to 20, 138-5 to 12) In short, testing the math behind STRmix is not the same as testing the code. There are five critical reasons why testing outputs of STRmix over whatever range of samples are in the studies is not a sufficient substitute for adequate code testing.

First, the ground-truth studies are also unlikely to cover most of the software code. As Mr. Adams explained, the total amount of testing of STRmix covered by those studies is minute, as STRmix performs "millions, billion, [or] trillions" of calculations "in a single deconvolution." (14T 154-23 to 155-1) "The total amount of human testing of STRmix entirely might amount to only a fraction of a single deconvolution performed by the system." (14T 155-2 to 4) Critical to software testing, and missing from all of these studies, which do not examine the code, is any "publication of why those specific tests were selected, what those specific tests were, what parts

of the code those tests exercised, and so on and so forth." (13T 138-5 to 12) So although these studies provide some information about whether STRmix's outputs comport with the expectations of forensic DNA scientists in some circumstances, they do not provide information about whether the software is operating correctly and do not test a sufficient amount of the code to enable anyone to assume the software will operate correctly in all circumstances. Although ground-truth studies test a range of DNA samples, that is not the same as testing enough of the code and testing how the code operates together under enough circumstances.

Second, the non-continuous nature of the software means that errors occur in unexpected ways that may not be captured in the ground-truth studies. Heimdahl Report, June 2024 (D-10) at 10-11; Pickett, 466 N.J. Super. at 299 (recognizing that the "software is non-continuous, meaning that correct results for the samples used in the validation studies do not preclude the possibility of erroneous results for others that do not match those samples"). As Dr. Heimdahl explains, a continuous system, in contrast, will operate predictably with different inputs. A crane that has been designed to lift boats and has been demonstrated to lift 100 pounds without failing can lift 10 pounds without failing. (D-10 at 10) But "[t]he same cannot be said for the software used to process the handling of the boats. If the software has not been programmed to handle certain cases (such as weekend transactions, red boats, or long owner names), it might fail entirely or produce erroneous results that could go undetected." (D-10 at 10) "[A]ny input could cause failure" in a non-continuous system. (D-10 at 11) That is why, as Dr. Heimdahl explains, these ground-truth sample studies are insufficient to demonstrate the risk of error with the use of STRmix: "they reveal exceedingly little about the quality of the software implementation of that computational solution and cannot rule out errors that could impact the result for any given set of inputs." (D-10 at 11) See also 15T 46-4 to 11 (Dr. Martin explaining that "[t]he problem is with

the samples that you don't have ground truth for, the whole point of bugs is that they're hard to detect. If they were easy to detect, they would have been caught, fixed and they wouldn't be there. So most bugs that I've looked at like they'll happen under very specific conditions, or a small percentage of the time based on things that are happening in the program."). Given the lack of independent verification and validation, and the documentation of such testing, there is no way to know whether the DNA mixture in any given case is one of those times or conditions where the software will fail.

Third, because there is no objectively correct LR for any calculation, it is harder to discern when the system does not work. Results on ground-truth samples would not demonstrate plausible but incorrect LRs. Let us imagine a sample where we know Person A is a contributor. It is impossible to say that an LR of 10,000 or 50,000 or 100,000 is objectively correct or not. These are all inclusive LRs for a person who should be included. The only measure for the LR's correctness in such a circumstance is whether it is the LR that STRmix's designers intended for STRmix to produce under those circumstances. That means that "plausible but erroneous result[s]" are extremely concerning to the software engineers, because such results would be very hard to find. (13T 53-19 to 24; 15T 93-21 to 23) Because there is no way to determine if an LR is correct for any given analysis, the only way to develop confidence in STRmix's reliability is through "rigorous adherence to software engineering" standards. (14T 178-1 to 25) See also 13T 53-21 to 24 (plausible but erroneous results are "insidious" and a developer of software "really need[s] this meticulous development process to make sure that those things don't happen"). These ground-truth studies generally consider STRmix to get to the right answer when it provides an inclusionary LR for a contributor and a non-inclusionary LR for a non-contributor. But STRmix does not simply produce a binary result, inclusion or exclusion; it produces a

numerical LR. Giving a binary check on reliability is insufficient because it does not actually assess whether the software is producing the LR that the developers believe it should assign under the circumstances of a specific analysis.

Fourth, and relatedly, STRmix's outputs cannot be manually verified by a human. As Dr. Buckleton explained, there are simply too many calculations run for any given sample for a human to check STRmix's work. (7T 159-12 to 21) A human "can't feasibly actually calculate these things by hand to make sure that you check your work, to make sure you get the right result on any particular sample." (15T 46-25 to 19) Therefore, none of the ground-truth studies allow the peer reviewers or publishers to check that STRmix got the right answer, other than for the binary of inclusion or exclusion. But because STRmix produces a number, the inability to check if that number is the one STRmix should have reached in a given case makes the analysis of these ground-truth studies insufficient.

Fifth, these studies, as discussed further below in subsection D.4, are done by STRmix's developers and by laboratories that want to begin using STRmix, and a few by independent researchers. As IEEE explains, "most peer-reviewed studies of probabilistic genotyping software are not independent, violating a fundamental tenet of IV&V, and making them insufficient to determine reliability." IEEE, Response to NIST, at 5. These studies cannot be considered independent for purposes of IV&V.

In sum, ground-truth studies, even those that are peer-reviewed and published, are not adequate evidence of the reliability of STRmix as a software system.

**4. Any verification or validation of STRmix that has occurred has not been undertaken by sufficiently independent testers.**

There has been no evidence produced that demonstrates the sufficient independence of STRmix's testers. Dr. Buckleton claims ESR is financially, technically, and managerially

independent of Forensic Science South Australia (FSSA) and Orbit Systems—the two other entities involved in the development and testing for STRmix—but that claim cannot be meaningfully investigated or substantiated. Mr. Caneiro wrote a letter requesting more information about the relationship between ESR, FSSA, and Orbit. ESR refused to provide that information. (7T 124-1 to 2) Because the State bears the burden of demonstrating reliability as the proponent of the software, the failure to provide any information weighs against the State, not the defense.

Moreover, the evidence produced shows that all of these entities are intertwined to some degree. ESR is both a developer and a tester of STRmix. (14T 44-20 to 22; Martin Report, July 19, 2024 (D-13) at 34) The STRmix unit within ESR is involved in any validation and verification activities. The STRmix unit in ESR equally has a close technical relationship with both FSSA and Orbit, the other chief developers of STRmix. (14T 44-24 to 45-6) Dr. Buckleton explained that there is significant collaboration between ESR, FSSA, and Orbit. (7T 127-2 to 7) In short, the information provided demonstrates that STRmix's independence is "[a]mong the lower tiers, possibly the lowest tier, which is called embedded independent verification and validation." (14T 46-13 to 15)

The claim that the developer and testers of STRmix are somehow immune from bias because they are government employees working for a governmental organization, and that therefore less independence of testers is required, must be rejected. First of all, as explained above, the need for independence is due to unconscious bias as much as conscious bias. People who work for an organization have a vested interest in the professional success or reputation of its products. People who work on a product also want to believe that product is a good one. Given the lack of independence, there is a real risk that STRmix testers are subject to an

"unconscious bias" that could undermine the testing efforts. (13T 75-2 to 17) That bias does not evaporate merely because one works for the government.

Second of all, ESR is a company with a profit motive, not just a commitment to science. As Dr. Buckleton explained, ESR has a board of directors, a CEO, and is "expected to operate in a business-like manner." (7T 121-6 to 13) "Financial performance" is one of the two key deliverables for ESR. (7T 138-7 to 8) In ESR's 2014 Four Year Annual Review, the writers note with concern that ESR's "financial performance has been deteriorating" and "key financial ratios remain below the shareholders' expectations." (D-500 at 5) The writers highlight the "promising start of commercialization of ESR's STRmix technology" to help meet ESR's revenue targets. (D-500 at 21) One attempt to brand ESR in order to raise money was a new tagline "The Science behind the truth." (D-500 at 16) The "science staff" were uncomfortable with "the proposed brand positioning and tagline." (D-500 at 29) The writers noted that senior leadership of ESR "urgently need to work with the staff to enable them to align the purpose of 'good science' with ESR's wider organization objectives" of raising revenue. (D-500 at 29)

To be clear, software made by companies for money can be reliable. The way for such software to be demonstrated to be reliable is by the provision of documentation that demonstrates the standards of IV&V have been met. That is the same way that software that is not made for profit can demonstrate its reliability. The standards are the same. But insofar as there has been an insinuation that the fact that Dr. Buckleton is a public servant or that the testers are public servants or that ESR is government entity means that the biasing influence of money is absent from the development of STRmix, that is simply incorrect.

**5. Coming close to complying with verification and validation standards is not sufficient to assure reliability.**

As explained above, demonstrated conformance to verification and validation standards for safety-critical systems is the only way to demonstrate the reliability of that system. Dr. Buckleton asserts that STRmix conforms with "modified or classical" IV&V. Buckleton Report, Oct. 2024, at 9 (D-20). Only classical IV&V is appropriate for level 4 integrity systems. (14T 43-9 to 15) The State pointed out many times that of all PGS, STRmix comes closest to complying with IEEE 1012. (14T 107-15 to 108-10) Although it is admirable that STRmix is trying to comply with IEEE 1012, STRmix does not meet the requirements for classical IV&V and therefore it cannot demonstrate its reliability. As Mr. Adams explained, "[s]tandards compliance is not, you don't get partial credit. You comply or you don't. And if there are deficiencies in the adherence then we need to identify what those risks are to make an assessment of how effective was the software development process, what level of confidence should we have in a system." (14T 159-3 to 14) STRmix does not comply with any standard of independent validation and verification and therefore we cannot have sufficient confidence in its reliability.

**6. Source code review in an adversarial setting is not a substitute for IV&V and does not guarantee the software is free of flaws.**

Source code review in an adversarial setting is not a substitute for IV&V and does not guarantee that software is free of flaws. The review in this case, of source code and of related documentation and development materials, has revealed that there is an insufficient basis to claim that STRmix is a reliable piece of software. The source code review was a necessary component of assessing the reliability of STRmix and provided important information for the court, but is not more important than the evaluation of whether STRmix has demonstrated that it meets software engineering standards for safety-critical software. For the reasons that follow, any claim that the source code review somehow demonstrates STRmix's reliability is incorrect.

**a. The source code review in this case was limited.**

First, the lack of any bugs found during the review of the code misses the point of the review of the related software documentation. As explained above, there are many other concerns about the reliability of software not encompassed in the binary finding or not finding of bugs in the code itself.

Moreover, the source code review allowed in an adversarial setting, including in this case, did not allow for the fullest testing of the code, so bugs that do exist would not necessarily be found in this context. The limits of such reviews is why Dr. Heimdahl explained that such a review was unlikely to reveal all the flaws in a system: "given the restrictions in the code review and the enormous efforts involved in doing a thorough review, having access to the code is an absolute necessity, but it's certainly no guarantee that the reviewer is going to be able to pass a judgment and say, yep, this code is good." (13T 108-18 to 109-1)

The constraints of this source core review in this case further limited what theoretically could be done in a general adversarial source code review. Neither Mr. Adams nor Dr. Martin was able to debug the code, "step through the code," or perform a dynamic inspection of the code, tasks that would have been very helpful in assessing its reliability. (14T 59-3 to 19; 15T 98-12 to 99-12) In fact, Mr. Adams, who reviewed the code and related documents for v2.5.11, declined to review v2.8.0. He explained that in asserting that STRmix is reliable, it is often cited that "the code has been reviewed four or five times," mostly if not completely by Mr. Adams. (14T 61-18 to 20) But Mr. Adams testified that the limits imposed on him in those reviews meant that they "cannot be effective, cannot be accomplished" meaningfully. (14T 61-21 to 22) On top of that, the protective orders he has signed and the threat of lawsuits by ESR mean that he "can't discuss them after the fact," which means he cannot prevent "misinformation going out in the world" about the reviews and their significance. (14T 61-23) Mr. Adams declined to review the

source code of v2.8.0, a task he would have been paid for, because he "feel[s] complicit with if I continue to engage in very limited, partial kinds of reviews of the system." (14T 61-23 to 25) In other words, these reviews are not enough to ensure STRmix's reliability and the claim that they are enough is so wrong that Mr. Adams can no longer be involved. Although there seemed to be some focus by the State at the hearing as to who was to blame for the limited source code review, the allocation of fault is not relevant to the factual conclusion that the source code review was in fact limited. Notably, Dr. Martin testified that he faced the same limitations Mr. Adams faced. The failure to find a bug is never proof that a bug doesn't exist; it is axiomatic that "[t]esting can only show the presence of errors, not their absence." Sommerville at 227 (internal quotation marks omitted). But it is even more unlikely to find an error under the circumstances of the reviews in this case.

In fact, all three software engineering experts agreed that the errors in STRmix are not limited to the "miscodes" published by ESR. Dr. Heimdahl and Mr. Adams agree that it is likely that STRmix contains undetected faults. (13T 146-1 to 11; 14T 172-13 to 20) Dr. Heimdahl explained that there is "no evidence of the meticulousness, the rigor necessary to make an argument that" errors are "just not going to happen or it's going to be sufficiently unlikely." (13T 146-1 to 11) Both Mr. Adams and Dr. Martin explained that based on Dr. Buckleton's own report, there have been more errors detected in STRmix than the "miscodes" listed on STRmix's website. (14T 172-13 to 20; 15T 52-21 to 53-15) Because "miscode" is not a term used in software engineering, it is unclear if they are somehow classified by Dr. Buckleton to be narrower than software faults. (13T 109-13 to 25, 144-9 to 1445-4; 14T 82-1 to 3; 15T 53-22 to 54-1) Any claim that there have been only 13 problems found in STRmix is shown to by false by Dr. Buckleton himself. Of concern to Dr. Martin is Dr. Buckleton's claim that all 13 miscodes

have been caught by users, a way to suggest that real IV&V is unnecessary. Errors should be caught by software testing, not by users. (15T 91-23 to 92-2) STRmix does anticipate issues to be continued to be caught by users—the risk mitigation document asserts that many risks associated with STRmix are mitigated by "training and post sales support." (7T 172-5 to 7 This is particularly concerning because, as explained above in subsection E.3, many errors are unlikely to detected by users because they arise in the form of plausible but incorrect LRs.

Dr. Martin readily testified that the overall quality of the code he reviewed was "quite good." (15T 66-2 to 10) But he explained that writing good code does not mean there are no errors: "The testing is really important even in code that's well written. Nobody intends to write bugs into their code. Good software developers write bugs all the time. You need to test the code because you want to find bugs that aren't obvious." (15T 66-24 to 67-18)

Of critical importance is that there is no evidence that, whatever these undetected faults may be, they do not impact Mr. Caneiro's case. The inability to step through the code or run it dynamically means that neither expert watched STRmix work as these samples were run.

**b. A source code review by a criminal defendant would never be able to replicate independent validation and verification.**

Further, a code review by a criminal defendant would never be able to fully replicate an IV&V effort. Verification and validation take up a huge amount of the budget for building a piece of software: up to 50%. (13T 28-10 to 11; 14T 50-23 to 25; Sommerville at 20) The Office of the Public Defender is unable to spend such a tremendous amount of money on a case. It is very important and necessary that two source code reviews occurred in this case. But it is not a substitute for IV&V. As discussed above, all three software engineers who reviewed STRmix's software development materials found these materials insufficient to demonstrate the reliability

of the software. While short code reviews by two of those engineers did not reveal any glaring errors, IV&V is not met by a source code review in an adversarial setting.

Because IV&V is the only way to determine that safety-critical software is reliable, a code review that does not find errors is not sufficient to demonstrate that safety-critical software is reliable. In other words, this code review was necessary but not sufficient for an assessment of STRmix's reliability.

**7. STRmix's reliability has been insufficiently demonstrated under software engineering standards.**

The unanimous opinion of all of the software engineering experts that testified is that STRmix v2.5.11 and v2.8.0 have not been verified or validated, independently or otherwise, and that therefore the reliability of STRmix v2.5.11 and v2.8.0 has not been demonstrated.

**F. The ability of any probabilistic genotyping system to reliably analyze complex DNA samples must be empirically demonstrated, not assumed.**

This case presents the first admissibility hearing about the reliability of STRmix, a novel type of DNA analysis software known as probabilistic genotyping software (PGS), in New Jersey. It is important to understand exactly how different probabilistic genotyping software, including STRmix, is from traditional DNA analysis. Because of the reliance on computer modeling and the challenging nature of the samples they attempt to interpret, "probabilistic genotyping software[] marks a profound shift in DNA forensics." Pickett, 466 N.J. Super. 270 at 276. The reliability of traditional DNA analysis does not govern this case. But the principles underlying forensic DNA analysis apply.

**1. All scientific methods have limits. The reliable use of any method requires finding those limits and adhering to them.**

A general scientific principle that must guide this case is that "[a]ll scientific methods have limits." Mixture Interpretation at 11. NIST, the non-regulatory agency "with a mission to

advance national measurement science, standards, and technology[,]" sets forth this critical principle in its recent Foundation Review of mixture analysis. Id. at 7. As testified at the hearing, both NIST in general and one of the lead authors of this review in particular, John Butler, are considered authoritative in the field of forensic DNA. (1T 88-11 to 12; 2T 971-3; 10T 12-22 to 25; 16T 77-2 to 80-7) As NIST explained in this review, "[t]o use a method appropriately, one must understand those limits, which are inevitably tied to the risk one is willing to accept either as an individual or as a society. This is especially important in forensic science, as critical decisions impacting life and liberty are often based on the results of forensic analysis." Id. at 11. Moreover, "[r]eliability is not a yes or no question, but a matter of degree. Understanding the degree of reliability of a method can help the user of that information decide whether they should trust the results of that method in any specific situation when making important decisions." Id. at 15.

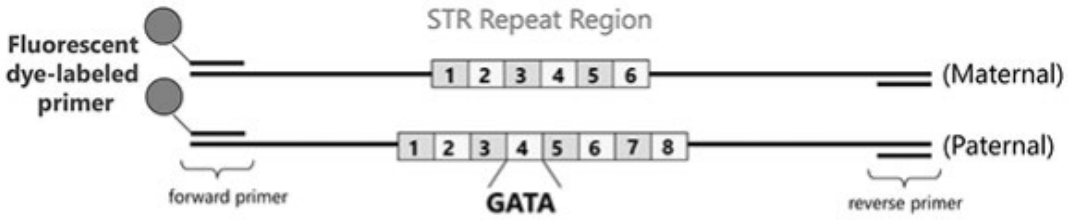
In case of DNA mixture analysis, without such limits, analysts do not know "when to stop attempting to interpret a mixture" and therefore cannot guarantee that they are using the method being used in that interpretation reliably. Id. at 28. Because all systems have limits to their reliable use, it is inappropriate to assert that any technique, including STRmix, is "reliable," without any parameters that define its reliable use, whatever that might be. For PGS, those limits must be established in relationship to the features that make different DNA samples more complicated to reliably interpret. Subsections F.3, F.4, F.5 infra. The way those limits are established is through the gathering of empirical evidence in the form of validation testing. Subsection F.8, infra.

**2. Basics of Forensic DNA Analysis**

Deoxyribonucleic acid, or DNA, is our genetic blueprint. John M. Butler, Fundamentals of DNA Typing 19 (2010). The entirety of the DNA in a cell, which is the "complete set of

instructions for making an organism," in this case a human, is referred to as the genome. Ibid. DNA is stored in chromosomes —there are 23 pairs in the human body. Id. at 24. Each person inherits one chromosomes of each pair from each parent. Id. at 6. A copy of both sets of chromosomes is contained in nuclei of cells that have nuclei. Id. at 23.

There are four nucleobases that comprise DNA, each represented by a letter: adenine (A), thymine (T), cytosine (C), and guanine (G). Id. at 20. DNA regions with repeat units of nucleobases that are 2 to 7 base pairs in length are called short tandem repeats (STRs). Id. at 148. Traditional DNA analysis examines the number of repeats on specific spot on a chromosome, called a locus. Ibid. The alternative possibilities for how many repeats there are at a genetic locus is called alleles. Ibid. For instance, the below chart shows a person who inherited a sequence of nucleobases, GATA, six times from their mother and eight times from their father:



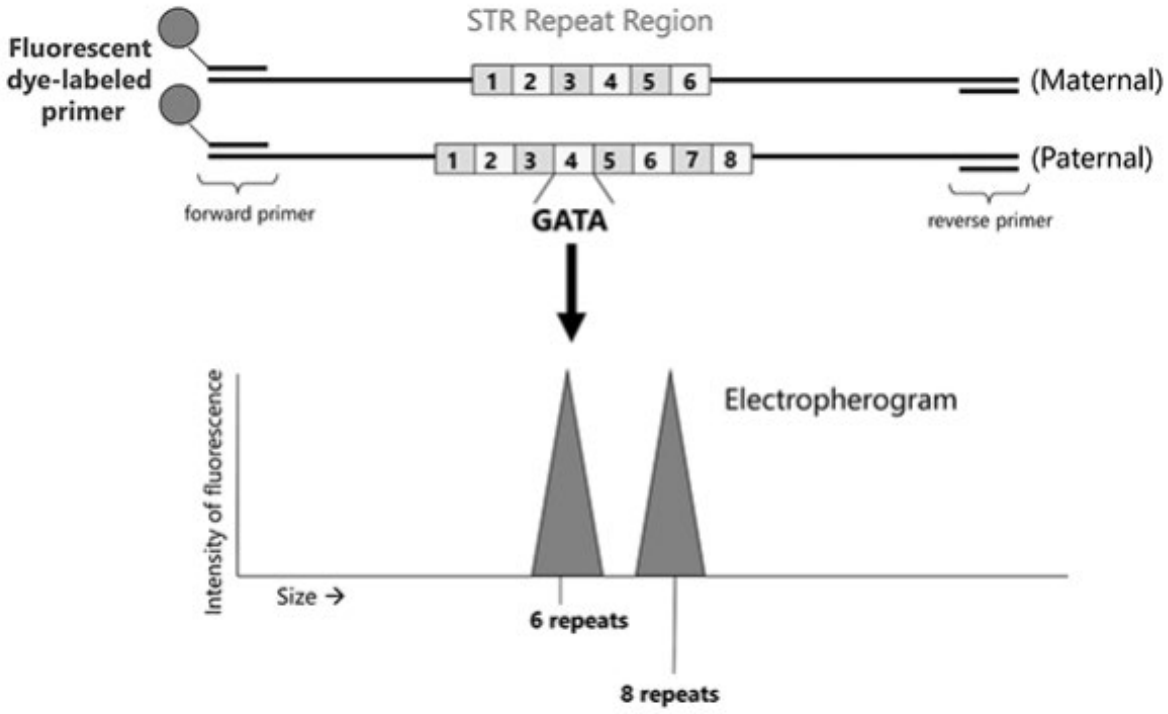

(D-604 (from Butler and Bell, Understanding Forensic DNA))

In the above example, at that locus, which would have a name such as "TH01" or "D2S1338," that person would be said to have a 6, 8 allele. A combination of alleles at each locus is a known as a genotype. A person's genotype—their genetic makeup—is reported for forensic purposes by the number of STRs at certain loci for each chromosome.

The basic steps of forensic DNA analysis are extraction, quantitation, amplification through the PCR process, detection through the generation of an electropherogram, and interpretation. (1T 47-15 to 49-4) Probabilistic genotyping and traditional DNA are no different in the extraction, quantitation, amplification, and detection steps. (1T 49-5 to 8) The difference is

in interpretation. (8T 41-7 to 12) The goal is of forensic DNA is to ascertain the genotype—the combination of alleles—that belong to each person whose DNA is in a sample.

The electropherogram is a visual representation of all of the alleles detected at a specific locus. The 6, 8 allele above would be represented as follows on an electropherogram:




(D-604) (From Butler and Bell, Understanding Forensic DNA)

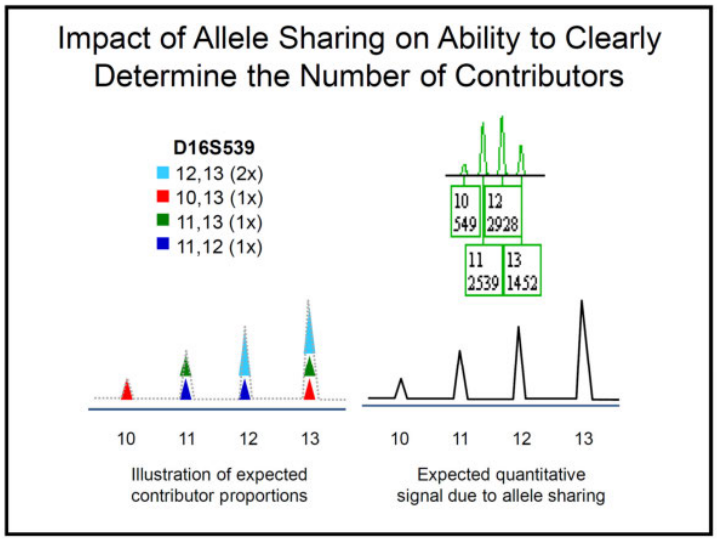
Each person has their own unique genotype, but because forensic DNA analysis only tests certain loci and not the whole genome, it doesn't always uniquely identify a person. Forensic DNA analysis involves comparing a known person's (or people's) genotype(s) to a sample from a relevant scene. Id. at 221. For a single-source sample, the analyst must determine if there are any mismatched alleles at a particular locus—if so, they could not have come from the same person. So, for example, if Sample A has a 2, 4 at a specific locus and Sample B has a 3, 5 from that same locus, the DNA did not come from the same person. If the genotype is the same, then it

must be determined how common that genotype is and how common the combination of all the genotypes in a sample are in order to determine how likely it is that a specific person is the source of the DNA.

**3. Samples that contain more than one person's DNA are more challenging to interpret.**

When there are multiple contributors to a DNA mixture, the challenge is to figure out the profile, or genotype, of each separate contributor. "In single-source samples, only a single genotype is possible at each locus"; however, [i]n a DNA mixture, it may not be clear which genetic components, called alleles, belong to which contributor." Mixture Interpretation at 12, 34.

To interpret a mixture, analysts count allele peaks, identify the number of potential contributors, and estimate the relative ratio of the individuals contributing to a mixture. Fundamentals of Forensic DNA Typing at 325. Because the mixture ratio is approximately preserved in PCR application across loci, a contributor would be expected to give a similar amount of DNA at each locus—so if there are consistently smaller peaks that are about a third of the bigger peaks, an analyst would be inclined to conclude that the smaller ones belong to one contributor and the bigger to another. Id. at 326. However, peak numbers and relative ratios can be misleading. The below image shows what is actually a four-person mixture which most analysts would believe to be a two-person mixture:




(D-609) (12T 49-2 to 52-10)

In sum, "with DNA mixtures, more than one genotype combination may be possible at each locus. This ambiguity is an important reason why DNA mixture interpretation is more difficult than testing single-source samples." Mixture Interpretation at 34. For the reasons explained further in the next subsection, with lower levels of DNA, contributors that contribute a very small portion of the total DNA in a sample, and high levels of allele sharing, as occurs with related individuals, it becomes very difficult to determine how many contributors are in a mixture and what each contributor's genotype is.

**4. Some DNA mixtures are very hard to reliably analyze.**

Although not all mixtures are complex, many are. Complexity is a term "for how difficult it is to determine who the contributors to a mixture are." (2T 35-25 to 36-3) Factors that make a sample complex are:

- Low-quantity of DNA from one or more minor contributors; (2T 37-24 to 39-2; 3T 70-9 to 17; 9T 105-19 to 106-6)

- Mixture proportions, including when contributors have made similar contributors or where there is an extreme imbalance in proportions; (2T 36-4 to 37-14; 3T 69-7 to 70-8; 9T 34-11 to 20; 9T 105-7 to 12; 12T 50-12 to 52-10)

- Degradation; (2T 36-21 to 37-2; 3T 69-4 to 6; 9T 34-8 to 10)

- Number of contributors; (9T 34-2 to 4)

- Degree of allelic overlap. (2T 37-21 to 38-4; 3T 70-17 to 71-21; 9T 105-16 to 16; 12T 46-14 to 47-8)

See also Mixture Interpretation at 12.

The more complex a sample is, the greater uncertainty with respect to measurement and interpretation of results." Ibid. See also 10T 105-14 to 16 (Dr. Coble explaining that with "lower amounts of DNA," measurements become more uncertain).

When contributors have given low amounts of DNA, "stochastic" (random) effects arise that create interpretation challenges, including:

- "Drop-in," or sporadic contamination. Drop-in refers to when pieces of DNA (alleles) that are not part of the crime scene sample are nevertheless detected during testing. John M. Butler, Advanced Topics in Forensic DNA Typing: Methodology 324-26 (2011).

- "Drop-out," which occurs when an allele that is in a sample is not detected by the testing machinery because the quantity of the DNA being tested is so small. Thus, pieces of DNA that belong to the DNA profile of a contributor to a sample are literally missing. Ibid.

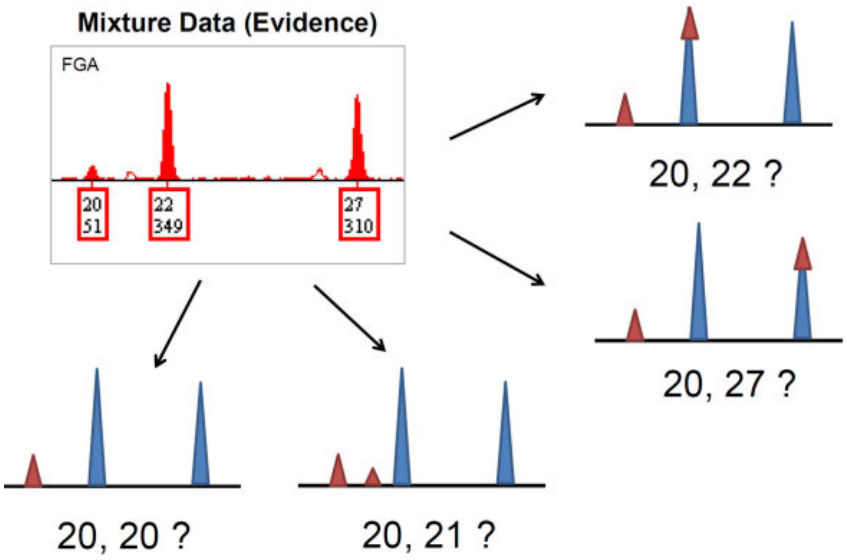
"Analysis of samples containing very small quantities of DNA tends to produce [electropherograms] with a higher proportion of artifacts due to stochastic variation or random sampling of DNA molecules." Mixture Interpretation at 33. "[W]ith low-quantity DNA samples, the resulting profile and [electropherograms] may vary in how accurately they reflect the original sample, which can lead to loss of genotype information from a true contributor to the mixture." Id. at 42. "Furthermore, in part due to stochastic variation, two low-quantity DNA samples collected from the same surface can produce DNA profiles with different peak heights and

therefore different ratios of alleles and possible genotype combinations. Analyzing the same low-quantity DNA mixture two or more times can also produce dissimilar DNA profiles with different degrees of stochastic variation[.]" Ibid.

Contributor proportions also matter. It can be almost impossible to deconvolute mixtures where the contributors have given similar amounts of DNA. But extreme mixture proportions also present a challenge. When a contributor has given a low proportion of total DNA in a mixture, that minor contributor's alleles may also be masked as stutter, an artifact of the PCR process that creates non-allelic peaks. Mixture Interpretation at 41. See id. at 38 ("Stutter products can be indistinguishable from true alleles of minor contributors and therefore can significantly impact DNA interpretation."). Peaks in the stutter position can also make it particularly hard to determine the number of contributors. Id. at 60.

There is uncertainty in every part of the process of traditional DNA analysis, from extraction to quantitation to amplification. (16T 163-10 to 164-1) But the uncertainty is more pronounced and more important the more contributors, less DNA, more extreme proportions, or more allele sharing you have: "[t]he difference between mixtures and simple DNA samples means that all the other variations and uncertainty inherent in DNA analysis can have a much greater impact on the outcome in complex DNA analysis." Mixture Interpretation at 25.

In addition to the amount of DNA contributed, the number of people who have contributed DNA can make a mixture complex. The more contributors that may be in a sample, the "more possible genotype combinations with any observed set of alleles," making deconvolution (separating each genotype out) more challenging. Ibid.




(D-607) (12T 43-1 to 46-7)

The above chart shows the ambiguity inherent in trying to figure out the genotypes present at a specific locus. The electropherogram shows DNA present at alleles 20, 22, and 27. Counter-clockwise from bottom left, the possibilities of possible genotype combinations are:

- one person contributed very little DNA (the minor contributor) and is a 20, 20, and the other person contributed much more and is a 22, 27.
- One person contributed 20, 21, but the 21 was not detected as a DNA peak by the electropherogram. The other contributor was 22, 27.
- One person contributed 20, 27 and the other 22, 27. The minor contributor's 27 allele is masked on the electropherogram by the major's 27.
- One person contributed 20, 22 and the other 22, 27. The minor contributor's 22 allele is masked on the electropherogram by the major's 27. (12T 43-1 to 46-7)

As the above demonstrates, it is very hard to accurately separate out the genotypes of the people who have contributed to a complex sample. "Interpreting the mixture requires an assessment of weighted possibilities for which alleles go together to form the DNA profiles of

the individual contributors, which are then compared to a person of interest (POI)." Mixture Interpretation at 12-13. Manual DNA interpretation undertaken by a human is largely unable to deal with samples any more complicated than a relatively simple two-person mixture. (1T 109-11 to 13; 5T 98-4 to 9; 9T 18-3 to 22-18)

**5. Related contributors create an intractable difficulty for DNA analysis.**

High allelic overlap is inevitable in the case of related contributors—we share much of DNA with our relatives. Id. at 58. The more closely people are related, the more DNA they share. Ibid. That allelic overlap increases the risk of false positives for a non-contributing relative of the actual contributor. It is undisputed that "[n]ot accounting for relatedness can increase the risk of falsely including a non-contributor relative in the DNA mixture." Id. at 58-59.

Relatedness causes other difficulties with DNA interpretation. Allele sharing makes it harder to determine how many contributors are in a mixture. Ibid. Overestimating the number of contributors generally increases the risk of false positives, and underestimating generally increases the risk of false negatives. (8T 104-10 to 17; 11T 54-9 to 15) Thus, when a mixture contains related people, meaning there is allele sharing in that mixture, the number of contributors may be underestimated, leading to false exclusions of true contributors (related or unrelated) in the mixture. As explained above, it is also harder to determine the number of contributors the more contributors there are and the less DNA there is in a sample. Ibid. at 61. Thus, a low-template sample with related people in it will make estimating the number of contributors quite difficult.

Dr. Buckleton testified that the "risks" of analyzing samples with related contributors can be "mitigated" but "will never be completely gotten rid of" by any method of DNA analysis, including STRmix. (7T 200-1 to 13) Dr. Buckleton has reiterated this problem in his written

work, explaining that "there are at least four effects of close relatedness between DNA donors for mixtures analyzed with STRmix:"

- "Underestimation of the number of contributors due to allele sharing."
- "The deconvolution may preferentially choose an alternate genotype explanation" that falsely excludes true related contributors.
- "High adventitious support" (high inclusionary LRs for non-contributors) "for a non-donating relative of the true sample donor occur more often than when the sample is compared to unrelated non-donors[.]"
- "When comparing non-donors of DNA to a mixture, the greater the fraction of donors that are related to the non-donor, the more frequently the non-donor will be adventitiously supported as a contributor." (In other words, for a mixture with multiple contributors, if more of those contributors are related to a non-donor, it is more likely that the non-donor will be falsely included).

  Tim Kalafut, John Buckleton, et al., Investigation Into the Effects of Mixtures Comprising Related People on Non-Donor Likelihood Ratios, and Potential Practices to Mitigate Providing Misleading Opinions, 59 FSI: Genetics 1, 3 (2022) (Da-1078).

In short, the challenge of accurately interpreting a DNA mixture that has related people within it is ineradicable.

**6. Probabilistic genotyping systems were designed in order to attempt to analyze complex mixtures.**

Probabilistic genotyping systems in general, and STRmix in particular, were designed to try to interpret complex mixtures that humans could not reliably interpret. (7T 167-8 to 11) One of the catalyzing events for moving towards PGS was a NIST study known as the Mix13 study, in which many laboratories falsely included a non-contributor in a challenging mixture. (10T 55-2 to 58-6) The results of Mix13 were presented at a meeting of DNA technical leaders, who were greatly upset by the results. (10T 58-8 to 10; 11T 67-21 to 68-12) The next day, Dr. Coble "gave a presentation on STRmix to show them that there is a way forward." (11T 68-13 to 16)

**7. Probabilistic genotyping systems, including STRmix, will produce false positive and false negative errors.**

Although PGS may overall be a more reliable method of analyzing complex DNA samples than traditional DNA analysis, any PGS, including STRmix, will err. These errors will be more common the more complex the DNA sample is, barring any other quirks of the model or software issues that could cause errors. For the purposes of analysis, when a PGS gives an inclusionary LR for someone who did not contribute to a mixture or an exclusionary LR for someone who did, those are errors.

**a. In probabilistic genotyping systems, false positives are inclusionary likelihood ratios for non-contributors and false negatives are exclusionary LRs for contributors.**

Probabilistic genotyping systems give an inclusionary LR when the information available about a DNA sample looks sufficiently like a person of interest's DNA. Dr. Buckleton referred to inclusionary LRs for non-contributors to a sample as "false positives," as providing "false inclusionary support," and "false inclusions." (7T 24-1 to 12 to 24, 163-21 to 164-9) Exclusionary LRs for contributors to a sample would appropriately be considered false negatives. See also 2T 94-8 to 11 (Ms. Ghannam explaining that a "false exclusion" is an exclusionary LR for a known contributor); 2T 115-13 to 116-2 (Ms. Naughton concurring the a false inclusion is when a non-donor was given an LR greater than 1 and a false inclusion is when a donor was given an LR below 1); 9T 47-7 to 13 (Ms. Thayer explaining that including a non-contributor should be considered a "type 1" error and excluding a contributor should be considered a "type 2" error); 14T 162-25 to 163-9 (Mr. Adams explaining that the "false positive rate" is the "rate at which a non-contributor is inappropriately included" and the "false negative rate" is the rate at which contributors are excluded); 16T 89-25 to 90-11 (Mr. Inman referring to STRmix generating false inclusions and false exclusions).

**b. Probabilistic genotyping systems will frequently give inclusionary likelihood ratios for non-contributors and exclusionary likelihood ratios for contributors.**

Complex samples present challenges to interpretation with probabilistic genotyping systems. A PGS attempts to account for all of the stochastic effects, the impact of stutter, and all of the other features that make it harder to interpret complex samples through its mathematical modeling. But the more complex the sample is, the more likely the PGS is to err. See John Buckleton et al., Response to: Commentary on: Bright et al. (2018) Internal validation of STRmix™ – A multi laboratory response to PCAST, 34 Forensic Science International: Genetics, 34: 11–24, 44 FSI: Genetics 1, 4 (2020) (D-1043) ("The adventitious match rate is determined almost completely by the profile (and more specifically on a per contributor basis, within the profile) and to a very much lesser extent by the software. . . . Hence the reported adventitious match rate would be influenced by the exact mix of samples investigated.").

STRmix cannot distinguish between a person's DNA and DNA from another contributor that is similar. (7T 161-24 to 162-3) The more allele-sharing there is between a person of interest's DNA and the DNA collected from a crime scene, the more likely it is that the evidence "will look like it came from someone that it did not." (7T 162-16 to 19) In Dr. Buckleton's words, "[a]llele sharing does cause a false positive." (7T 164-9) See also Buckleton et al, The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity, 64 J. Forensic Sci. 393, 398 (2019) ("In a large set of mixtures compiled from 31 laboratories, all large (over 10,000) LRs for nondonors were investigated; in all instances, the nondonors had high allelic overlap with the profile. This is the correct result.") (D-1044).

The less DNA available from any person, the less information available about that person's genetic profile, the less it is possible to tell one person's DNA from another's. This is an immutable fact of genetics. Many people will share the same alleles at some loci; some alleles

and allele combinations are more common than others. John Butler, Fundamentals of DNA Interpretation 230 (2010) (D-1). Because many people have the same alleles at some loci; forensic DNA gets discriminating only when there are enough alleles that these accidental similarities are dwarfed by differences. (7T 159-15 to 159-1)

When there is not enough information about sufficiently unusual alleles, forensic DNA is not discriminating; it cannot tell one person from another very well. It is dangerous to infer too much from high similarity between two profiles if relatively few alleles are examined. Dr. Reich illustrated the importance of having sufficient genetic information to discriminate between potential contributors: he testified that when a nine-locus profile was developed for a defendant in a death penalty case and compared to all of the profiles in the Illinois State database, which contained several hundred thousand individuals at the time, "they found hundreds of DNA profiles which were the same at nine. They weren't the same at 13, but [at] nine loci they were the same. That same approach was used in some other cases later on and they identified profiles that were the same at 10 and 11 and 12 [loci], all in the State database." (12T 209-7 to 210-7).

The fact that many people share some alleles means that many people will be given inclusionary LRs for mixtures they did not contribute to when those contributors are low-level (and therefore there is not information about them at the majority of loci) or there is much allele sharing. Dr. Buckleton explained, this is a necessary, immutable fact of DNA, and even the best PGS will run into this fact: "A DNA profile with limited information"—meaning not a lot of DNA—"will not allow high discrimination and will typically give inclusionary support for many genotypes. Thus, as the information in a profile decreases, more adventitious inclusionary support is observed." John S. Buckleton et al., Are Low Lrs Reliable?, 49 Forensic Sc. Int'l: Genetics at 2 (D-1063). "Adventitious support" means an inclusionary LR for someone who

didn't contribute due to accidental (that is what adventitious means) genetic similarities. "Adventitious matches may also occur when comparing mixtures with individuals who are closely related to contributors or due to inbreeding effects." Ibid. See also 7T 24-14 to 18 (false inclusions occur when "[a] person who's a non-donor just happens to have the right combination of alleles to fit this mixture"); 7T 107-18 to 25 (Dr. Buckleton: "[R]eally bad mixtures, five person mixtures with low peaks have high false inclusion rate and really tidy single source samples have an enormously low false inclusion rate."); 11T 33-18 to 25 (Dr. Coble: false inclusions will occur with any PGS system as many people "have a lot of alleles that are out there in the population and just by random chance you may find, especially when you have a low level sample where you only—instead of looking at all 20 STR markers you are only looking at a handful, three or four, then yes you expect to find potentially people who would have likelihood ratios greater than one" who did not actually contribute to a mixture). As Ms. Reed explained about an LR of 18.5 "given the low amount of data" that was present in the crime scene profile "probably the majority of this room would be included in this profile." (5T 78-12 to 14)

As general matter, overestimating contributors creates the risk of generating false inclusions and underestimating to false exclusions. (8T 104-10 to 17; 11T 54-9 to 15) The more complex a sample, the harder it is to appropriately estimate the number of contributors. (2T 38-20 to 39-6; 3T 74-23 to 72-5, 89-20 to 90-2, 102-1 to 13; 5T 58-8 to 16, 88-11 to 23; 7T 164-7 to 14; 9T 29-2 to 15, 108-18 to 109-1; 11T 55-24 to 56-22)

In sum, the risk of error when analyzing complex mixtures does not go away when probabilistic genotyping is used, as opposed to traditional analysis. Some PGS may be more reliable than manual interpretation, but that does not make any given PGS reliable, let alone error free.

**8. Because the error rate of probabilistic genotyping systems will vary across sample types, reliability must be established across those sample types.**

As explained above, the degree of the reliability of PGS "depends on sample complexity." Mixture Interpretation at 15. Therefore, a system cannot be considered reliable or unreliable "without considering context"—the kind of sample being analyzed and the amount of information available about the ability of that PGS to analyze that kind of sample reliably. Ibid. In other words, no PGS could be unqualifiedly "reliable." It could only be considered reliable across a specific range of samples of varying complexity—as is created through decreasing amount of DNA, extreme contributor proportions, and relatedness, among other factors—that testing has shown it consistently analyzes correctly.

**a. Both developmental and internal validation are necessary to demonstrate the reliability of the use of probabilistic genotyping systems in casework.**

The ability of a PGS to analyze samples reliably is measured through validation studies. In the forensic DNA context, validation means something completely different than in software engineering. (7T 167-17 to 20) In forensic DNA, a validation study tests a method against different kinds of samples to see if the outputs are as expected. As explained above, software engineering validation is a term of art that refers to something different, and because PGS involve both software and forensic DNA analysis, both forms of validation are necessary.

In the field of forensic DNA analysis, validations studies are mandatory before the use of any new method. There are two types of validation studies: developmental and internal. As discussed further below, developmental validation studies must demonstrate the range of samples that the software as designed can reliably analyze, and internal validation studies must demonstrate the range of samples that the software as implemented in any given laboratory can reliably analyze. In the words of NIST, the reliability of any method of DNA mixture

interpretation, including PGS, "cannot be established without validation testing using known samples of similar complexity" as those encountered in casework Ibid.

**b. Developmental validation must establish the outermost bounds of the reliable use of probabilistic genotyping systems.**

Developmental validation is "the acquisition of test data and determination of conditions and limitations of a new or novel" method used on forensic samples. Federal Bureau Of Investigation, Quality Assurance Standards for Forensic DNA Testing Laboratories 3 (2011) (D-204). In other words, developmental validation is supposed to determine if and when a new technique produces reliable results. As NIST explains, developmental validation is necessary to show that a forensic test method is "fundamentally valid." National Institute of Standards and Technology, Views of the Commission: Validation of Forensic Science Methodology (2016) (D-1204). See also Human Factors at 337 ("Developmental validation refers to the process of determining the conditions under which newly translated DNA methodologies work well and establishing the limitations of the technology.") (emphasis added).

Developmental validation can support a claim of reliability only as to the types of samples tested in that validation. Mixture Interpretation at 100 ("Reliability statements based on aggregate performance across many types of samples and many different probabilistic genotyping software (PGS) systems do not provide the information needed to judge the degree of reliability of the measurement and interpretation in a particular case of interest."). ANSI/ASB Standard 18 requires that a developmental validation "shall include case-type profiles of known composition that represent (in terms of number of contributors, mixture ratios, and total DNA template quantities) the range of scenarios that would likely be encountered in casework." (S133/D212 at 3) Laboratories can go no farther in using a PGS than the developmental

validation went: "Case type profiles that fall outside the range of conditions explored in developmental validation shall require additional developmental validation studies." Ibid.

**c. Internal validation studies are necessary to establish the reliability of a method and are a mandatory prerequisite to their implementation in any laboratory.**

Developmental validations are necessary but not sufficient for the reliable use of PGS. "Even when methods have foundational validity, application in an individual case may or may not be reliable." Id. at 8. That is why internal validation is necessary.

Internal validation is the "acquisition of test data within the laboratory to verify the functionality of the system, the accuracy of statistical parameters, the appropriateness of analytical and statistical parameters, and the determination of limitations of the system." ANSI/ASB Standard 18, Standard for Validation of Probabilistic Genotyping Systems 2 (2020) (S133/D212). Every DNA expert at this hearing testified that internal validation studies are a mandatory prerequisite to the implementation of a new method. (1T 13-6 to 7 to 19-3, 101-8 to 106-14; 2T 11-4 to 13-1; 2T 56-19 to 59-19; 4T 9-14 to 16; 6T 68-25, 70-6; 7T 34-3; 8T 10-2 to 5, 21-10 to 12, 33-14 to 17, 48-1 to 21, 53-11 to 13; 10T 79-18 to 23; 11T 42-11; 12T 67-3 to 7; 16T 85-20 to 23, 142-2 to 6). An internal validation study is how a laboratory demonstrates that it can reliably use a method in its laboratory. (1T 13-6 to 7 to 19-3, 101-8 to 106-14; 2T 11-4 to 13-1; 2T 56-19 to 59-19; 4T 9-14 to 16; 6T 68-25, 70-6; 7T 34-3; 8T 10-2 to 5, 21-10 to 12, 33-14 to 17, 48-1 to 21, 53-11 to 13; 10T 79-18 to 23; 11T 42-11; 12T 67-3 to 7; 16T 85-20 to 23, 142-2 to 6)

Every DNA expert testified that each laboratory must conduct an internal validation for each method it seeks to implement itself. (1T 13-2 to 16; 2T 77-9 to 79-24; 5T 17-1 to 13; 6T 54-13 to 19, 82-24 to 83-3; 8T 46-11 to 48-21, 53-11 to 13; 10T 79-7 to 14, 80-7 to 9; 12T 67-3 to 7) One laboratory's internal validation cannot substitute for another laboratory's, because each

laboratory must demonstrate its own ability to reliably use the method. (3T 64-18 to 25, 77-22 to 24; 5T 17-1 to 13; 6T 54-13 to 19; 8T 48-1 to 21, 68-15, 70-10; 10T 80-7 to 9, 40-1 to 4; 12T 69-6 to 11, 122-6 to 14, 173-17 to 174-6) As Ms. Ghannam explained, this is because each laboratory has "different instructions, different equipment, different setups, different procedures. So even though the developmental validation studies have said that this procedure work, you have to see if it works in our laboratory." (1T 20-24 to 21-5) In other words, the same piece of software used by different laboratories may not yield the same results.

The FBI QAS, which each laboratory discussed in this case is accredited to and is required to follow, requires that a laboratory perform its own internal validation. (1T 13-2 to 16; 12T 173-12 to 174-6); FBI, Quality Assurance Standards for Forensic DNA Testing Laboratories 18 (2011) (S5/D204). The SWGDAM Guidelines on probabilistic genotyping, which as witnesses explained are not mandatory but still have very significant influence on laboratories in the United States (1T 40-15 to 22; 6T 73-2 to 19; 10T 97-11 to 19), also require each laboratory to perform its own internal validation. SWGDAM Guidelines at 4 (S-6/D-202).

**d. Internal validation studies establish the limits of a laboratory's ability to reliably use a technique, including probabilistic genotyping systems.**

The internal validation describes the outer boundaries of what a laboratory can claim it can reliably analyze. An internal validation "will establish the function limit of a particular system" within the laboratory. Human Factors at 55. Casework profiles should only be analyzed that fall "within the validation range." Ibid. Internal validations, and the development of protocols that stem from them, are a mandatory requirement of a laboratory's accreditation.

**i. Standards, guidelines, and best practices require the use of internal validations to set the limits of what kinds of samples a laboratory will analyze.**

Because of the difficulty in reliably analyzing complex samples, "complex mixtures and low-level contributors should be evaluated thoroughly during internal validation, as the data from such samples generally help to define the software's limitations, as well as sample and/or data types which may potentially not be suitable for computer analysis." SWGDAM Guidelines at 8 (emphasis added). As Ms. Thayer explained, SWGDAM requires that validation samples be representative of casework samples. (9T 37-11 to 17) To review complex samples in case work, "you need to validate the method on complex samples as well." (9T 37-18 to 21)

ANSI/ASB Standard 18 requires that for internal validation, "the laboratory shall evaluate both the appropriate sample types (i.e., number of contributors, mixture ratios, and template quantities) and the number of samples within each type to demonstrate the potential limitations and reliability of the software." (S133/D212 at 3) See also Mixture Interpretation at 99 ("PGS models may or may not work satisfactorily when applied to data that are unlike scenarios considered in the internal validation training set. Identification of those scenarios in which the performance of a specific method is judged to be inadequate will assist in establishing operational limits for the types of samples that may be reliably interpreted and also point to areas where the measurements or models require improvements.") (emphasis added).

All but one DNA expert asked agreed with the well-established proposition that a primary purpose of an internal validation summary is to figure out the limits of the reliable use of a method. (1T 94-24 to 85-8, 92-16 to 93-25; 2T 12-10, 18-14 to 21, 35-21; 3T 45-1 to 24, 64-14 to 68-1; 8T 72-3 to 8, 9T 11-8 to 12, 23-13 to 20; 10T 24-5 to 10, 79-18 to 24). Mr. Inman explained that in a validation study, "if you can break the system then you have an idea of where you should stop if you will or perhaps be in a danger zone if you see the sign posts of what you've done to break it." (16T 60-9 to 12)

Dr. Buckleton was the only expert who did not agree that one purpose of a validation study is to establish limits on the reliable use of a method. He opined that NIST has "driven us to the low end, NIST has got this fear that there's something wrong at the low end. And it turns out we don't have a bound there." (6T 83-20 to 84-2) In other words, Dr. Buckleton does not believe there is any situation in which STRmix cannot be reliably used and that therefore no limits can or should be established through internal (or developmental) validation studies. This position seems to contradict previous statement he has made: "An understanding of the models within each of the program[s] and their limitations is required in order to validate interpretation software." John Buckleton et al., A Series Of Recommended Tests When Validating Probabilistic DNA Profile Interpretation Software. 14 FSI: Genetics 125 (2015) (emphasis added).

**ii. Standard operating procedures require the setting of a limit on what kinds of samples a laboratory will analyze.**

From the data gathered from the internal validation studies, laboratories should proactively set limits on what samples can and cannot be analyzed in casework, set forth in Standard Operating Procedures (SOPs). As NIST explains, "[v]alidations attempt to test samples reflective of casework, and SOPs use this information to provide a framework for analysts' tasks and steps." NIST, Human Factors at 28. Clear SOPs not only make sure the laboratory as a whole has determined its capabilities and limitations (and is sticking to them), but they also "arm analysts with the tools needed to make educated and empirically supported decisions and reduce inter- and intra-analyst variability." Ibid. Together, internal validation studies and SOPs find and define the limits of what samples can be reliably analyzed in a specific laboratory. Using an internal validation to inform SOPs is a requirement of the FBI Quality Assurance Standards. FBI QAS Standard 9.1 ("The laboratory shall have and follow analytical procedures supported by the internal validations and approved by the technical leader"); FBI QAS Standard 9.6 ("The

laboratory shall have and follow written guidelines for the interpretation of data that are based on and supported by internal validation studies).

**iii. Establishing a limit means establishing a boundary of what kinds of samples a laboratory will and will not analyze.**

Although all of the State's experts who currently conduct DNA analysis agreed that an internal validation must find the limits of a laboratory's reliable use of STRmix, the Bode experts testified that they did not feel that the boundaries of what they tested established any limits for the laboratory (3T 91-9 to 13; 5T 95-20 to 96-4) and, as discussed further in subsection H, analyzed samples outside of these boundaries. Dr. Coble agreed that the internal validation must establish limits on a laboratory's use of STRmix. Dr. Coble explained that "the real purpose of doing a validation study is to try to determine the limits of the software." (10T 79-18 to 24) However, he then undermined this position, saying that in determining whether the use of STRmix is reliable "you have to look at the whole" rather than follow objective limits. (10T 110-20 to 21) This claim that an internal validation study should set limits but that those limits need not be objective or adhered to renders the concept of a limit meaningless.

A limit means that laboratories should not analyze samples or report results that go beyond what they are validated to reliably analyze or report. The plain meaning of the word as commonly understood is of a boundary, something that restricts action. See also Merriam-Webster, Limit ("(a) something that bounds, restrains, or confines; (b) the utmost extent"), https://www.merriam-webster.com/dictionary/limit. A limit is not a limit if it does not constrain actions. SWGDAM, ANSI/ASB Standard 18, and the National Institute of Science and Technology used that word on purpose; it should be understood in accordance with that purpose.

At the hearing, the defense DNA experts echoed the meaning of limit as used by NIST, SWGDAM, and ANSI/ASB. As Dr. Inman explained, the idea of an internal validation study "is

to make the sample as complex as you expect to find within your casework and to mimic those as much as possible and I would say to go beyond . . . . The other part of that is that there will come a time when you do see something that you have not covered and there you need to both recognize it and back away from it and say like I don't know what to make of this." (16T 74-11 to 75-1) Dr. Reich explained that it is inappropriate for a laboratory to analyze casework samples more complex than it demonstrated it can reliably analyze in its validation study: "Technically and scientifically, if a sample/DNA profile falls outside of the range of samples/DNA profiles that Bode Technology tested for the reliable, reproducible and accurate analysis using STRmix, then there would be no support for deriving any forensic DNA conclusions from the particular samples that fall outside of the tester range used in validation." (D-15 at 3) See also Human Factors at 197 ("In examining and identifying the limitations of a method or technology," through an internal validation study, forensic science service providers "will identify the boundaries in which to operate and the signs when they are approaching those limitations.").

In short, the defense experts agree with the accreditation bodies, standards bodies, and the agency that is committed to the reliability of forensic science that validation studies establish a hard limit on the types of samples PGS can reliably analyze. Even UCPO, which has been using STRmix for longer than NJSP, has developed a hard limit of this kind. UCPO laboratory will not report any results with any contributors of less than 2% to the total mixture. (D-24) The only people who seem to disagree, contending that a limit is not a limit but an amorphous suggestion that can be ignored for unexplained reasons are the people who have either built this software or use this software for business, and therefore have an incentive to use it as much as possible.

Notably, all of the DNA analysts who testified in this case do believe that their internal validation studies established some hard limits. All the DNA analysts testified that they believe

the number of contributors is a hard limit established by their validation studies; they would not attempt to analyze a mixture with more contributors than were tested in those studies, which is 4. (3T 68-10 to 16; 4T 8-21 to 9-2) There was no explanation presented at the hearing for why the number of contributors tested in the validation study was considered to be a "hard limit" by Bode and NJSP when no other hard limits seemed to exist. There is no logical reason for that to be the case. Just as the contribution proportions and total template amounts can be incorrect, so can an analyst's number-of-contributor estimate, as the true number is always unknown and unknowable in casework. (3T 89-16 to 19; 7T 78-14 to 18; 9T 28-18 to 21) Yet number of contributors is a hard limit, as it should be. It's the lack of other limits that is inconsistent with the purposes of an internal validation study and the relevant guidelines and standards for PGS.

An analogy might be helpful. Imagine a ruler that has been demonstrated to be able to measure distances as short as a tenth of an inch. No one has ever checked if the ruler can measure even smaller distances. If someone brought out that ruler and claimed to be able to accurately measure a thousandth of an inch, that claim should be rejected: just because an instrument can do one thing does not mean it can reliably do a harder thing. And someone who uses rulers a lot saying, "It feels like this ruler can reliably measure a thousandth of an inch," would not be sufficient to demonstrate that it could actually do so. Without empirical testing and validation of the claim that it can do the harder thing, it is at most an assertion of hope, but not scientifically sound. See also Mixture Interpretation at 73 ("An important hallmark of science is to develop reliable theories and methods based on empirical data[.]") (emphasis added).

**e. It is possible for internal validation studies to study the impact of relatedness on a probabilistic genotyping system's reliability.**

As found above, subsection E.5, relatedness has many impacts on the reliability of PGS. Two types of relatedness issues need to be considered in an internal validation: (1) the impact of

a non-contributor being related to the true contributor; and (2) the challenge of properly deconvolving and interpreting a sample with related contributors within in it. It is quite possible for internal validation studies to study the effect of relatedness on the performance of probabilistic genotyping systems, and STRmix in particular, in their laboratories. In fact, many of them do.

**i. Internal validations reveal that non-donors related to the real donors to a sample are at a significant risk of being falsely included in that sample at a high likelihood ratio.**

The results from internal studies that examined related contributors give much reason for caution. Using the DNA from one sibling relationship, the Sacramento lab made a total of 29 comparisons with sibling non-contributors. (D-108 at 29) The lab tested a total of 24 three-person mixtures against a non-contributor sibling of a true contributor. Eleven of those 24 comparisons led to numbers above 1—that is, led to false positives. (D-108 at 29) The highest false positive, 59.5 trillion, occurred in a comparison with a sample indicated to be 32.5 picograms. (D-108 at 29) Another false positive, 642 billion, occurred with a sample indicated to have 15.625 picograms. (D-108 at 29)

Similarly, the Palm Beach County Sheriff's Office's internal validation study tested both mixtures comprised of related people and tested related non-contributors against samples a relative contributed to. The summary revealed false positive LRs as high as $10^6$ for non-contributors that were related to true contributors to a mixture. (D-107)

Los Angeles County's internal validation study revealed so many false positive inclusionary LRs for related individuals that it concluded that "[a]nalysts should use caution when interpreting mixtures believed to be comprised of first-order relatives." (D-104 at 61) The underlying validation data, when obtained by a defense attorney and analyzed by an outside organization, revealed high levels of false positives for related contributors that were removed

from the validation summary. People v. Collin T., Notice Of Motion To Preclude Expert Testimony Or For A Limited Remote-Video Hearing under Frye/Wesley at 33 (N.Y. Sup. Ct. June 18, 2020) (D-1217).

A review of nine different laborites that did analyze related contributors in their internal STRmix validations found high false positive LRs for non-contributors that were related to people in the mixture:

| DNA Crime Lab | Highest Non-Related | Highest Related | Difference as a factor of |
|---|---|---|---|
| LA County Sherriff's Dept. | $10^{1}$ | $10^{15}$ | 100,000,000,000,000 |
| Sacramento Cty. D.A.'s Crime Lab | $10^{1}$ | $10^{13}$ | 1,000,000,000,000 |
| Palm Beach Cty. Sheriff's Office | $10^{6*}$ | $10^{17}$ | 100,000,000,000 |
| Las Vegas Metropolitan Police Dept. | $10^{3}$ | $10^{13}$ | 10,000,000,000 |
| Colorado Bureau of Investigation | $10^{2}$ | $10^{10}$ | 100,000,000 |
| Jefferson Cty. Regional Crime Lab | $10^{2}$ | $10^{9}$ | 10,000,000 |
| DNA Labs International | $10^{2}$ | $10^{7}$ | 100,000 |
| Wisconsin State Crime Lab | $10^{2}$ | $10^{7}$ | 100,000 |
| Oregon State Police Portland Metro | $10^{4}$ | $10^{6}$ | 100 |

In Defense Of, The Kinship Problem, https://indefenseof.us/issues/kinship-problem (last visited Oct. 21, 2024) (D-1216).

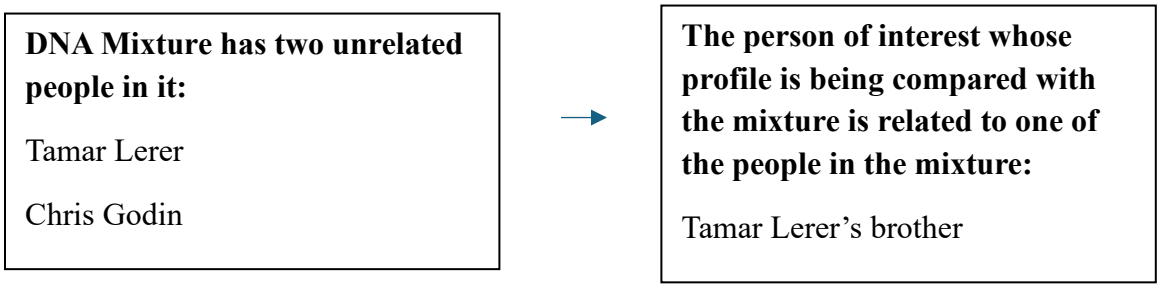
New Jersey State Police Laboratory is attempting a relatedness study of STRmix now. (9T 32-12 to 15) This demonstrates that not only is such a study possible, it is desirable.

**ii. The use of the likelihood ratios produced by STRmix is inappropriate for samples with related contributors.**

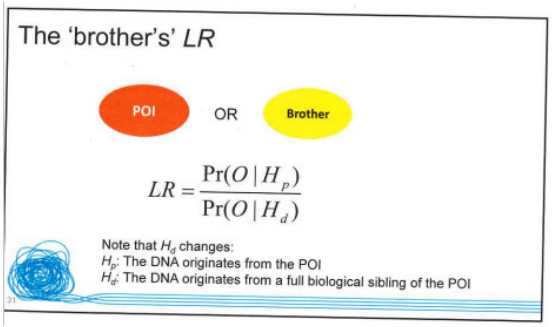
Although laboratories can test mixtures comprised of related people in their validation study, the resulting LR is never the appropriate one to report. Therefore, studying that circumstance can help a laboratory understand what the risks are of their interpretations in terms of trends, but the LR reported by STRmix is not accurate in those cases. That is because apart

from the general problems DNA interpretation (manual and PGS) has with deconvoluting mixtures that may falsely include people who did not contribute to a sample, the STRmix program cannot properly account for mixtures containing related individuals. This is a different issue, although potentially overlapping, than the one that confronts STRmix when a non-contributor is related to a contributor. The problem of adventitious matches when a non-contributor is related to the true contributor looks like this:

In that case, Tamar's brother is at a high risk of false inclusion. The sibling LR can do some accounting for this problem by changing the




math in the alternate hypothesis, although that problem cannot be completely solved, as explained at the hearing: there is always an enhanced risk of a false positive when the non-contributing person of interest is related to the true donor. See subsection F.5. The modified LR does this by considering, in the alternative hypothesis, that the person of interest's sibling was the true contributor, and comparing that to the hypothesis that the person of interest was:




STRmix Full User Training Workshop Guide at 202 (2024). See also Duncan Taylor et al., Considering Relatives When Assessing The Evidential Strength Of Mixed DNA Profiles. 13 FSI: Genetics 259 (2014) (D-1003).

It should be noted that STRmix considers the possibility that only one relative is the true contributor. The sibling LR cannot account for the possibility that one of five brothers could have contributed to a mixture.

The problem that arises when related people are in the mixture is different, and it looks like this:

| **DNA Mixture has two related people in it:** | → | **The person of interest whose profile is being compared with the mixture is related or unrelated to anyone in the mixture:** |
|---|---|---|
| Tamar Lerer<br>Tamar Lerer's brother | | Tamar Lerer's sister<br>**Or**<br>Chris Godin (N.B. Tamar and Chris are unrelated) |

Regardless of whether the POI is related to the true contributors, no LR produced by STRmix can properly account for the fact that there are related people in the mixture. STRmix will always treat the two people in the mixture as unrelated for the purposes of generating an LR, and that is wrong when the people in the mixture are related. If Tamar Lerer's sister is being compared to a mixture comprised of related individuals who are related to her, the sister is at

elevated risk of a false inclusion. But the likelihood ratio produced by STRmix is inappropriate for mixtures that contained related people even if the person compared to that mixture is unrelated to anyone in it. While mathematical solutions exist within the literature that properly account for this situation, STRmix has not incorporated those solutions, as Dr. Coble explained.

Dr. Coble explained that STRmix "cannot do the test" of a mixture "with two siblings." (11T 41-23 to 42-7) Instead, "[i]t treats the two people" within the mixture "as unrelated." (11T 41-23 to 42-7) If you want to do a test that generates a result for a mixture with related individuals, "you have to use a different kind of software. A software that takes kinship into account. But what -- when you have a sample that you swab whether you know they are brothers or not STRmix is going to run them as non-related individuals." (11T 41-23 to 42-7) Coble did agree that "obviously we should do studies with relatives, I think it is a good idea," to see the effect of the software "treating people as non related[.]" (11T 42-16 to 21)

Mr. Inman echoed this explanation. STRmix has the ability to generate LRs when the alternate hypothesis is that someone related to the person of interest contributed to the mixture. (16T 82-22 to 83-16) However, "your likelihood ratio for a random unrelated or against a random unrelated person is not even remotely descriptive of the weight to assign to any likelihood ratio that you get" for related people. (16T 81-15 to 18) The problem is fundamental to the way the likelihood ratio is calculated: it "depends on how common" certain genotypes are in the population, which is not an accurate reflection of commonality among related people. (16T 81-19 to 82-13) "[S]o clearly a likelihood ratio based on random individuals does not describe a situation and really is inappropriate to use in a situation where you have relatives within the mixture itself." (16T 82-14 to 17)

Studying how STRmix performs with mixtures of related people may give additional insight into the limitations of the program and it is certainly possible to do. But such a study cannot fix the fundamental problem that the likelihood ratio cannot account for the presence of related people within a mixture.

**9. Even with the best-calibrated PGS being used appropriately, many non-contributors would yield an inclusionary likelihood ratio if their profile was run against a complex sample.**

Turing's Rule is a mathematical relationship that explains the hypothetical error rate for a PGS. Turing's Rule posits that in ground-truth experiments for a "well calibrated" system that is "performing well," a PGS will produce an LR greater than x for 1 in x false donors. Human Factors at 83. This means that theoretically, in a well-calibrated system that is performing well, an LR of 1,000,000 will occur for every 1 in one million non-donors. Ibid. The other 999,999 should yield an LR of 0. Ibid.

This rule, of course, assumes that a model is well-calibrated and performing well, which would need to be proven for any given system. But even assuming that, if Turing's Rule holds, the best PGS in the world would frequently generate false positives. As Mr. Inman explained, a falsely inclusionary LR of 10,000 should occur once for every 10,000 non-donors run against a sample. (16T 89-19 to 90-11). Assuming New Jersey's population is 8,000,000 (an undercount used at the hearing), 800 non-donors in New Jersey would yield a false inclusionary LR of 10,000. (16T 89-19 to 90-11). In Monmouth County, which has about 640,000 people, 64 people in Monnmouth County would be falsely included at an LR of 10,000. United States Census, QuickFacts, Monmouth County, New Jersey, https://www.census.gov/quickfacts/fact/table/monmouthcountynewjersey/PST045223.

Even in this perfect system, low false positive LRs will occur with even greater frequency. Turing's Rule theorizes that an LR of 4 for a non-donor should occur for every 4 non-donors. A false inclusionary LR of 100 will happen once every 100 donors.

Thus, even assuming a perfect PGS that is operating as best as its can, errors will occur with frequency. Compounding the risk of error is reliance on human judgment in any PGS, including STRmix. That reliance and its risks are discussed further below.

**10. Probabilistic genotyping systems, including STRmix, rely on human analysts to make discretionary decisions in operating the software.**

Although probabilistic genotyping systems are pieces of software, human analysts are necessary parts of the process that leads to a PGS output. Human analysts make decisions, have to spot errors by understanding diagnostics, and are subject to cognitive bias and other limits, as all humans are. The role of the analyst and the risks that brings to the reliable use of PGS are essential to understand.

**a. Human analysts make decisions about STRmix inputs**

Analysts are required to make many decisions about STRmix inputs. For instance, analysts must decide whether a sample is suitable for analysis in STRmix. Human Factors at 41. They must decide which profiles to run and, in some laboratories, which to visually exclude. (1T 51-21 to 25; 5T 6-18 to 7-11) They must decide which hypotheses to give STRmix in order to create a likelihood ratio. Mixture Interpretation at 51; Human Factors at 28. And they must tell STRmix how many contributors are in a given mixture. Laura Russell et al., A Guide To Results And Diagnostics Within A STRmix™ Report, WIREs Forensic Sci. 2 (2019) (D-1051); (8T 101-17 to 24); Mixture Interpretation at 25. In short, human judgment is essential to the use of STRmix from the very first step.

**b. Probabilistic genotyping systems, including STRmix, rely on human analysts to use their understanding of diagnostics to spot errors**.

After data has been input and a PGS has been run, there are indicators that are supposed to alert an analyst to any issue with the analysis. These are called diagnostics, and they must be interpreted by analysis. See also National Institute of Standards and Technology, Forensic DNA Interpretation and Human Factors: Improving Practice Through a Systems Approach 72 (May 2024) (D-7) Human judgment is essential to understanding these diagnostics and fixing any problems they may indicate.

**c. STRmix diagnostics are new to DNA analysts making these discretionary decisions.**

Despite how important the diagnostics are, they are quite new to most analysts using PGS. As to STRmix in particular, the risk of error is supposed to be mitigated by the human analyst's review of a dozen different diagnostics that represent various aspects of the model's performance during a given run. Russell, supra, at 5-7. Analysts are expected to combine information about genotype weights and mixture proportions, along with allele and stutter variance, Markov Chain mixing and convergence, total number of iterations and the rate at which those iterations were accepted, and others to determine whether the answer provided by STRmix seems correct to them. Ibid.

The particulars of how to interpret these diagnostics, both individually and collectively, are left to the "intuition" of the analyst running the program. Id. at 5 Dr. Buckleton believed that one of the primary diagnostics, mixture proportions, allows an analyst to "eyeball it" and see if it agrees with the mixture proportions that STRmix has suggested. (6T 58-17) Yet Dr. Coble testified that analysts cannot discern the contributor percentages from an electropherogram. (10T 92-15 to 18) How an analyst can eyeball an electropherogram to see if a metric they cannot discern comports with STRmix's estimate for that metric is unclear.

In addition to the primary diagnostic of genotype weights, there are more complicated secondary diagnostics. "The secondary diagnostics are new concepts to most analysts and are more difficult to check for intuitiveness. There is no 'right' value for each of the secondary diagnostics although there may be a range of expected values dependent on the profile presentation and complexity." Russell, supra, at 5-7. See also 6T 59-10 to 12 (Dr. Buckleton explaining that secondary diagnostics are new to the analyst); 7T 172-13 to 17 (same). Dr. Buckleton testified that combining these diagnostics is "always complicated." (7T 173-9 to 11) In other words, analysts are supposed to use their intuitive understanding of new concepts with no correct value in order to catch issues with the program. Every STRmix analyst testified that they these diagnostics are new to them since they began to use STRmix and struggled to define many of the diagnostics or explain their significance. (2T 8-5 to 20, 22-8 to 13; 3T 60-8 to 61-25; 5T 102-10 to 104-11; 9T 39-18 to 40-19, 101-25 to 102-20)

**d. Human analysts are subject to cognitive bias and another cognitive limitations.**

It was undisputed that analysts and their judgment are a critical part of the use of STRmix. (1T 34-15 to 18; 3T 101-9 to 17; 4T 7-24 to 8-24) Because humans are central to the use of PGS, including STRmix, "the strengths and limitations of human cognition will be central to forensic casework." Human Factors at 12. In particular, that means that "there is always a chance that expectations, task-irrelevant information, and preexisting prejudices can affect decision-making." Ibid. Of particular concern are cognitive bias and contextual bias. Ibid. All humans are subject to bias. Ibid. Critically, "[t]asks requiring more cognitive effort are generally more susceptible to bias. Greater cognitive effort is required when results are complex and data are ambiguous, when there are time pressures, when a large amount of information must be combined and processed, or when decisions are discretionary." Id. at 14.

The evidence available shows that the use of STRmix varies widely among laboratories and analysts. One study, co-authored in part by three employees at Bode, demonstrates the lack of consistency in the use of PGS. Lauren M Brinkac et al., DNAmix 2021: Laboratory Policies, Procedures, And Casework Scenarios Summary And Dataset, 48 Data in Brief 1 (2023) (D-1210). This study examined the decisions made before interpretation across 83 forensic laboratories. Id. at 3. A few examples demonstrate just how non-uniform the choices laboratories have made about using PGS are. It found that just over half of the participating labs will not interpret samples with less than 10pg of DNA, but half will. Id. at 21. Half of the laboratories require a minimum number of loci with data in order to interpret a DNA mixture, ranging from 2-15 loci with data; half of them have no such requirement. Id. at 21. Default analytical threshold values ranged from 40-200 RFU. Id. at 8.

Another study found that when presented with a four-person mixture with allele drop-out, analysts selected numbers of contributors ranging from 3 to 5, which resulted in LRs ranging from 102 to 1028, and multiple false exclusions. Mixture Interpretation at 96.

Thus, the unavoidable fact that human analysts' subjective decisions influence STRmix outputs demonstrates the necessity of objective protocols, rigorous training, and demonstrated proficiency. When a human analyst is using her subjective judgment, it is essential that procedures "be carefully defined" to attempt to limit the effects of "human error, inconsistency among examiners, and cognitive bias." President's Council of Advisors on Science and Technology, Forensic Science in Criminal Courts: Ensuring Scientific Validity of Feature-Comparison Method 49 (2016) ("PCAST Report") (D-6).

In other words, even a perfect PGS is going to sometimes yield erroneous results because of the human using it. It is important to mitigate the risk of human error in use of PGS.

**11. The outputs of probabilistic genotyping systems are likelihood ratios, numbers that should be reported quantitatively.**

As explained at the beginning of these findings of fact, PGS produces a likelihood ratio as its output. That output is numerical. However, SWGDAM has developed a "verbal scale" that gives qualitative equivalents to a range of LRs. (1T 65-2 to 6; S-9) This is the verbal scale:

| ***LR* for $H_p$ Support and 1/*LR* for $H_d$ Support** | **Verbal Qualifier** |
|---|---|
| 1 | Uninformative |
| 2 – 99 | Limited Support |
| 100 – 9,999 | Moderate Support |
| 10,000 – 999,999 | Strong Support |
| ≥1,000,000 | Very Strong Support |

NJSP and Bode have both adopted the verbal scale, although NSJP has modified it to have any LR less than 1000 be reported as "inconclusive." (1T 67-14 to 68-4) Bode reports all LRs greater than 1 according to the verbal scale. Despite the reliance on this scale, no evidence was provided at the hearing to support its use.

There has been no empirical support provided for the arbitrary cutoffs established by the verbal scale. Neither Dr. Reich nor Mr. Inman believed its use is appropriate, which shows its lack of acceptance in the forensic DNA community. (12T 60-4 to 61-23; 16T 87-11 to 88-23) In their writings, Dr. Coble and Jo-Anne Bright, a co-developer of STRmix, agree that "'[t]here has been some justifiable criticism that LRs are not understood by our audience. The use of words to represent the strength of evidence has been proposed as a way to supplement numerical LR evidence. The assignment of words to a numerical LR scale is, of course, arbitrary…and there are a number of different scales used around the world for different jurisdictions[.]'" Mixture

Interpretation at 50 (quoting Jo-Anne Bright and Michael Coble, Forensic DNA Profiling: A Practical Guide to Assigning Likelihood Ratios (2020)).

The verbal scale is also misleading because it can tend to make weaker associations seem very strong. Dr. Reich explained that an LR of 43,000, for instance, a number in the "strong support for inclusion" on the verbal scale is a "almost meaningless association," because it involves so little information about a person's genetic profile. (12T 94-7 to 96-8) See also 12T 63-1 to 12 ("[W]hen we have a likelihood ratio of a number that has 26, 27, 28, 30 zeros behind it and that's a full profile likelihood ratio, which essentially approaches the random man excluded value as it should, when we compare that number I need you to think of the sets of three zeros repeated 20 plus times. And you're going to compare that with say 10,000 or 100,000 to make it easier, two sets of three zeros. The difference between those numbers is an astronomically big value for which the thousand, 10,000 and 100,000 shrink to a meaningless comparison."). UCPO reports that there is an elevated chance of false positives for LR under 3,000, which is in the "moderate support range" of the verbal scale. In Germany, the recommendation is that an LR less than a million be reported as inconclusive. John Buckleton et al., A diagnosis of the primary difference between EuroForMix and STRmix™, 69 J. Forensic Sci. 40 (2024).

In short, no evidence was provided to justify the use of the verbal scale at this hearing, and the evidence that was provided does not show an empirical justification for it and demonstrates a real risk of the verbal equivalent being misleading.

**12. All forms of DNA interpretation, including probabilistic genotyping, are constrained by fundamental principles of genetics and human judgment, and must be appropriately implemented.**

In conclusion, the interpretation of DNA mixtures is hard. It is harder when there is less DNA, when one contributor has given a very small proportion of the total DNA, or when there is

significant allelic overlap, as occurs with related individuals. These difficulties cannot be circumvented by PGS. Therefore, it is necessary for any type of PGS and any laboratory implementing that PGS to demonstrate, through a validation, where the system works well and when it stops working so well. Only by knowing that can appropriate limits on the use of PGS on different sample type be established. Because humans are integral to the operation of PGS, the utmost care must be taken to give them the most objective guidelines for their work and mitigate the effects of cognitive bias. Last, great care must be given to accurately report the likelihood ratio, which is hard to do, and which is not appropriate in all situations.

## G. There is limited evidence to support the foundational reliability of STRmix v2.5.11 and v2.8.0.

Below are factual findings that relate to each Olenowski factor and other related, relevant findings. The conclusions of law that stem from those findings are discussed in the Conclusions of Law section that follows. The following sections address the insufficient testing of STRmix, the unknown error rate, the almost total lack of independent peer-reviewed publications establishing its reliability, the lack of standards governing the performance of STRmix, and the general consensus of forensic DNA analysts that STRmix can only be used across a range of samples it has been shown to be able to reliably analyze. In addition to information that goes directly to the Olenowski factors, this section finds facts relevant to the reliability of STRmix, as opposed to PGS in particular: the lack of limits on its appropriate use set by its developers and the inappropriateness of the LR it reports for mixtures comprised of related contributors.

### 1. There has been limited testing of STRmix across the range of samples STRmix is used on and almost none of that testing is independent.

Although many tests of STRmix have been run, these tests are insufficient to demonstrate the reliability of STRmix. They are not independent, they do not address the versions used in this case, and they do not test STRmix across the full range of samples STRmix is used on.

**a. The published developmental validation of STRmix is not independent and is limited in its scope.**

There has been limited developmental validation of STRmix in general and in particular of the versions used by the State in this case. The one developmental validation proffered by the State, Developmental Validation Of STRmix, Expert Software For The Interpretation Of Forensic DNA Profiles, 23 FSI: Genetics 226 (2016) (S151/D1034), was written by the developers and purveyors of STRmix, including John Buckleton, Jo-Anne Bright, and Duncan Taylor. There are several problems with this validation.

First is the interest of the authors. As PCAST has explained, studies that seek to establish scientific validity "should be performed by or should include independent research groups not connected with the developers of the methods and with no stake in the outcome." PCAST Report at 78-81 (Pa 8). NIST, Human Factors, at 202 ("In addition to a lack of peer-reviewed publications with validation data available for independent review, many of the available published studies include members of a commercial product development team and are therefore not independent"). This study is not independent.

Second, neither of the versions of STRmix used in this case were tested in this study. Every version of STRmix is at least somewhat different from the one before—that is the point of creating new versions. Although the developers would hope that each version is an unalloyed improvement, there is no way to know whether the changes have introduced new issues. See also Martin Declaration, July 2024, at 4 ("To put it plainly, testing or reviewing one version of source code is insufficient to draw conclusions about other versions of source code; different versions of the code contain material differences in the source code. Source code and related materials of each version of a specific software must be reviewed separately.") (D-13). In his initial report, Dr. Buckleton asserts that "[t]he developmental validation process has been repeated for each

software version released." (D-21 at 1) These repeat validations, and their results, have not been publicly released anywhere and therefore provide no support to the State's burden of establishing any version of STRmix's reliability.

Third, the validation study does not provide enough information about STRmix's performance across a range of samples to actually assess its reliability. Although the developers of the program happily conclude in that article that "STRmix is suitable for its intended use," there is insufficient information given to support that conclusion. Developmental Validation at 238. No error rates are given for the samples concerned, discussed further below in subsection 2. The authors mention that, with more DNA, STRmix gives a "high LR for true contributions and a low LR for false contributors," but does not mention how often or high those false inclusive LRs are. Id. at 299. Each graph shows false inclusionary LRs for noncontributors under different circumstances, notably increasing as template amount decreases. But no detailed information is given about how and when these errors occur. More discussion of error rates is presented in subsection 2, infra.

Fourth, the developmental validation did not analyze samples similar to the ones presented in this case. There were only 31 samples tested in the developmental validation, which were amplified in triplicate. Id. at 229. The lowest amount contributed by a minor contributor was 9% for three samples. Ibid. Related contributors were not tested and degree of allele sharing was not otherwise addressed.

Fifth, no limitations for STRmix were established by the developmental validation. In other words, after seeing the performance of STRmix under these circumstances, the authors did not conclude that there are circumstances under which STRmix does not operate reliably. As explained above, the question for PGS is not simply whether it is reliable at all, but, if it is

reliable at all, across what range of DNA samples it is reliable. See Mixture Interpretation at 8 ("Reliability statements based on aggregate performance across many types of samples and many different probabilistic genotyping software (PGS) systems do not provide the information needed to judge the degree of reliability of the measurement and interpretation in a particular case of interest. We believe it is inappropriate to transfer any global reliability statements to a specific case because of the number of variables that affect DNA mixture interpretation."). The mere fact of lots of tests being run is insufficient to say that STRmix is foundationally valid in any or all circumstances. As Dr. Reich put it, there has been a lot of testing, but the essential question "is how broad and deep is that testing." (12T 115-5 to 8)

As set forth above in subsection F.8, the developmental validation study could support the reliability of analysis only of casework samples sufficiently similar to enough study samples that were reliably analyzed in that study. A piece of software that has been shown to reliably analyze a two-person mixture made of 400 picograms of DNA cannot be assumed to be able to reliably analyze a 6-person mixture made of 40 picograms of DNA. Many kinds of samples similar to such a complex sample would have to have been tested and reliably analyzed before such a claim could be made.

**b. Internal validation studies cannot compensate for insufficient developmental testing.**

The developmental validation is insufficient to demonstrate the reliability of STRmix v2.5.11 or v2.8.0. Its shortcomings cannot be remedied by reference to internal validation studies for six reasons.

First, as found above, subsection F.8.d, internal validations are not substitutes for developmental validation, which is a mandatory prerequisite for establishing the validity of any forensic method, including STRmix.

Second, these internal validation studies are not independent. ESR does much of the work for them. (2T 76-11 to 20; 6T 80-9 to 81-7; 8T 88-6 to 90-7) Moreover, those who run them are laboratories who have already committed to bringing STRmix online and therefore have a vested interest in STRmix being found to be reliable. For this reason, internal validations are often not focused on reliability, instead focusing on proving utility, as is demonstrated by the lack of error rates provided and limits established by the Bode and NJSP internal validations, discussed further below. See also Mixture Interpretation Supplement at 48 ("Historically, forensic DNA laboratories have conducted mixture studies during their internal validation experiments with emphasis on robustness (does the test produce a result?) and detectability (can minor alleles in a two-person mixture with multiple mixture ratios be detected?) rather than reliability (was interpretation of the mix data accurate and consistent if repeated?)").

Third, these internal validation studies are generally not peer-reviewed or even publicly available. Mixture Interpretation at 95. That most of these studies are inaccessible other than through litigation undermines their reliability and helpfulness because "[d]issemination is a critical part of the scientific process because it exposes our work to peer review and allows scientists to build upon the contributions of others. A study isn't complete until it's been published." Mixture Interpretation at 26.

Fourth, internal validation studies that are publicly available almost often do not contain enough information for anyone to independently review the results. Id. at 16 ("There is a growing body of scientific literature on DNA mixture interpretation. However, supporting data provided in the scientific literature is not always sufficiently detailed for an independent review of claims."); id. at 95 ("Numerous PGS studies have been published in peer-reviewed journals . . . . [M]any of the publications did not contain information and details to aid independent review

of the data in them."); 14T 149-4 to 7 ("A lot of the actual underlying data are not published. So the analyses can't be independently performed for a lot of these studies.").

Fifth, they are all substantively similar, so they do not provide sufficient challenging testing of STRmix. 12T 68-16 to 70-13 ("[I]f all of the tests of the method are being done in the same way by the same source, you're going to naturally get the same kind of results, and you're not going to do a thorough and critical evaluation of all of the possibilities that you would want to do before you use this method on the very significant samples that come into crime labs."); 16T 76-4 to 15 ("[I]f you simply do the same experiment over and over, if everybody does for example, you know a series of two, three and four person mixtures that have the same mixture ratios, you're not doing different experiments, you're doing the same experiment and basically testing reproducibility which by the way, is an important thing but it's not the same thing as saying well, the more samples and the more laboratories that look at the same samples the more confident we are that we've covered the sample space. That's just, that just doesn't, that's logically not connected.")

Sixth, because there is no objectively correct LR, these studies are limited in what information they can relay about STRmix's reliability. With the samples assessed, often it's known if the LR should be above or below one, but it's unknown what LR the model, if it were performing well and being used correctly, "should" be producing. See also 14T 179-13 to 23 ("All of the publications, the published articles about STRmix are, the vast majority of them are end to end system tests, exactly this black box. Where they say here's the included data. We don't know what the LR is supposed to be. Oftentimes we know if it's supposed to be above one or below one, but we don't know exactly what it's supposed to be.").

The "31 Laboratory study," which was often suggested as proof of reliability at the hearing, (7T 95-1 to 5, 99-1 to 101-12; 10T 68-16 to 18; 12T 138-13 to 140-18; 13T 131-2 to 15) suffers from all of the flaws discussed above. Jo-Anne Bright et al., Internal Validation of STRmix – A multi laboratory response to PCAST, 34 FSI: Genetics 11 (2018) (S146/D1041). Multiple STRmix developers are authors, including Dr. Buckleton. The underlying data are not published. (14T 152-6) It examined the use of only STRmix v2.5.02. Id. at 12. The studies used only simulated non-donors and the Caucasian allele frequencies. Ibid. The study provides a table of "large inclusionary LRs for false contributors and percentage of overlapping alleles," but no error rates. Id. at 15. It notes that "[n]on-contributors that share most of their alleles with the mixture's donors can typically still be excluded because the peak heights make their inclusion unlikely." Id. at 18. In short, it's an article written by people who developed STRmix and who already use it that notes that STRmix produces both false positive and false negative errors, but does not produce enough information to assess when and how often those errors occur or for independent evaluation of the data.

But the data that is in the article shows that it is insufficient to validate the use of STRmix across all sample types because there is not enough data and not enough was tested. There is evidence that STRmix falsely includes people who share many alleles with the real contributor, as would be expected based on how STRmix works, but no error rate presented across degree of allele sharing and other factors that make a sample more complex to help establish under where the limit of reliable analysis is. The authors concluded that "the greater the allele sharing, the less the power there is to discriminate a true contributor from a non-contributor," meaning there will be more false inclusions, but those are not sufficiently quantified. A Multi-Laboratory Response at 20. Moreover, only one of the 31 laboratories tested any mixtures with related contributors and

that was just one mixture, so the study provides almost no information about the reliability of STRmix when analyzing mixtures comprised of related people. Id. at 13. Further, it is impossible to specifically discern the mixture proportions run through STRmix or the picograms based on the published article. There is no evidence that sample types similar to the ones run in this case were tested.

In short, there has barely been any independent testing of STRmix, and none of the testing that has occurred is sufficient to address how each version of STRmix performs across the range of samples it is used on.

**2. There is almost no information about STRmix's error rate across the range of samples STRmix is used on.**

There is no rate of error presented for STRmix—not in the developmental validation and not specifically for versions 2.5.11 and 2.8.0. Because of this total lack of error rates, including a lack of error rates across samples of different complexity, there is no evidence of if or when STRmix is reliable. The lack of information is about error rates is concerning for five reasons: there is literally no false positive or false negative rates to consider; there is no evidence for that error rate as samples get more complex, which is known to increase error rates; we do not know that sufficient samples of each kind have been tested in order to reach a reliable error rate; and we know that false positive and false negative occurs rates do in fact occur.

**a. No false positive or false negative rates have been provided.**

First, no error rates have been provided by the State. There are no false positive rates—how often noncontributors are incorrectly assigned inclusionary LRs—or false negative rates—how often contributors are incorrectly assigned exclusionary LRs—in the developmental validation or either internal validation presented in this case. We know that STRmix does, and forever will, have a rate of error. As explained in a paper that Dr. Buckleton coauthored, error is

inevitable: "False inclusions and false exclusions may occur as result of a combination of specific software, multiplex and operator factors." Developmental Validation at 231. No matter how good STRmix is and how well it is implemented in a laboratory, there will be false positives and false negatives produced by it: "There are no modeling improvements that could ever be made which will eliminate all LRs that falsely favour inclusion. This is because the phenomenon causing these results is not a modeling phenomenon, but is due to the available biological data." Id. at 232. In other words, as Dr. Buckleton candidly admits, the software is going to make mistakes. What he won't admit is how often such mistakes are made and under what conditions. As a result of the lack of independent testing and the refusal of STRmix to release this information, there is almost no information availability about STRmix's error rate.

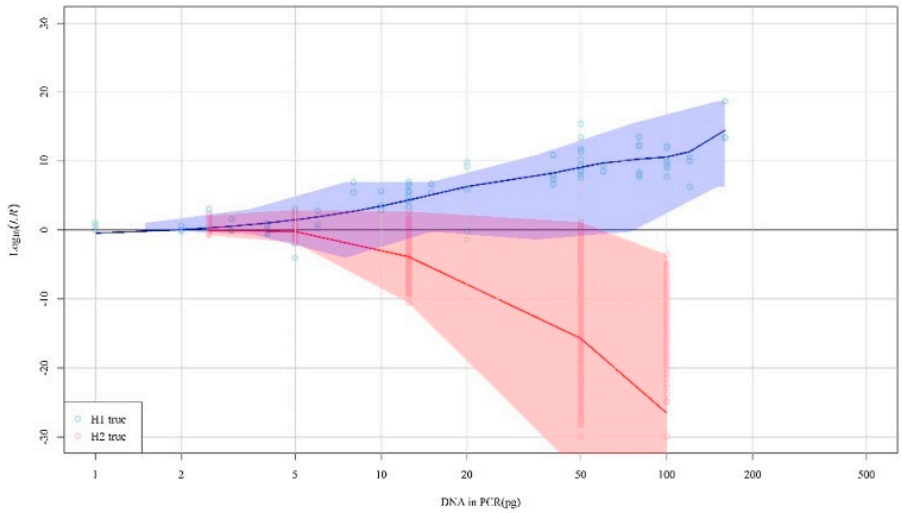
**b. There is no evidence of what the error rates would be across mixtures with various features.**

Second, no error rate has been provided for different kinds of samples. We know that STRmix will predictably have more errors in more complex samples. In other words, the error rate is not constant across samples of various complexity. Mr. Inman explained that defining what kind of samples produce an error rate too high for the reliable use of STRmix is the key concern with its use: "The probabilistic genotyping software can't do any better than the information that it's fed and so we simply need to understand where along this process the uncertainty is great enough that whatever the software does there's going to be too much uncertainty for us to believe that the inference we're going to make is basically error free. It won't be error free ever but just, there's too much uncertainty for us to really decide whether this person is a contributor or not so." (16T 163-1 to 9)

Dr. Buckleton agrees that STRmix will produce false positives at higher rates for more complex samples: "A DNA profile with limited information will not allow high discrimination

and will typically give inclusionary support for many genotypes. Thus, as the information in a profile decreases, more adventitious inclusionary support is observed. . . . Adventitious matches may also occur when comparing mixtures with individuals who are closely related to contributors or due to inbreeding effects." John S. Buckleton et al., Are Low Lrs Reliable?, 49 Forensic Sc. Int'l: Genetics at 2 (D-1063).

The limited data available shows false positive and false negative errors. For instance, the developmental validation shows that as the amount of DNA contributed decreases, the rate of false inclusions increases, but no rates or actual numbers are provided. In the chart below, taken from that study, all red above the 0 axis would be a false inclusion and all the blue below the 0 axis would be a false exclusion:




Developmental Validation at 231. (S-151/D-1034)

As is apparent, more errors occur as there is less DNA in the sample. That makes sense because the less DNA, the more complex the sample is. But the fact that there will be more errors as DNA samples get more complex does not mean that the information about how often it leads to errors in the LR does not need to be demonstrated.

Providing this error rate is quite possible. Dr. Buckleton, in fact, explicitly recognizes that error rates vary across the kinds of samples tested and explicitly refuses to give those rates. As he said at the hearing, "[w]e can actually measure it for you. So for instance, we could have done one or a few of these samples and empirically measured the false inclusion rate but that's not being done." (7T 108-2 to 10) NIST also recognizes that it is possible and advisable to produce this kind of information: "Studies can, however, estimate the percentage of time the LRs are . . . providing adventitious exclusionary or inclusionary support. Sometimes, data may be favorable to H1 even when H2 is true. This happens not just due to adventitious matches from a high degree of allele sharing among contributors in DNA mixtures, but also due to limitations of models, particularly with low LRs." Mixture Interpretation at 77 (internal citation omitted).

The lack of error rates seems to entirely stem from Dr. Buckleton's resistance to giving such rates. As he writes in one article, data from a STRmix studies "should not be used to assign a false inclusion rate (we prefer the term adventitious match rate) for STRmix™ or any other software. The adventitious match rate is determined almost completely by the profile (and more specifically on a per contributor basis, within the profile) and to a very much lesser extent by the software. . . . Hence the reported adventitious match rate would be influenced by the exact mix of samples investigated." John Buckleton et al., Response to: Commentary on: Bright et al. (2018) Internal validation of STRmix™ – A multi laboratory response to PCAST, 44 FSI: Genetics 1, 4 (2020) (D-1043). But which samples produce errors at what rate is exactly the issue.

Everyone, even Dr. Buckleton, knows that STRmix will yield errors and that the rate of error will increase in different circumstances, depending on the kind of sample analyzed. Without knowing the rate of error in these varying circumstances, how can this Court determine what

kinds of samples have too high of an error rate to be used in criminal court? The inevitability of error makes the unfettered use of STRmix more concerning, not less.

**c. Even if error rates have been provided, there is no evidence that the tests have been run on large enough samples such that those rates are known to be representative.**

Third, in order to trust that an error rate is a reliable reflection of the actual chance of error, the testing to establish that error rate on different kinds of samples must be sufficient. It is well established that a large enough sample size is necessary for the results of any testing to be reliable. The idiom a "sample size of 1" is used to explain that one person's opinion is not enough to draw a conclusion. That one is not enough is beyond doubt. The question is how large does a sample size need to be in order to have what degree of confidence that it's the conclusion to a test?

There are statistical methods to determine how many samples must be run to have a particular degree of confidence in the outcomes of a study. See e.g., Lawrence Joseph et al., Bayesian And Mixed Bayesian/Likelihood Criteria For Sample Size Determination, 16 Statistics in Medicine 769, 769 (1997) (discussing various methods for determining "the optimal number of experimental units," which is "an essential step" in study design) (D-1206) No evidence that those methods have been employed to determine whether there has been sufficient testing by STRmix of different kinds of DNA samples to have confidence in the results have been provided, although such statistical methods do exist. (16T 101-24 to 102-7)

**d. There is evidence of significant rates of error.**

The limited information available shows that error rates when STRmix is used in a laboratory cannot be assumed to be insignificant. The FBI's internal validation study found a 6.1% false negative rate overall, a .1% false positive rate overall, and a 1.7% false positive rate when an additional contributor was incorrectly assumed to be included. Moretti et al., Internal

Validation Of STRmix For The Interpretation Of Single Source And Mixed DNA Profiles, 126 FSI: Genetics 126, 138, 141 (2017) (D-1037). Although the State attempts to cast the false positive rate as the only rate that matters, as even its experts agreed, a high false negative rate matters as well: a third-party guilt defense can rise and fall on a false exclusion. (7T 118-3 to 20) If a possible alternate suspect was incorrectly excluded as a contributor by STRmix—a false negative—a defendant would have a significantly harder time convicting a jury that that alternate suspect was the true perpetrator.

Moreover, the "overall" false positive rate does not reflect the false positive rate of any particular kind of sample: the rate of false positives in the FBI's study increased, predictably, when the complexity of the mixtures increased through lower template or higher numbers of contributors. Id. at 131-133. Four-person mixtures, for instance, had a false-positive rate of 4%. Id. at 132.

Moreover, this validation suffers from the same defects as most of the other studies—it is co-authored by many STRmix developers, including Dr. Buckleton, it was produced by a laboratory that already had purchased and wanted to use STRmix, its underlying data are not provided, and what is in the article is insufficient to determine the limits of the STRmix's reliability across different sample types.

One of the few studies that no STRmix developer seems to have been involved in also found significant rates of error. This study found that across all types of mixtures—with varying complexity—there was a 1.35% false positive rate. Sarah Riman et al., Examining performance and likelihood ratios from two likelihood ratio systems using the PROVEDIt dataset, 16 PLoS One 9, 10 (2021) (D-1082). 2.07% of true contributors were incorrectly excluded. Ibid. The false positive rate for four-person mixtures was 2.76%. Id. at 11.

In short, STRmix generates false positives and false negatives frequently and with increasing frequency as samples get more complex. The limited information available about these rates underscores the need for more information about them before STRmix can be deemed reliable.

**e. Trends do not sufficiently describe the error rate of STRmix across sample types.**

A number of other facts about STRmix have been gestured at to excuse the lack of error rate. These facts, while potentially informative, do not substitute for error rates of STRmix across sample types.

First, trends are not error rates and cannot substitute for error rates. Two trends have been pointed at by the State to attempt to dismiss the need for error rates: that STRmix seems to generally include real donors and exclude real donors, and that LRs trend towards 1 when profiles are more complex, meaning false positives would generally be low. As to the first, Dr. Inman explained, the mere fact that STRmix often produces correctly inclusionary or exclusionary results on ground-truth samples is insufficient to extrapolate that it would do so with casework samples unless there is enough testing of varied enough genotypes and across complex enough samples. In case work "the real question then becomes well, what if you have the wrong person which way would they trend. And that depends entirely on a combination of what the profile has and what your person of interest has true or not true. A trend is a trend. It's not every case." (16T 171-10 to 15) As NIST explains, "[d]emonstrating a trend though is not the same as showing the reliability of a specific LR in DNA mixtures with various levels of complexity." Mixture Interpretation at 80. That is the point of testing the gamut of samples rigorously during a validation; to make sure there has been enough testing of enough different kinds of samples, not just to point at a trend that has come up in limited testing and rely on it.

As to the second, although it is true that LRs trend towards 1 when profiles are more complex and that therefore false positive LRs are more frequently going to be low does not tell us enough about how often errors occur and when they occur and whether a specific sample is more likely to generate an erroneous result. Although it is good, if true, that errors with low LRs will occur more frequently than errors with high LRs, they are still errors. And we need to know how often these errors are made, especially since STRmix reports all LRs, and the laboratory is left with the discretion to report or not report them. For instance, NJSP does not report LRs under 1000, due to this risk of false positives. (D-53 at 8-6) Bode does. (Bode Technology Case File (D-66) at 4) ESR has not set parameters at which LRs should not be reported. Moreover, "[h]igh" and "low" are qualitive judgments. They are not specific enough to understand the risk of error across sample types, not just the subjective assessment of the error's magnitude

Nor does Turing's Rule, a mathematical relationship that explains a trend, substitute for information about error rates. The only error rate produced by the State is Turing's Rule. According to Dr. Buckleton, "STRmix™ will produce an LR greater than x from about 1 in x false donors." (Pa 7 at 25) This is insufficient for two reasons.

First, it assumes the very thing that needs to be proven: that the model "is performing well." A bad model or a model being misused by an analyst is not going to give a false positive LR of 100 only 1 in 100 times. That frequency of error is the best-case scenario for a probabilistic system, not the rate of error for every probabilistic system in every circumstance. The State has not demonstrated that this best-case scenario holds in STRmix v2.5.11 or v2.8.0 and across the range of different kinds of samples.

Second, Turing's Rule does not speak to the overall rate at which STRmix will err and provide an inclusionary LR for a non-contributor or an exclusionary LR for a contributor.

Instead, it speaks to how often errors of certain magnitudes will be made. For the reasons explained above, that general trend is insufficient to demonstrate the risk of error.

**3. No standards substantively govern STRmix's performance.**

Many standards and guidelines documents were discussed at the hearing. Most are guidelines, which are not mandatory for a laboratory to follow. (1T 29-1 to 7; 6T 73-4 to 12). The available standards and guidelines do create requirements, or at least suggestions, that developmental and internal validation studies determine the limits of STRmix's reliable use across factors of various complexity, which, as discussed further in subsection H, Bode and NJSP have not done. These standards, however, do not set any thresholds for performance. They do not create substantive requirements on STRmix's use to render it reliable, but rather operate as a checklist for completion of certain tasks, not the quality of the output of the laboratory. See also Mixture Interpretation at 28 (discussing the concerns of the forensic DNA community that there are few "'standards' with 'teeth' (impact or real influence), rather than general guidelines").

As thoroughly discussed in subsection D and E, IEEE 1012, the leading software engineering standard that was discussed, imposes requirements and STRmix was shown by all accounts to fail to abide by the standard.

**a. FBI Quality Assurance Standards**

The FBI QAS are standards, which means they must be followed by laboratories accredited to the FBI QAS. But these standards are not specific for PGS and therefore do not have any specific requirements for the reliable use of PGS. (1T 39-22 to 40-4) The QAS does require laboratories to perform internal validations, to use those validations to define guidelines for mixture interpretation, and to have written analytical procedures. FBI QAS (S5/D204), Standard 8.3 and 9. Both Bode and NJSP are accredited to ANSI/ASB 18. (2T 66-21 to 67-2; 8T 18-18 to 19-21)

**b. Scientific Working Group on DNA Analysis Methods Guidelines**

The Scientific Working Group on DNA Analysis Methods, Guidelines for the Validation of Probabilistic Genotyping Systems 2 (June 2015) (D-202) are guidelines, not binding standards. (6T 73-2 to 10) As the experts discussed, these guidelines are influential on forensic DNA laboratories. (6T 73-16 to 19; 10T 97-11 to 19)

As discussed above, SWGDAM guidelines suggest that each laboratory conduct its own internal validation, that the sample types "test the system's capabilities and identify its limitations," and pay special attention to "various contributor ratios," "various total template quantities," and "[s]haring of alleles among contributions." Ibid. Although SWGDAM recommends that such testing be done, it does not establish minimum results of this testing to demonstrate reliability. Neither the developmental validation nor either internal validation at issue in this case explored allele sharing as SWGDAM suggests. SWGDAM does not establish minimum results of this testing to demonstrate reliability.

**c. ANSI/ASB Standard 18 for the Validation of Probabilistic Genotyping Systems.**

ANSI/ASB is a standard, meaning it is mandatory for laboratories accredited to this standard to meet its requirements. (6T 76-5 to 22; 10T 97-20 to 23) As explained above, subsection F.8.d.i, it too requires developmental validations and internal validations that span the totality of the casework done by a laboratory. ANSI/ASB does not establish minimum results of this testing to demonstrate reliability. Both Bode and NJSP are accredited to ANSI/ASB 18. (2T 66-21 to 67-2; 8T 18-18 to 19-21).

**d. The International Society for Forensic Genetics Recommendations on the Validation of Software Programs Performing Biostatistical Calculations for Forensic Genetics Applications.**

The ISFG recommendations are neither a standard nor a guideline in the United States. (10T 64-6 to 18) The recommendations do suggest that IEEE-1012 applies to PGS. (S131/D203

at 2). The ISFG also recommends internal validation studies to analyze samples representative of casework in terms of, "as a minimum, the number of contributors, mixture ratios of contributors and DNA template amounts." Id. at 3. It also recommends laboratories to test "(1) true donors and non-donors and/or (2) related and unrelated individuals across a range of situations that span or exceed the complexity of the cases likely to be encountered in casework." Id. at 5. Although it recommends thorough testing, ISFG does not establish minimum results of this testing to demonstrate reliability. Neither the developmental validation nor either internal validation at issue in this case explored related individuals as ISFG suggests.

**e. Forensic Science Regulator of the United Kingdom, Software Validation for DNA Mixture Interpretation**

These are neither a standard nor a guideline in the United States. (14T 161-18 to 162-12) Although it recommends thorough testing, FSR does not establish minimum results of this testing to demonstrate reliability.

**f. ISO 17025**

ISO 17025 is a standard, not a guideline, but it is not particular to forensic laboratories, to DNA analysis, or to PGS. (1T 26-3 to 7; 8T 19-3 to 15)

**g. Audits of accreditation standards do not operate to ensure reliability.**

Insofar as any internal validation studies are audited by accreditors, these are largely superficial and do not operate as a check on the reliable use of STRmix. Even as to the FBI QAS or the ANSI/ASB 18, two relevant standards, there is no need for a laboratory to show that the results of its testing demonstrate an ability to reliable use STRmix, just that the testing occurred. As NIST explains, "an audit of a validation study under the FBI QAS requirements involves a 'yes', 'no', or 'N/A (not applicable)' response to a series of questions, such as 'have internal

validation studies included, as applicable: precision and accuracy studies? sensitivity and stochastic studies? mixture studies?...'" Id. at 71-72. "However, in an audit, there is no mechanism to assess performance reliability, only whether or not a specified type of study has been conducted and documented in records retained by the laboratory and made available to the auditor.' Id. at 71.

Put another way, both the FBI QAS validation requirements and the SWGDAM validation guidelines are "task-driven rather than performance-based. In other words, the requirements and guidelines may be treated by some as a checklist of studies that need to be completed to satisfy requirements rather than a demonstrated performance of the accuracy or reliability of results obtained using the method." Supplemental Document to Mixture Interpretation at 42. But this approach does not further the goal of reliability: "Performance-based approaches are preferable over checklists of validation studies conducted because they can provide information on the limitations of the method." Id. at 48. Taking a driver's test does not help assure your fitness to be on the roads unless you passed it. But without any standards for passing or failing, all these standards do is ensure the test was taken, not that any of the laboratories are safe on the road.

Dr. Reich echoes NIST's explanation of the minimal impact of the accreditation standards and related audits. He explained that there is "a wish that the laboratory standards in the quality assurance document from the FBI or the International Standards Organization standard 17025 are somehow related to the scientific integrity and scientific standards in the laboratory. But they're not." (12T 102-9 to 14) Rather, "the standards are an administrative requirement that the laboratory has certain documents and does [a] certain amount of documentation. But it is not a standard of care as would be applied let's say in medicine or veterinary medicine and it doesn't

describe the scientific level of the laboratory. It describes the administrative and documentation level of the laboratory and those are not necessarily linked." (12T 102-14 to 22) See also 12T 103-3 to 9 ("The standards describe what the lab has to have in place. It doesn't tell them whether they're within the norm of the field. It doesn't tell them whether they've implemented it as accurately as they could. That's not what the standards are doing and that's not how the laboratory is audited against those standards."). As multiple laboratories losing their accreditation shows, accreditation does not ensure reliability. (16T 160-3 to 4)

**h. Analyst intuition is not a replacement for effective standards.**

The designers of STRmix and the laboratories implementing it rely on analyst intuition to discern errors and to use STRmix appropriately. Subsection F.10. This reliance on subjective judgment does not serve as an adequate substitute for real standards. Subjective judgments are "more susceptible to human error, bias, and performance variability across examiners." PCAST Report at 47. That is why it is essential to have clear, objective standards guiding discretion. The need for standards is especially acute given how unfamiliar the DNA analysts are with STRmix's diagnostics and their limited knowledge of how it actually works. None of the standards discussed above create objective guidelines to channel analyst discretion, but instead leave it to each laboratory. STRmix has also refused to set objective boundaries on the use of STRmix. As discussed further in subsections H.2 and H.4, neither Bode nor NJSP have objective SOPs that sufficiently channel their analysts' discretion or that take any steps to mitigate the impact of cognitive bias on their case work.

**4. There are almost no independent, peer-reviewed publications that assess STRmix's reliability.**

Much has been written about STRmix, but when those publications are examined carefully, it becomes apparent that almost none of them are independent, peer-reviewed

publications. Instead, they are almost entirely written by STRmix's developers or people who work in forensic laboratories. Almost all of them are published in journals that Drs. Buckleton and Coble are or have been on the board of, including one journal with such a high self-citation rate that it was delisted from the journal rankings for a time. Almost all of the publications are in forensic science journals and none of them are published in engineering journals.

Out of the 101 articles provided by Dr. Buckleton to support the reliability of STRmix (S-129a), only 16 are written by people totally unaffiliated with STRmix's development. In other words, 84% of these articles are written at least in part by people with a vested reputational and professional interest in STRmix's success.

Of the 16 articles not written by STRmix developers, ten have nothing to do specifically with the reliability of STRmix. Of the six remaining, two are written by law enforcement that has purchased STRmix and uses it for casework, which also gives those agencies a vested interest in STRmix being found reliable. Duke and Myers, Systematic Evaluation Of STRmix Performance On Degraded DNA Profile Data, 44 FSI: Genetics (2020) (D-1092); Noel et al., STRmix™ Put To The Test: 300 000 Non-Contributor Profiles Compared To Four-Contributor DNA Mixtures And The Impact Of Replicates, 41 FSI: Genetics 24 (2019) (D-1089); Greenspoon et al., A Tale of Two, supra. Of the four remaining, one is not a peer-reviewed publication, but a master's thesis. Diana Orozco, TrueAllele and STRmix: A Comparison of Two Probabilistic Genotyping Software Programs in Forensic DNA Profile Analysis, University of California, Davis (2023). One does not test STRmix for the first time, but uses data from another study. M. McCarthy-et al., Low Lrs Obtained From DNA Mixtures: On Calibration And Discrimination Performance Of Probabilistic Genotyping Software, 73 FSI: Genetics (2024). That study does note that data from

all four PGS reviewed, including STRmix, produce false inclusions and inclusions when the LR is less than 1000.

One contains almost no information about STRmix's reliability, explaining only the general circumstances under which STRmix performs better and worse:

> As DNA amount per contributor decreases and the complexity of DNA profiles increases (due to an increase in NOCs), Log10(LR) produced from known donors or non-donors trend downwards or upwards towards 0, respectively. This demonstrates, as expected, that when relevant information in the mixture profile decreases, the degree of discrimination between true and non-true contributors using Log10(LR) decreases as well
>
> Sarah Riman et al., Exploring DNA Interpretation Software Using The Provedit Dataset, 7 FSI: Genetics Supplement Series 724 (2019) (D-1093).

Therefore, there is only one peer-reviewed article written by independent researchers about STRmix's reliability: Sarah Riman et al., Examining Performance And Likelihood Ratios From Two Likelihood Ratio Systems Using The Provedit Dataset, 16 PLoS One 9 (2021) (D-1096). It was written about STRmix v2.6.0, which is not the version used in this case. Id. at 2. It found the 1.35% false positive rate and the 2.05% false negative rate for STRmix v2.6.0 for all the samples together. The false positive rate over doubled for four-contributor samples: in those circumstances, the false positive rate was 2.76%. It noted that a different PGS gave higher LRs for correct inclusions than STRmix. Id. at 16.

Moreover, articles about STRmix are almost always published in journals that people invested in STRmix's acceptance are on the board of which cite their own work to make their assertions seem broadly supported. Dr. Coble is on the board of the Journal of Forensic Science and of Forensic Science International: Genetics ("FSI: Genetics") and WIRES Forensic Science. (10T 28-4 to 7) Dr. Buckleton has been on the editorial board of the Forensic Science International: Genetics since it was first released. (7T 151-18 to 153-20) Most of Dr. Buckleton's articles are published in FSI: Genetics. (7T 152-15 to 16) Duncan Taylor, one of the developers

of STRmix, who often publishes about STRmix, has the highest self-citation rate of any forensic science research; Dr, Buckleton has the ninth most self-citations, with 30% of his work citing itself. (11T 73-4 to 75-12) FSI Genetics was removed from the journal impact factor in 2019 because of the extraordinarily high self-citation rate. (11T 75-20 to 76-4) In 2014 and 2015 FSI Genetics had self-citation rates of 55% and 61%, respectively. (11T 76-24 to 77-21)

In sum, mostly STRmix is written about by people who are very invested in STRmix in journals that circularly rely on prior articles by the same people to continually justify their assertions. As found in subsection D.3 internal validation studies are neither independent nor peer-reviewed. And as found in subsection 1.b, none of these studies, published or not, provide sufficient information for independent examination of its conclusions.

**5. The unlimited use of STRmix is not accepted in the field of forensic DNA analysis.**

It is true that many laboratories use STRmix. It is unclear if the majority of laboratories in the United States do, but if so, it is just barely the majority. 11T 84-8 to 11 (There are around 212 laboratories that do forensic DNA testing in the US and only "somewhere on the order of 130, 140" have brought on STRmix or TrueAllele."). But with the exception of Dr. Buckleton's testimony at the hearing, no DNA analyst has ever suggested that it is appropriate to use STRmix on every kind of mixture type.

As explained above, every single other DNA expert witness, both for the State and the defense, agreed that there need to be limits established within which STRmix's reliable performance has been demonstrated. NIST agrees: "[G]uidance on how to identify limitations with these PGS systems is sparse or non-existent. Knowing that complex models may at some point begin to produce unrealistic results, the identification of these points of caution or concern need to be defined by assessing performance from whether same-source and different-source sample profiles are appropriately 'included' or 'excluded' in situations with types of samples

similar to the case in question.". Mixture Intepretation Supplement at 50 (internal citations and quotation marks omitted).

William Thompson, the special master in United States v. Lewis, discussed further below, agrees. As Dr. Thompson explains, "The validation research published to this point has done an excellent job of demonstrating how well the programs work under a broad range of circumstances. The research has been less successful at establishing the limits of reliability." Thompson, Uncertainty in Probabilistic Genotyping, at 13. "Finding the limits will be important to courts evaluating the admissibility of PG results in cases like this one where labs may be working near or even beyond those limits." Ibid. See also Tim Stelloh and Brenda Breslaur, Previously Unusable DNA Sample Now Evidence in the Quadruple Murder Trial of N.J. Uncle, NBC News (Dec. 27, 2024), https://www.nbcnews.com/news/us-news/paul-caneiro-murder-trial-dna-rcna185278 (quoting Dr. Thompson as saying that "STRmix works well if it's used under the conditions it has been tested for. But, he said, labs can run into trouble when they analyze complex types of samples that haven't been validated, especially if they're relying on tiny amounts of DNA. . . . 'There's a long history of people getting enamored of the technology and taking it a little bit too far and not quite understanding what they're doing. And I think that could definitely happen with regard to probabilistic genotyping. Probably it already has happened.'").

In short, some non-negligible portion of US forensic DNA laboratories use STRmix. That means forensic DNA analysts find it to be helpful and hopefully they believe it is reliable as well. But with the exception of Dr. Buckleton, even the staunchest supporters believe STRmix cannot be used for every kind of sample and without a demonstration of reliability as to all of the kinds of samples a laboratory wants to analyze. "General acceptance" is not a blanket term. STRmix, at

best, is generally accepted for use over the kinds of samples laboratories have demonstrated they can reliably analyze.

**6. No limits have been identified for the reliable use of STRmix by the developers.**

Despite the almost universal recognition in the forensic DNA community that STRmix can only be used within the limits that its reliability has been established, no limits about been identified by its developers.

**7. Humans are a necessary part of PGS which means proficiency and cognitive bias impact the ability to properly use it.**

Found in subsection F.10, human analysts are an essential part of the reliable use of STRmix. They must use it properly and be able to discern any errors that emerge. No evidence has been presented by STRmix or the State that analysts are proficient at using STRmix at all, let alone on complex mixtures that resemble casework.

## H. Even assuming the foundational reliability of STRmix, the use of STRmix in this case has not been demonstrated to be reliable.

Even if STRmix is foundationally valid, it may or may not be reliably used in any given laboratory. In this case, there is evidence that Bode and NJSP did not use STRmix in a reliable manner.

As a threshold matter, neither laboratory established the boundaries within which it is able to reliably use STRmix. Both Bode and NJSP provided internal validation summary documents with insufficient information to assess the performance of STRmix across sample types. Without demonstrating what each laboratory can analyze reliably, the laboratories cannot demonstrate that any of the casework samples have been reliably analyzed.

In the alternative, even if the samples that were tested in those summaries were taken to be the limits of what can be reliably analyzed—which should not be the case because there is insufficient evidence of STRmix's performance in those tests, merely that the tests were taken—the analyses in this case fall outside of those limits. In particular:

- Neither of the laboratories has demonstrated, through their internal validation, that they can reliably analyze samples with related contributors.
- Bode analyzed samples with a lower percentage of DNA contributed by the minor contributor that it tested in its validation study.
- Bode analyzed samples with a lower amount of DNA contributed by the minor contributor that it tested in its validation study.
- Bode analyzed samples containing first-degree relatives, contrary to its SOP.
- NJSP reported a "source attribution" for a contributor after using STRmix, in violation of the standards of the forensic DNA community.

Each of the validation studies and the relationship of the casework samples to the samples tested in these studies is exampled below.

**1. Bode's internal validation.**

Bode conducted its validation study with significant ESR involvement. (2T 76-11 to 20) Four people spent "at least" hundreds of hours doing Bode's validation, exclusive of ESR time. (3T 107-8 to 21) Bode gets paid per sample it analyzes through STRmix. (5T 98-12 to 99-5) Below the types of samples that were tested, as set forth in the summary, are reviewed.

**a. In its validation study, Bode did not attempt to test any sample of similar complexity to the samples it analyzed in this case.**

In order to run its sensitivity and specificity studies—testing STRmix's ability to exclude non-contributors and include contributors—Bode created "known template amounts" to run through STRmix in the validation study. (2T 79-7 to 9, 99-15 to 24, 109-6 to 11, 112-19 to 21) Only 55 unique mixtures were created. (D-60 at 4, Table 2a) Bode created varying mixture

proportions, which ranged from mixtures in which two contributors gave equal amounts of DNA to mixtures in which the minor gave around 5% of DNA. (D-60 at 4, Table 2a) The amount of DNA contributed by any contributor ranged from 400 picograms to 25 picograms. (D-60 at 4, Table 2a) Bode tested mixtures that contained two to four people's DNA. (D-60 at 4, Table 2a)

**i. In its validation study, Bode did not test any sample in which the minor contributor contributed fewer than 25 picograms of DNA.**

In its validation study, Bode did not test any sample in which the minor contributor contributed fewer than 25 picograms of DNA. (2T 82-24 to 25; 5T 95-5 to 7) Only 11 samples contained a minor of 25 picograms. (D-60 at 4, Table 2a) With duplicates, 22 samples total were run where the minor contributed 25 picograms (but only 11 unique mixtures). (D-60 at 4, Table 2a) One of those 25 picogram samples was removed from analysis. (D-60 at 4, Table 2b) The next lowest amount of DNA tested was 50 picograms. (D-60 at 4, Table 2a)

**ii. In its validation study, Bode did not test any sample in which the minor contributor contributed less than 5% of the total mixture.**

In its validation study, Bode not intentionally create samples to test that had less than approximately 5% of a minor contributor. (3T 109-23 to 110-4) Ten samples had a minor of around 5%, which came from two and four person mixtures. (D-60 at 4, Table 2a) With duplicates, 20 mixtures were created where the minor contributed about 5% of the mixture. (D-60 at 4, Table 2a) Of those 20, six were removed from the study for various reasons, so there were only 14 samples with 5% from a minor contributor. (D-60 at 4, Table 2b)

**iii. Bode's validation study reveals significant discrepancies between the mixture proportions deliberately created by Bode and STRmix's estimation of those proportions.**

Naughton created a memorandum that asserted that Bode did test samples below 5% in its internal validation study. One of the samples in that memorandum was a mixture that Bode intended to be 5% and that STRmix interpreted at 3%. (2T 118-24 to 119-1) Therefore, this

sample does not support the claim that Bode validated a sample that the minor gave less than 5% of the DNA to. If anything, it supports the claim that STRmix cannot reliably analyze samples which contain less than 5% DNA from a minor contributor.

The other two samples which are referenced in the letter do not further this claim either. TRN1777-0961 and TRN 177-0962, which are referenced in the letter and described at the hearing, are challenge samples made through deliberate contact with an item at the laboratory. (3T 50-1 to 53-13) That means the ground truth of who contributed what to the sample is unknown, although there is some idea of who at least some of the contributors should be. (3T 104-3 to 5) Therefore it is impossible to know if STRmix did a good job analyzing those samples. But more importantly, STRmix did not generate an LR for the minor contributor in either sample. Rather, the STRmix output says "no conclusions on low contributor," and annotated next to it is "manual interpretation results." (D-27 at 2) In other words, STRmix returned no results for the minor contributor in those samples, but instead the LRs in the chart referenced at the hearing are for the major contributor. Thus, as Dr. Reich explained, the memo does not further the claim that Bode validated samples in which the minor contributor contributed less than 5% of the DNA in a sample. (12T 86-24 to 87-21; Da 15 at 2-4)

There were many samples run through the validation study with extreme discrepancies between the mixture proportions estimated by STRmix and those intended to be created by Bode. A contributor intended to compromise 28% of a mixture was estimated to comprise 16% of that mixture by STRmix. (3T 80-1 to 81-2) A contributor intended to comprise 9% of a mixture was estimated to comprise 38% of that mixture by STRmix. (3T 81-15 to 82-18) Another contributor intended to comprise 9% of a mixture was estimated to comprise 33% of that mixture by STRmix. (3T 85-14 to 25) Another contributor intended to comprise 5% of a mixture was

estimated to comprise 30% of a mixture by STRmix. (3T 86-14 to 87-5) Either Bode analysts are very bad at making accurate mixtures—a problem caused by poor pipetting or other human errors—or STRmix is bad at estimating mixture proportions.

**iv. In its validation study, Bode tested only two samples in which the minor contributor contributed 25 picograms of DNA and 5% of the total mixture.**

In its validation study, Bode tested only two samples in which the minor contributor contributed 25 picograms of DNA and 5% of the total mixture. (D-60 at 4, Table 2a)

**v. In its validation study, Bode did not test the impacts of relatedness.**

In its validation study, Bode did not test mixtures that contained the DNA of related individuals nor tested the impacts of relatedness on the LR given to a non-contributor to a mixture that is related to the true contributor. (3T 101-23 to 25; 5T 86-22 to 24)

**vi. In its validation study, Bode did not validate any of the likelihood ratios used to hypothesis the real contributor is a relative of the person of interest.**

STRmix has the ability to created LRs with the assumption that the actual contributor had a number of familiar relationships to the person of interesting—sibling, cousin, etc. (7T 87-16 to 25; 8T 9-1 to 9; 9T 8-1 to 9, 72-18 to 23) Bode did not validate any of the related LRs and does not use the related LR, even though it might sometimes be the most appropriate one to run. (5T 84-21 to 24)

The failure to validate and use the related LR has a significant impact on case outcomes. One of the Bode samples gave an LR for 446 million for Paul, but a sibling LR of 5.9. (11T 46-24 to 47-16) In other words, the DNA evidence on that sample is only 5.9 times more likely to have come from Paul than from a brother, as opposed to 446 million more likely to have come from Paul than an unknown, unrelated person. (11T 50-7 to 13)

**b. Samples in this case tested by Bode were outside of the limits of the validation study or close to those limits.**

Many samples tested were beyond or at the limits of what Bode tested in its validation study. The following samples were below, at, or very close to the 5% contributor proportion, according to STRmix's estimates:

- E02b1 had a minor contributor that contributed 6% of the sample.
- E04a1 had a minor contributor that contributed 3% of the sample.
- E06a1 had a minor contributor that contributed 5% of the sample.

(Reich Report, Aug. 2024 (D-15) at 2-4; S-44; S-69; S-81 4T 66-18 to 68-25, 80-5 to 6, 87-16)

The following samples were below, at, or very close to 25 picogram minor contributor DNA amount, according to STRmix's estimates:

- E02b1 had 19.81 picograms of DNA attributed to the minor;
- E03b1 had 17.47 picograms of DNA attributed to the minor;
- E04a1 had 34.11 picograms of DNA attributed to the minor;
- E06a1 had 37.425 picograms of DNA attributed to the minor
- E07a1 had 25.194 picograms of DNA attributed to the minor

(Reich Report, Aug. 2024 (D-15) at 2-4; S-44; S-59; S-69; S-81; S-95; 12T 90-10 to 91-10)

Sample E02b1 therefore was just at the lowest contributor percentage of DNA tested by Bode in its validation and below the lowest amount of DNA tested by Bode in its validation. Sample E04a1 was below the lowest contributor proportion tested and slightly above the lowest amount of DNA tested. Sample E06a1 was just at the lowest contributor percentage of DNA tested by Bode in its validation and just above the lowest contributor percentage tested.

Every sample tested by Bode was theorized to have related contributors—the only potential contributors considered were the five Caneiros. (Bode Case File (D-66) at 5) Yet not a single sample tested involved related contributors. There is "a lot" of allele sharing in these profiles. (7T 87-22 to 25) Keith and Paul, for instance, share 70% of their alleles. (Bode Case

File (D-66) at 9) In E02b1, there are only there loci where Paul has alleles that none of the other [redacted] does not have. (Reich Report, February 2024 (D-14) at 10).

Below is the table of Bode results of what Ms. Reed believe was a mixture that are outside or close to the limits of what the validation summary tested:

| **Sample No.** | **DNA Template below or close to 25pg?** | **Contributor % below or close to 5%?** | **Related contributors in mixture?** |
|---|---|---|---|
| E01a1 | | | Yes |
| E02b1 | Yes | Yes | Yes |
| E03a1 | | | Yes |
| E03b1 | Yes | | Yes |
| E04a1 | Yes | Yes | Yes |
| E06a1 | Yes | Yes | Yes |
| E06b2 | | | Yes |
| E07a1 | Yes | | Yes |
| E10a1 | | | Yes |

**c. It was inappropriate for Bode to test and report results for samples outside of the boundaries of its validation study.**

As explained at length above, validation studies are supposed to set the limits of what can be reliably analyzed by a laboratory. Given how few samples were analyzed at those limits, how few samples came close to lose limits on both dimensions, and the failure for Bode to demonstrate that the number of samples tested of each kind was sufficient to draw reliable conclusions, there cannot be confidence in the reliability of the analysis of the samples below the limits and the samples right at the limits of what was studied.

STRmix results that went outside of the limits of the validation study are not reliable and should not be reported. Reporting results of these samples doesn't "conform to the Bode validation." (12T 171-2)

Because no assessment of STRmix's reliability when analyzing mixtures of related contributors was performed in the validation studies, these samples should never have even been

put into STRmix. Bode did not study either of the issues that arise with relatives: the possibility of false inclusion for a non-donor related to the real donor and the ability to properly interpret a mixture with related contributors. All of the potential contributors in this case are related to one another in a myriad of ways: (1) there is the possibility both that the real contributor is related to one of the Caneiros who is not a contributor but was run through STRmix; (the first problem identified above) (2) that the mixtures contain two Caneiros and person run against the sample is either related or unrelated to the Caneiros (the second problem identified above); or the mixtures contain two Caneiros and a third person, whether related or unrelated. Because allelic overlap can lead to underestimating contributors which can lead to false exclusions, subsection E.5, Reed's and unwritten subjective belief that the mixtures were made of two people does not obviate the need to understand the effects of relatedness, underestimation, and the subsequent interpretation.

As Dr. Inman explained, the analyst should have "stepp[ed] away from that immediately" from these samples because they contain related contributors. (16T 87-7) Dr. Buckleton notes the risks that mixtures composed of related individuals pose to the reliable use of STRmix and suggests some tactics to mitigate these risks, none of which Bode followed: the use of Mx priors, the use of conditioning, the use of "specialist software such as DBLR or MixKin" or "by simulating non-donor relatives and comparing them with the profile." John Buckleton, et al., Investigation Into the Effects of Mixtures Comprising Related People at 10 (D-1078).

Moreover, because mixture proportions or DNA amounts are below the limits of what was tested or too close to determine if STRmix performs adequately at those limits, results for samples E02b1, E03b1, E04a1, E06a1, and E07a1 should not have been reported. Reich Declaration, August 2024, at 3 (D-14) ("Bode has not established that it can reliably and

accurately analyze samples in that range.") Bode should not have analyzed either the mixture proportions or template amounts below the limits tested or the ones very close to those limits. As Dr. Reich explained, because of the uncertainty in measurements for low-level amounts of DNA, for mixture proportions that are close to the limits tested "you can't really tell whether you're at or below or just above the amount in the validation[.]" (12T 90-15 to 16) In order to be confident in testing samples right at the limits of what was tested, much more information about Bode's ability to reliably analyze DNA samples with contributors at those levels would have to be done. 12T 92-9 ("We don't know whether that's an accurate template amount. I don't think they made a mathematical error from the quantification but the quant is variable. And so we don't know what the actual amount of DNA is. Are we just under? Are we under by twice, three times? Are we above it and we're not aware of it? We don't have a validation study that goes beyond this so that we're more sure that this sample conforms."). See also 12T 206-11 to 21 ("[T]here in the Bode reports they absolutely have samples that are at the very border of what they have tried to test in this space, right? And because the numerical values of the quantitation, how much DNA we get are not precise. That is the method problem we have. Then they are at that edge, the conservative prospective would be either push your validation beyond it so we know what it is or take those out of the report because they are pushing the envelope and I do not believe in the criminal justice system we should be allowed to push the envelope.").

In short, accreditation standards, guidelines that are followed in the field, and bodies of experts all agree that the validation study has to establish the limits of the kinds of samples a laboratory can analyze in casework. What a laboratory has not tested in its validation study it cannot purport to be able to reliably analyze. And what is very close to what a laboratory has

tested also cannot be reliably analyzed without proof that sufficient samples of such a complexity were analyzed and the results were reliable in the validation.

**2. Bode's Standard Operating Procedures give very little objective guidance on what kinds of samples analysts should analyze. The little guidance given was not followed in this case.**

As found in subsection F.8.d.ii, SOPs are supposed to codify the limits found in the internal validation study in order to sufficiently guide analyst discretion and further uniform, reliable use of STRmix. Bode's SOP's are too meager to meet that goal. But the few guidance given by those SOPs were violated in this case by Bode's handling of related contributors.

**a. Bode's Standard Operating Procedures have very few limits on what samples to run and results to report.**

Bode has only two limits established on the use of STRmix: (1) No more than four contributors (3T 68-13 to 19; 4T 8-22 to 23); and (2) "genetic representation at at least three testing locations or loci in order to deem that component of the mixture interpretable." (3T 91-1 to 8) It is unclear what data the second limitation comes from, but it does not seem to come from any data disclosed in their validation summary. (12T 212-16 to 213-5) The first limitation comes from the fact that Bode did not test more than four contributors in its validation study.

Although the number of contributors is a hard limit, there were no limits established on minor template or contributor portion. No explanation was given for why the number of contributors should be a hard limit but template amount or contributor ratio is not.

**b. Bode's Standard Operating Procedures note increased risks of false positives on the kinds of samples analyzed in this case.**

Although it does not contain any objective guidance, the SOP does warn about the risks of analyzing samples that were ignored in this case.

**c. Despite the risk of unreliable results when first-degree relatives may be in a mixture, recognized in its Standard Operating Procedures, Bode ran samples with potential first-degree relatives without any explanation or risk mitigation.**

Bode's SOP recognizes that mixtures compromised of relatives are particularly hard to reliably interpret. The SOP explains that an "evidence profile may be unsuitable for interpretation if it appears there is a mixture profile from first-degree relatives." (D-60 at 9) First-degree relatives include an individual's parents, siblings and offspring. (3T 126-15 to 24; 4T 107-23 to 108-5) There is no other written guidance about assessing whether a mixture of related people is suitable for analysis.

Ms. Naughton testified that the SOP "give[s] guidance to our DNA analysts to evaluate the profile and the profile complexity and the case scenario to determine if relatedness could be affecting the interpretation. If it is not affecting the interpretation, even if it is a first-degree relationship, they are permitted to interpret that -- the profile." (3T 75-3 to 15) The only guidance given to determine whether allele sharing is affecting in an interpretation is as follows: "The analyst should scrutinize the data to ascertain if the peak height ratios indicate allele sharing to the extent that the deconvolution may not be intuitive." (3T 101-1 to 25, 125-2 to 20) There is no definition of intuitive or factors given to guide that intuition.

Many of the samples analyzed in this case potentially have first-degree relatives. E07a1 has multiple siblings and offspring/parent pairs and E10a1 has siblings ([redacted] and [redacted] (Bode Case File (D-66) at 5). E05a1 was only analyzed as having a niece ([redacted] and uncle (Paul), (Bode Case File (D-66) at 5), but an inclusionary LR was generated by Ms. Reed on the stand for the that niece's biological parents (Keith and Jennifer) as potential contributors. (5T 68-15 to 20, 71-16 to 21, 77-15 to 78-8)

These samples were run and reported without any documentation on the choice to visually exclude, on the choice of how many contributors, or on why Ms. Reed did not believe the allele sharing could present a problem for interpretation. (5T 22-11 to 23, 59-17 to 60-11)

**d. Despite the risk of unreliable results when profiles with very limited data are analyzed, recognized in its Standard Operating Procedures, Bode analyzed samples where there was very limited profile data about the minor contributor.**

The SOP also recognizes that samples in which the minor contributor contributed very little DNA are more likely to lead to a false positive. The SOP says that "[p]rofiles with limited data are likely to result in low LRs for true contributors and the chance of an adventitious match (positive LR when comparing to a non-contributor) is increased." (D-60 at 9) Yet E02b1 reveals only three loci where Paul does not share alleles with [redacted] (Reich Report, February 2024 (D-14) at 10) Despite the noted risk of false inclusions in these circumstances, noted in Bode's own SOP, it analyzed and reported a result for E02b1 anyway. There is no documentation as to why this choice was made.

**e. Rather than SOPs, Bode relies significantly on the DNA analyst's intuition.**

Given the lack of objective limits and guidance in the SOPs, there is significant reliance on the "intuition" and discretion of an analyst. (3T 101-2 to 25; 4T 7-24 to 8-14; 5T 104-12 to 25) The Bode SOP about the use of STRmix says: "Final discretion regarding interpretation is left to the analyst as experience is a significant factor in interpreting STR profiles." (D-62 at 2) It also states that "[t]he LRs generated by STRmix should be assessed at each locus for their intuitiveness." (D-62 at 18)

Despite this reliance on subjective discretionary choices, Ms. Reed explained that Bode does not require analysts to document any of their decision-making, which leads to an inability to later explain and review those decisions. (5T 22-11 to 23, 59-17 to 60-11) This is contrary to best practice, which requires that "When discretion is afforded to an analyst, documentation in the case notes or case file is necessary to record the reasons or justifications behind the decisions made. SOPs alone do not provide enough information regarding the discretionary decisions made for a particular case or sample." Human Factors at 30. The combination of a subjective standard

that is insufficient to guide analyst discretion and no requirement to contemporaneously explain the use of that discretion creates a situation fraught with the risk of error and bias.

**f. Despite analyst exposure to potentially biasing information, the Bode SOPs have no bias mitigation protocols.**

The Bode SOPs contain no bias mitigation techniques. Yet analysts are presented with potentially biasing information. For instance, before beginning her analysis Ms. Reed knew what happened, including that there were two juvenile victims, and knew that Paul Caneiro was the suspect. (5T 112-3 to 113-6) She knew this while making the decision not to even test Keith Caneiro's DNA on many samples. (5T 112-23 to 113-6) Yet Ms. Reed does not believe she is impacted by cognitive bias. (5T 114-18 to 20)

**g. The Bode SOPs do not contain an uninformative range.**

Despite the universally recognized risk that false positives are more likely to occur at low LRs, there is no mention of an uninformative range in Bode's STRmix interpretation SOPs. (D-62) The case file notes that the only uninformative LR is 1. (D-66 at 4) That means it reports all LRs above 1 as support for inclusion and all LRs below 1 as support for exclusion.

**h. Bode inappropriately conducted visual exclusions in this case instead of running all relevant people through STRmix.**

Reed visually excluded many Caneiros from many samples before running the samples through STRmix to compare them only to the people she determined would be considered. The choice to visually exclude means she did not run STRmix to generate an LR for those people. At the hearing, some of the persons visually excluded were run through STRmix with inclusionary results. For instance:

- For E06b1, Ms. Reed visually excluded Keith, but when run through STRmix he has an inclusionary LR of 9730. (5T 22-11 to 45-1)
- For E05a1, Ms. Reed visually excluded every Caneiro but Paul and she believed. contributor 3 was unsuitable for comparison. When Keith is run through STRmix

an LR of 60 for contributor 3 is produced. (5T 68-23 to 72-4) When Jennifer is run through STRmix an LR of 18.5 for contributor 3 is generated. (5T 77-15 to 78-8)

Ms. Reed did not document why she chose to visually exclude anyone. (5T 22-14 to 17) She initially said that the exclusion of Keith for E06b "may have been an error on Bode's part," both by Ms. Reed and by the technical reviewer. (5T 51-3 to 12) When confronted with the inclusionary LR in court, she said "I feel like I'm just freaking out." (5T 51-12)

Later Ms. Reed testified that she believes she would not exclude Keith if the analysis was done today, based on her evolving understand of the appropriateness of visual exclusions: "[i]t wasn't like an SOP change." (5T 53-11-12) These exclusions were therefore not supported by Bode SOPs. Dr. Buckleton testified that he doesn't believe visual exclusions are appropriate. (7T 19-1 to 5)

**3. In its validation study, New Jersey State Police did not test the impacts of relatedness.**

As with the Bode validation, the NJSP validation was a labor-intensive process that significantly involved ESR. The parameter document was almost completely done by ESR. (8T 52-17 to 20) ESR did the sensitivity and specificity part of the validation summary. (8T 88-6 to 90-7) NJSP did not technically review all of ESR's calculations for that section. (9T 25-6 to 16) NJSP has spent at least $400,000 on STRmix. (9T 49-14 to 25)

NJSP did not test either of the two impacts of relatedness on STRmix's reliability in its internal validation. (9T 32-1 to 7) NJSP is doing a relatedness study now. (9T 32-10 to 14) No discussion of allelic overlap and its relationship to error exists in the validation summary.

Other deficiencies with NJSP's validation study were identified at the hearing. NJSP also created mixtures with varying contributor ratios and DNA contribution amounts of its sensitivity and specificity study. The lowest amount of DNA by a contributor tested is 6.25pg. (8T 96-5 to 8) This is less than the amount of DNA in a cell,ha so that's always going to be less than a

complete copy. (9T 36-19 to 22) Only 11 or 12 samples have 6.25 picograms and it was unknown how many genotypes were represented by those samples. (9T 53-14 to 23; 12T 80-22 to 81-8) That is not "a lot of samples at this very low template." (12T 81-17)

Only 1000 known non-contributor were tested, which was insufficient to determine "how the program is going to work over time and over the variety of samples that forensic laboratories receive." (12T 74-11 to 15 In its study, NJSP used one database with computer generated profiles. (12T 75-24 to 76-3) NJSP also used staff for known-contributors, which is inadequate because "staff may not be representative of the type of samples they receive, the diversity of the individuals whose samples they receive" (12T 77-11 to 14)

The limits violated in this case was NJSP's decision to test a mixture comprised of related individuals and to test a person of interest who might be related to the true contributor against a mixture without ever having tested these kinds of samples in its validation study.

**4. New Jersey State Police Standard Operating Procedures give very little objective guidance on what kinds of samples analysts should analyze.**

Much like Bode, NJSP's Standard Operating Procedures contain almost no objective guidance to analysts as to what kinds of samples are unsuitable for comparison or what results are unsuitable for reporting.

**a. New Jersey State Police Standard Operating Procedures did not establish limits on suitability for STRmix analysis.**

Before implementing STRmix, NJSP used to have objective limits on the suitability of mixtures for analysis. A sample needed to have at least 400pg of DNA, and the major contributor had to have at least an average peak height of 400 RFU. (9T 19-17 to 20-15) SOPs before STRmix were objective for suitability for analysis and comparisons. (9T 22-3 to 23-7)

But since moving to STRmix, there are only two limits for analysis now. One is that no more than four contributors can be analyzed; that is the maximum number of contributors

analyzed in the validation study. (9T 30-12 to 16) The only hard limit for quantity is "data" at 7 loci in the whole sample. (9T 68-7 to 15, 26-13 to 27-21) There is no explanation for where the 7 loci limit comes from or why number of contributors is a suitable hard limit but template amounts or contributor amounts are not.

There is one limit established on reporting in the NJSP SOP. NJSP set its uninformative range at any LR below 1000. (8T 92-23 to 93-2) That means any LR below 1000 will be reported by the analyst as inconclusive rather than inclusionary.

**b. New Jersey State Police Standard Operating Procedures contain insufficient guidance on analyzing and reporting mixtures comprised of related individuals.**

Despite both NJSP analysts recognizing that the analysis of related contributors increases the risk of error, there is no objective guidance to an analyst faced with a mixture comprised of related contributors. NJSP's SOPs say to take into account relatedness when assigning number of contributors, but the SOP "does not get any more specific. It's more of an experience thing and something that's taught and discussed during training." (9T 31-8 to 10) None of that information taught at training about this issue is written down anywhere. (9T 31-11-12) The SOPs contain no other guidance on analyzing related individuals.

**c. Rather than Standard Operating Procedures, New Jersey State Police relies significantly on the DNA analyst's intuition.**

NJSP relies on the analyst's intuition and subjective judgment to determine if there are issues with a STRmix analysis. (8T 83-18 to 84-20, 97-12 to 18) Analyst intuition is wholly subjective and reference to it is insufficient to channel analyst discretion.

**d. Despite analyst exposure to potentially biasing information, the NJSP Standard Operating Procedures have no bias mitigation protocols.**

There is no mention of cognitive bias in NJSP's SOPs and no evidence was presented at the hearing that NJSP employs any bias mitigation produces. (Bode STRmix Interpretation SOP, D-66)

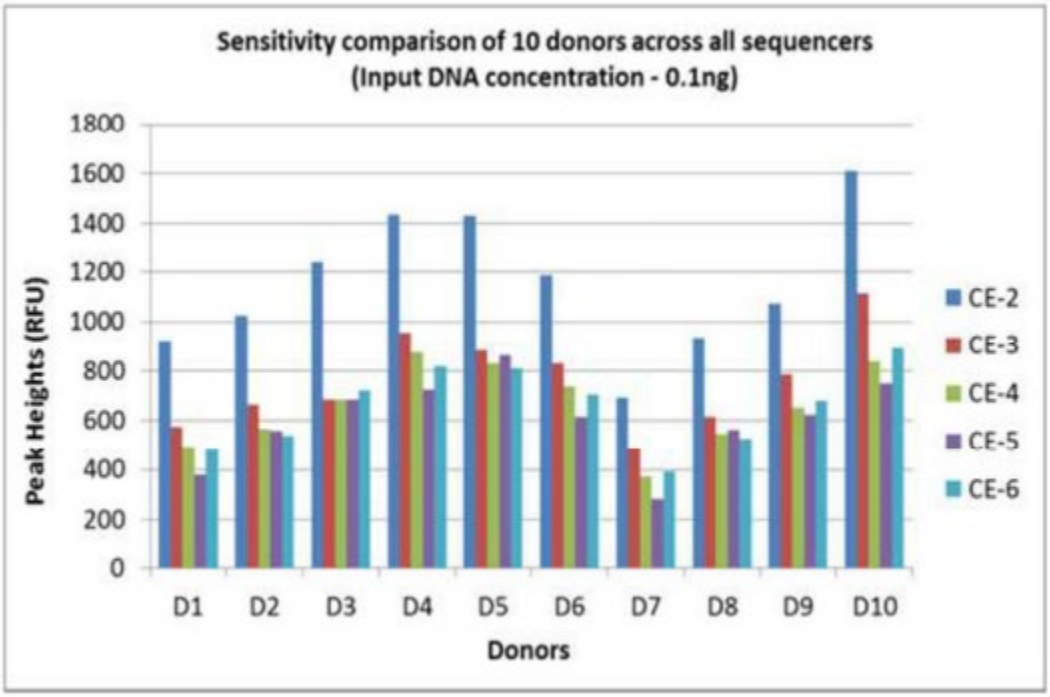
**5. Both Bode and New Jersey State Police conducted their sensitivity and specificity studies using mixture proportions and amount of DNA in picograms to assess performance.**

Sensitivity and specificity studies are designed to measure false positive and false negative errors. As the NJSP validation explains, "sensitivity is defined as the ability of the software to reliably resolve the DNA of the known contributors" and provide an LR greater than one, while "specificity is defined as the ability of the software to reliably include non-contributors." (S-162A/D50 at 13) Both laboratories conducted their sensitivity and specificity studies across a range of mixture ratios and amount of DNA template per contributor, as measured in picograms. (S-147/D-60 at 4, Table 2a; S-162A/D50 at 13-15) In other words, both laboratories chose to use contributor proportions and DNA amount measured in picograms to assess the performance of STRmix in their laboratories.

There was some discussion in Dr. Coble's reports and in his testimony that peak heights, as measured by RFU, are more important than template amount or contributor proportion to determine suitability for analysis. (Coble Report, Oct. 2023 (S-186/D-22) at 9-10; 10T 108-1 to 18) That opinion contradicts the SWGDAM guidelines and ANSI/ASB 18 standards, which explicitly requires laboratories to test STRmix's performance over a range of template amount and mixture proportions and do not mention RFU. SWGDAM Guidelines at 8-9 ("Internal Validation should address . . . mixed specimens" with "various contributor ratios (e.g., 1:1 through 1:20, 2:2:1, 4:2:1, 3:1:1, etc.)" and "various total DNA template quantities"); ANSI/ASB Standard 18 at 3 (developmental and internal validation "shall . . . include case-type profiles of

known composition that represent (in terms of number of contributors, mixture ratios and total DNA template quantities.") (emphasis added).

It is possible that peak heights would be a helpful addition to understanding the limits of the reliable use of STRmix, but it is not a substitute for these other metrics, both because of the requirements of these standards and because "peak height turns out to be a not a fixed value for the amount of DNA that we put into it. It's proportional roughly, but that roughly is an issue, because it varies a fair amount, up to 50 percent. And that makes correlating peak height with firm conclusions much more difficult." (12T 54-11 to 16) To that point, Bode's own internal validation shows a wide variation peak heights when the same exact amount of DNA is analyzed (an amount of DNA as low as 100 picograms, which is much higher than the lowest analyzed in this case). That variation is across both the individuals who contributed and across the capillary electrophoresis machines used:




(D-611; 3T 92-14 to 96-10)

The variation in RFU suggests it is not a superior, unusually accurate metric that should obviate the need for any other metric. In any event, both laboratories used picograms and percent contributors as explicit factors tested in their validation study and did not perform experiments with samples of varying RFUs. The performance and reliable use of STRmix can be assessed only based on the information provided by the summary, not by another factor that was not addressed and no information is presented about in the study.

**6. Neither internal validation summary provides sufficient information to assess the reliability of STRmix as used by these laboratories across different sample types.**

Neither internal validation summary provides sufficient information to assess the reliability of STRmix as used by these laboratories across different sample types. There is very little information in the Bode validation summary from which to draw any conclusions about the case types Bode can analyze reliably using STRmix. It is impossible to discern reliability limits from the charts provided in the document. (12T 85-15 to 20) Although there are false inclusions and exclusions in the validation summary, no error rates are provided. Bode, for instance, writes in its internal validation: "At high template, STRmix correctly and reliably gave a high LR for true contributors and a low LR for non-contributors." (Bode, Internal Validation Summary, D-60 at 15) "Reliably," "high," and "low" are all qualitative terms that are a matter of opinion.

Although the NJSP internal validation summary demonstrated that STRmix does generate false inclusions and exclusions, no error rate was provided. NSJP reported in its study that the highest inclusionary LR for a known non-contributor for non-degraded samples was 471, highest LR for degraded was 573. (8T 92-6 to 7to 95-3 to 6) Were more non-contributors run, there would likely be more false inclusions. (12T 75-5 to 10) Although Ms. Thayer claimed there were no false exclusions (8T 97-7 to 11) she considered only an LR of 0 to be a false exclusion, although an LR less than 1 for a contributor is appropriately considered a false exclusion.

The NJSP validation summary document is conclusory in similar ways to Bode's: "With the exception of these lower template high order mixtures, the plots in Figure 6 demonstrate that STRmix was able to reliably distinguish between true donors and non-contributors in these degraded samples." (NJSP, Internal Validation Summary, D-50 at 19) How low is the template when STRmix gets it wrong? How many contributors? What does "reliably distinguish" mean—no false positives, or just a number acceptable to the authors?

Mr. Inman explained that neither validation summary alone "provide[s] sufficient information to allow for a full evaluation" of the "breadth" and "depth" of the internal validations, whether the conclusions of the studies were "supported by the data generated" in the stud[ies], and whether the limitations of STRmix in these laboratories "were adequately determined and articulated in full[.]" (D-18 at 2-3; D-19 at 2-3) Mr. Inman explained that these summaries do not provide sufficient information about "what kinds of samples did they analyze and not merely in a summary form but what were the combination of types, exactly. How might that impact whatever conclusion they drew. You know would you think that they, that what was obtained was reasonable for what they put into it based on you know your own knowledge and previous work. Were there outliers that were discarded." (16T 99-9 to 17) These documents did not provide sufficient information to determine under what conditions false inclusions and exclusions occurred (16T 100-15 to 19), whether the laboratories tried hard enough samples (16T 100-24 to 25), whether they tried enough samples of each type (16T 101-8 to 13), and where the "edge cases" in which STRmix does not operate reliably in these laboratories are. (16T 101-7 to 13)

Exactly the kind of incomplete information presented in these validation summaries has been noted by NIST to be a key force in preventing a complete analysis of the reliability of PGS.

Information in internal validations "is often treated in aggregate and displayed as scatter plots. Without specific details about the samples, including the assigned LR values and metadata about the complexity of the mixture such as the degree of allele sharing, then reasons for differences cannot be independently assessed." Mixture Interpretation at 90. While many samples are tested in these studies, "it should be noted that important details are sometimes missing. For example, when differences in assigned LRs were observed in these publications, a reader typically cannot access the assigned LR values nor know anything about the degree of allele sharing in the mixture without the contributor genotypes." Ibid. See also Mixture Interpretation at 93 ("[S]ummaries typically do not provide data points (e.g., LR values) and associated information and metadata (see Box 4.1) necessary to assess the degree of reliability and performance under potentially similar case scenarios."); id. at 104 ("When aggregate graphs are provided in publications (e.g., Taylor 2014) or validation summaries do not include useful metadata for the data points displayed, an independent reviewer cannot assess or correlate the data and samples used to generate them.").

The State did not present any of the information necessary for a full, independent review by this Court of the reliability of STRmix in these laboratories. The defense attempt to do so was thwarted by an insufficient amount of time, compounded by receiving validation data in formats that were challenging and time-consuming to analyze. (16T 103-2 to 7) However, the magnitude of the undertaking of providing the data not presented by the State would have been significant. Four people spent "at least" hundreds of hours doing Bode's validation. (3T 107-6 to 21) There were 7 people involved in NJSP's validation within NJSP and it was "hundreds, if not thousands of hours of work," not including ESR's time. (9T 24-6 to 11) Mr. Inman would not have been able to complete his solo analysis of such a tremendous quantity of DNA for many more months.

(16T 134-17 to 19) But the burden is not on the defense to sift through the State's evidence and demonstrate the accuracy of the scientific method it is choosing to rely on. That is the State's burden, as proponent of the evidence.

**7. The laboratories did not stay without the boundaries of the reliable use of STRmix, whatever those undefined boundaries might be.**

Bode and NJSP both conducted internal validation studies that tested a range of mixtures. The results of these internal validation studies were presented in summaries that contained no usable data of the system's performance across sample complexity, that give qualitative commentary about the performance instead of quantitative information about its error rates, and that was insufficient for any independent person, including this Court, to draw any conclusions about the boundaries of STRmix's reliable use.

Compounding the concerns created by the conclusory validation summaries is that the laboratories tested samples and reported results in this case that were not tested at all in those studies. Neither of laboratory tested related individuals, yet each of these mixtures are theorized to contain related individuals. Bode did not test mixtures that were intended to have a minor that contributed less than 25 picograms or 5% of the total DNA, yet reported results where STRmix estimated that the minor fell below or just at those levels. The lack of regard for the validation studies—although each Bode and NJSP employee testified that the study was supposed to establish limits on the reliable use of STRmix—resulted in analysis that cannot be considered reliable. Moreover, the analyses in this case required a significant amount of human judgment that was not sufficiently challenged by objective SOPs or documented appropriately. Whatever STRmix's utility may be, and however well it is deployed in some laboratories, the evidence in this case shows that Bode and NJSP did not apply it reliably to this case.

## PROPOSED CONCLUSIONS OF LAW

### THE STATE'S PROFFERED STRMIX EVIDENCE FAILS TO MEET NEW JERSEY'S ADMISSIBILITY STANDARDS AS SET FORTH IN STATE V. OLENOWKSI AND N.J.R.E. 702.

As our Supreme Court has recently reaffirmed, "[r]eliability is critical to the admissibility of expert testimony." State v. Olenowski, 253 N.J. 133, 150 (2023) (Olenowski I)." "[A]n expert opinion that is not reliable is of no assistance to anyone." Ibid. (internal quotation marks omitted). To ensure that only reliable expert evidence is admitted at trial, trial courts perform an important role as gatekeepers. Id. at 154. In that role, they are guided by N.J.R.E. 702 and our caselaw interpreting it.

N.J.R.E. 702 governs the admissibility of scientific and technical evidence. In order for evidence to be admissible under N.J.R.E. 702, three requirements must be met:

> (1) the intended testimony must concern a subject matter that is beyond the ken of the average juror;
>
> (2) the field testified to must be at a state of the art such that an expert's testimony could be sufficiently reliable; and
>
> (3) the witness must have sufficient expertise to offer the intended testimony.

Id. at 153. The State has not demonstrated that prong (2) is met in this case.

Prong (2) requires that trial courts "assess the reliability of the theory or technique in question." Olenowski I, 253 N.J. at 152. In other words, "[f]or an opinion to be admissible under N.J.R.E. 702, the expert must utilize a technique or analysis with 'a sufficient scientific basis to produce uniform and reasonably reliable results so as to contribute materially to the ascertainment of the truth.'" State v. J.R., 227 N.J. 393, 409 (2017) (quoting State v. Kelly, 97 N.J. 178, 210 (1984)). The determination of whether the evidence proffered is reliable is two-

fold. First is a foundational review, which "entails a preliminary assessment of whether the reasoning or methodology underlying the testimony is scientifically valid[.]" Olenowski I, 253 N.J. at 147 (emphasis in original) (internal quotation marks omitted). Second is an as-applied review, in which the court must determine "whether that reasoning or methodology properly can be applied to the facts in issue," and were in fact reliably applied. Ibid. (emphasis in original) (internal quotation marks omitted). See also In re Accutane Litig., 234 N.J. 340, 378 (2018) (explaining that adopting Daubert helps to ensure "that only reliable and reliably applied expert testimony enters New Jersey's courts"); Fed. R. Evid. 702 (expert testimony admissible only when it is "the testimony is the product of reliable principles and methods" and when "the expert has reliably applied the principles and methods to the facts of the case."). As the proponent of the evidence, the State has the burden to "clearly establish" that the testimony is sufficiently reliable under N.J.R.E. 702. State v. Cassidy, 235 N.J. 482, 492 (2018).

The DNA evidence in this case fails to meet these standards for admissibility. First, the State has failed to demonstrate the foundational reliability of the two versions of STRmix used. Second, the State has failed to prove the as-applied reliability of the results generated by each laboratory. For either of these reasons, the DNA evidence must be excluded.

**A. The State has failed to prove that STRmix v2.5.11 and 2.8.0 are foundationally reliable.**

The State seeks to present DNA evidence generated from two versions of STRmix, v2.5.11 and v2.8.0. The State has failed to prove that these pieces of software are foundationally reliable.

Under Olenowski I, 253 N.J. at 464, in assessing the reliability of a technique, this Court is guided by a non-exclusive list of four factors:

(1) whether the scientific theory or technique can be, or has been, tested;

(2) whether it has been subjected to peer review and publication;

(3) the known or potential rate of error as well as the existence of standards governing the operation of the particular scientific technique; and

(4) general acceptance in the relevant scientific community.

Ibid. (internal quotation marks omitted). Before turning to these factors, it is important to bear two critical points in mind.

First, the State has the burden to "clearly establish" the reliability of the evidence as the proponent of the evidence. The defense bears no burden at all. Cassidy, 235 N.J. at 492. That means the State does not get the benefit of the assumption that any omissions in the record or in the testing would support reliability.

Second, there are two relevant communities whose opinions are relevant to reliability in this case—the forensic biology community and the software engineering community. Our appellate courts have already recognized that when it comes to probabilistic genotyping software, there is "more than one scientific community to consider." Pickett, 466 N.J. Super. at 302. The State cannot demonstrate reliability by relying solely on people with expertise in DNA, but must also demonstrate reliability as understood in the "computer science community to which [PGS] also belongs." Id. at 323. The insistence on consideration of both communities is not merely semantic—it is essential to ensuring that only reliable evidence is admitted in court. Probabilistic genotyping software, like STRmix, is making the impossible possible by purporting to reliably analyze DNA samples that no human could analyze. Because the final answers are beyond the scope of human comprehension, in order to trust that the answers are reliable, we must know that the software was built reliably and is used reliably by the humans involved. Software that is not built in compliance with software engineering norms is not reliable.

The State has not demonstrated STRmix to be foundationally valid because STRmix has been inadequately tested, its error rates are unknown, it does not adhere to the standards of the software engineering community, there are insufficient standards for its use in the DNA community, the software engineering community does not generally accept it to be reliable, and there is insufficient peer-reviewed publication of STRmix by independent authors. For these reasons, STRmix cannot be used in New Jersey criminal trials.

**1. STRmix v.2.5.11 and 2.8.0 have been inadequately tested.**

STRmix v.2.5.11 and 2.8.0 have been inadequately tested. It is important to bear in mind that there are two kinds of testing necessary to demonstrate the reliability of the software: testing as defined by the DNA community and testing as defined by the software engineering community. On both fronts, STRmix has been inadequately tested.

As found above, STRmix v2.5.11 and v.2.8.0 have failed to meet the standards for testing safety-critical software. Finding of Fact E. Those standards are the only way the reliability of software can be established, Finding of Fact D, and all experts agreed that those standards were not met in this case.

Three software engineers have reviewed STRmix and have found it fails to meet software the engineering standards to demonstrate reliability. There is no software engineer to testify otherwise. Verification and validation are to software what procedural due process is to the law: it doesn't guarantee that no mistakes are made, but if it is provided then we can be confident in the outcome of the system. Here, we cannot be confident in the outcome of the system because there has been no independent verification and validation.

Implementing the most rigorous testing standards are not a guarantee that a piece of software is unflawed. As IEEE writes in its disclaimer, meeting these standards is necessary but not sufficient to ensure the reliability of a system and to avoid adverse consequences: Meeting

these minimum standards cannot guarantee that there will be no software failures or adverse consequences. IEEE 1012-2016 at 21. But it is the bare minimum in the software engineering field to appropriately minimize those risks. And there is nothing preventing a company, such as ESR, from going beyond this floor in its own testing. STRmix does not meet the standards that any other piece of software that can impact life, liberty, or even a significant amount of money, meets. It is not exempt from these standards. The failure to meet them is a failure to demonstrate its reliability.

Nor has there been sufficient DNA testing of STRmix v2.5.11 and v.2.8.0. As found above, Finding of Fact G.1.a, the testing that has occurred is almost entirely by the developers of STRmix or interested parties, and it does not assess these versions in particular.

Perhaps most fatally, the reliability of STRmix has not been tested across the range of samples analyzed in casework, Finding of Fact G.1. Particularly relevant to this case, the testing that has been done has not established the reliability of STRmix when used to analyze mixtures made of related contributors or when the true contributor may be related to a non-contributing person of interest. As the Supreme Court of the United States has held, it is essential that the data marshalled in support of the reliability of a method be sufficiently relevant to the case at hand to be relied upon. In Joiner, the Supreme Court upheld the lower court's exclusion of the expert's opinion that a chemical called polychlorinated biphenyls (or PCB) caused cancer based on a review of studies of PCBs in animals. General Electric v. Joiner, 522 U.S. 136 (1997). The Court explained that "whether animal studies can ever be a proper foundation for an expert's opinion was not the issue. The issue was whether these experts' opinions were sufficiently supported by the animal studies on which they purported to rely. The studies were so dissimilar to the facts presented in this litigation that it was not an abuse of discretion for the District Court to have

rejected the experts' reliance on them." 522 U.S. at 144–45. How STRmix performs with a two-person sample of 300 picograms with a mixture proportion of 80% to 20% DNA is not probative to how it performs with the complex mixtures it is used on. For reasons discussed further in the next section, the error rates revealed by the testing are insufficient to demonstrate across which type of samples, if any, STRmix v2.5.11 and v2.8.0 can be reliably used.

**2. There is insufficient evidence of STRmix's error rate to demonstrate its reliability.**

In determining whether a "particular scientific technique" is reliable "the court ordinarily should consider the known or potential rate of error." State v. Olenowski, 255 N.J. 529, 594 (2023) ("Olenowski II"). There is no rate of error presented for STRmix—not in the developmental validation and not specifically for versions 2.5.11 and 2.8.0. The lack of information about the error rates is fatal to the admissibility of the STRmix.

As explained above, the error rates for STRmix should be known across samples of varying complexity. Even an aggregate error rate, which has not been provided, would not be sufficient. There should also be a rate of error or uncertainty for likelihood ratios in order to properly contextualize them; no such information was given in this case. Last, the risk of software error is not sufficiently minimized. No amount of testing of STRmix with ground-truth samples could substitute for appropriate IV&V, which is the only way to sufficiently decrease the risk of error in safety-critical software.

**a. The false positive and false negative rates are unknown.**

As found in Finding of Fact G.2.a, no false positive or false negative rates have been provided in a developmental validation for any version of STRmix, let alone the versions used in this case. We do not know how often these errors will occur across which kinds of samples, Finding of Fact G.2.b. and G.2.c., despite even the developers' acknowledgement that more complex samples will lead to more false positives.

Gesturing at the assertion that false inclusionary LRs are generally low—even if true—doesn't make them somehow insignificant or harmless. Many people will be falsely included in a STRmix analysis when the mixture is complex. There is no reason to think the jury would discern that a likelihood ratio of 100 is just not that much, even though it certainly sounds like a lot (it would be explained to the jury as follows: "The DNA evidence is 100 times more likely if defendant and victim contributed their DNA than if the victim and un unrelated and unknown individual contributed their DNA." That certainly sounds to a lay person as though the defendant is 100 times more likely to be guilty than not guilty, no matter how much of a misunderstanding of Bayes' theorem that entails). A low likelihood LR is especially misleading because the jury will not be told by the analyst that defendant is only one of many people who could be contributors to this mixture according to STRmix.

Focusing on the assertion that false positive LRs are more frequently going to be low also obscures the fact that high false positive LRs will occur sometimes. This is an acute concern in a case like this where a mixture contains related individuals and the person of interest could be related to the true contributor. The fact that error is unavoidable means that STRmix can only be used under circumstances where it has been demonstrated that the likelihood of error is sufficiently low.

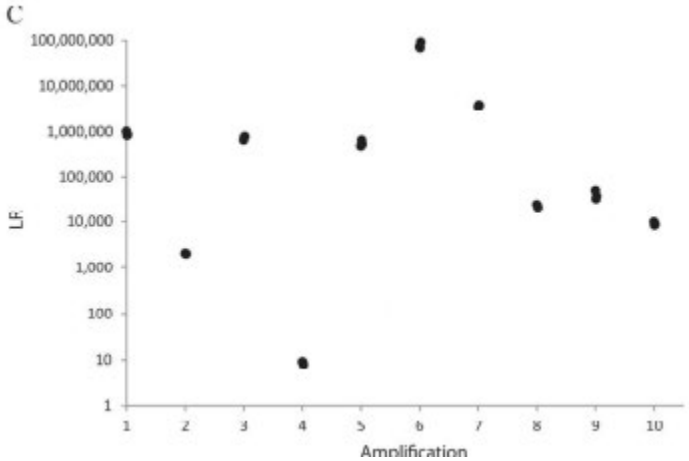
The lack of error rate is particularly important because jurors assume DNA analysis is close to infallible. One study found that jurors believe the rate of error in DNA analysis is 1 in 1,000,000,000 (one in one trillion). Jonathan J. Koehler, Intuitive Error Rate Estimates for the Forensic Sciences, 57 Jurimetrics 153, 163 (2017). See also Lisa Smith et al, Understanding Juror Perceptions of Forensic Evidence: Investigating the Impact of Case Context on Perceptions of Forensic Evidence Strength, 56 J. Forensic Sci. 409, 413 (2011) (finding that mock jurors

perceived the greatest increase in the probative value of DNA evidence when presented in a criminal context when the DNA evidence was moderate or weak). See also PCAST Report at 152 ("To be considered reliable, the [false positive rate] should be less than 5 percent and it may appropriate that be considerably lower, depending on the intended application.") (emphasis added). Because DNA is considered the "gold standard," juries are apt to trust it reflexively. That makes it even more important that any DNA evidence presented at trial is reliable

**b. The risk of error is not limited to false inclusions and exclusions, but incorrect likelihood ratios.**

Not only is there no error rate, but there is also no way to know how often STRmix produces plausible but incorrect LRs. G.2.e. Subtle errors that yield incorrectly overinflated LRs that are correctly inclusionary or exclusionary are not easy to catch and are not even captured by the number of false positives or negatives. Because there is no one, true, correct LR for any given sample, knowing whether the LR produced reflects the correct operation of a specific PGS as designed will often be impossible.

The State seems to attempt to dismiss the question of the correctness of the LR, focusing only on whether the LR is appropriately inclusionary or exclusionary in ground-truth studies. But STRmix does not produce a binary in or out answer. STRmix produces an LR. It is the correctness of that output that is at issue. But as found above, Finding of Fact C.3., LRs produced by STRmix can vary widely based on analyst choices, based on the model used, based on the machines used in a specific laboratory, and even based on analyst skill in pipetting. One study found that in the same laboratory amplifying the same DNA sample ten times on the same equipment and running each amplification through STRmix produced LRs ranging from 10 to 100,000,000. (12T 58-7 to 60-4):




(D-610)

That is a huge variation in LR. There's no indication of which LR is the one STRmix should have produced if used properly and no indication of a range of error in LRs that could be presented to a factfinder to contextualize the LR. A trial in which inculpatory DNA evidence is produced with a LR of 10 looks significantly different than a trial in which inculpatory evidence is produced with an LR of 100 million. That range is too substantial to ignore.

Because of the potential range of LRs produced by any given analysis, "the recipients" of LR assessments "in a given case need guidance on what to do situations where variations among different LR assessments could potentially impact the outcome of trial." Mixture Interpretation at 76. One way to address the range of potential LRs for any given sample due to "sampling variability, measurement errors, and variability in choice of assumptions and choose of models" is by "express[ing] the effect of such variabilities on an LR value by offering an interval estimate." Steven P. Lund and Hari Iyer, Likelihood Ratio as Weight of Forensic Evidence: A Close Look, 122 J. Research of Nat'l Inst. Standards & Tech. (2017) . A laboratory could also run a sample more than once by different experts and present the range of outputs to the jury. Mixture Interpretation at 76. Perhaps there are a number of approaches to accurately convey the uncertainty associated with a given LR to a jury. But it is misleading and violates N.J.R.E. 403

for a jury to be told that the LR for a given set of hypothesis is 10,000 without being told that it could have been 10 if it was run again.

**c. Testing STRmix by running DNA samples does not capture the risk of software error.**

Software errors can exist independently from the problems with a mathematical model. Finding of Fact D.6. Because of this, the DNA error rates do not capture the full range of possible error. The concern from the software engineering perspective is that the software may be failing in ways we can't predict. That is because this kind of "software is non-continuous, meaning that correct results for the samples used in the validation studies do not preclude the possibility of erroneous results for others that do not match those samples." Pickett, 466 N.J. Super. at 299.

With software, proper V&V is the only way to appropriately minimize the risk of error. Software works until it doesn't. Radiation therapy kills cancer patients, Bloomberg terminals crash the stock market, and a Boeing airplane crash kills 200 people because of a lack of sufficient software testing. (D-13 at 16) And these catastrophic errors occurred in industries with actual regulatory agencies and requirements for V&V, as opposed to the forensic science field, which is almost entirely unregulated and in which the defense finds itself standing alone attempting to act as a regulatory agency. How frequently do planes need to fall out of the sky for the risk of error to be unacceptably high? Every 1,000 flights? Every 10,000 flights? Every 100,000 flights? Why should our tolerance for stripping a person of their liberty—possibly for the rest of their lives—be any higher? Why would we not require the people selling this software to do everything possible to mitigate the risk of that catastrophic failure—in this case, by thoroughly testing their software according to software engineering standards?

**d. Without a known rate of error, STRmix is not admissible.**

The State is seeking to admit evidence that jurors find uniquely compelling generated by a system without a known error rate. If admitted, people are going to go to prison for the rest of their lives because of this evidence. That cannot happen if we don't know how often and when it produces erroneous results.

It was asserted many times that STRmix generating false positives for complex samples is STRmix working as expected. (7T 101-13 to 20, 107-3 to 109-1; 11T 33-18 to 25) Assuming that that assertion is correct and the errors are the result of limits of genetics and not flaws in STRmix's mathematical model or software implementation, that does not mean STRmix is admissible. Imagine a parachute built to the best standards by the smartest people. Imagine that that parachute always works as intended. Imagine that because of the law of gravity and other immutable laws of physics, as the height from which a person jumps is higher, the risk of death is higher. The designers of the parachutes are blameless of this unchangeable truth, and nothing could be done to improve this parachute. Before anyone is going to jump off a building, the very first question they would ask is: how often do people die when jumping at what height? If, at 120 feet, the risk of death is 20% or 50% or 70%, a person would have to think very hard about if they wanted to jump with that parachute. The specifics of that error rate matter. If the developers of the parachute merely said, "yes when you jump from high enough you are going to die but that is not the parachute's fault, that it is actually the parachute working as expected, and we will not tell you how high and we will not tell you the risk," no one would use that parachute. But that is exactly what STRmix is doing here.

It is quite possible that STRmix has acceptably low rates of error on easy samples. But we don't know what those samples are. We don't know when those rates of error become unacceptably high for the criminal justice system, even if they are a sign that STRmix is "working as expected." The job of STRmix developers is to make STRmix and put out the information necessary for this Court to make the decision about if and when it operates with a low enough rate of error to be admissible. That decision is this Court's. But STRmix refuses to put fourth that information. That refusal prevents this Court from acting as a gatekeeper and means that STRmix cannot be admissible in any circumstances.

In sum, the error rates are inherently knowable, they are simply not provided. This is a sharp contrast to Olenowski, in which our Supreme Court accepted less rigorous information about error rates for the DRE program, which is "less testable and the error rate less knowable" than the ideal. Olenowski II, 255 N.J. at 598 (internal quotation marks omitted).

**3. There are insufficient standards governing the use of STRmix.**

On the DNA end, there are insufficient standards for validation of PGS software. Finding of Fact G.3 On the software engineering end, there are robust standards that are not being employed by ESR. Sections D and E. For either reason, this factor weighs against the admissibility of STRmix v2.5.11 and v2.8.0.

Dr. Buckleton points to DNA standards for validation, most notably the SWGDAM standards. However, all of these standards merely lay out what sort of mixtures a laboratory should test during validation, not what the results need to be in order to determine that a laboratory can interpret those sorts of mixtures reliably. It would be like arguing that requiring that one sits for the bar exam, but without any score for passing that exam, would be sufficient to ensure the quality of lawyers. The point of a test is not the taking, it's the passing.

The lack of standards also means that there is a wide variety of choices laboratories implementing STRmix make, which means outcomes in each laboratory are going to differ. Within each laboratory, there are Standard Operating Procedures (SOPs), which should standardize analyst discretion, but they are very subjective. The range of laboratories' implementation of STRmix and analysts' use of STRmix of the core reasons why, as found above, neither a developmental validation study nor a different laboratory's internal validation is sufficient proof that STRmix is reliable in a specific laboratory. Finding of Fact F.8.

Further, as found above, there are rigorous and detailed standards for the development of software. Both Dr. Buckleton and the ISFG both believe those standards should apply to STRmix. However, those standards are simply not being met by STRmix's developers.

**4. There are almost no independent, peer-reviewed publications that assess STRmix's reliability.**

Our courts have repeatedly recognized the importance of peer review in reliable science. Reviews by other experts are, in part, what "push the scientist to explain his or her work clearly and which raise questions that might not have been considered." National Research Council, Strengthening Forensic Science in the United States: A Path Forward 112 (2009) (D-9/S-141); see also State v. Henderson, 208 N.J. 208, 242 (2011) ("The peer review process is a method of quality control that ensures the validity and reliability of experimental research."); State v. Pickett, 466 N.J. Super. 270, 312-13 (App. Div. 2021) ("The peer-review process entails a review for accuracy and quality of a scientific paper, in which a scientist describes his or her research and conclusions, and it is either accepted or rejected by two anonymous members of the relevant scientific community.").

Publications about STRmix are largely written by the developers and published in journals that developers sit on the board of that have a striking tendency to cite their own work.

Finding of Fact G.4. Our Supreme Court has made clear that the evidence provided by experts who have an interest in the acceptance of a technique cannot suffice to demonstrate that a technology is generally accepted. In Windmere Inc. v. Int'l. Ins. Co., 105 N.J. 373, 380-81 (1987), the Court held that the proponent of voiceprint analysis did not establish its general acceptance. In coming to that conclusion, the Court noted that the only witnesses who testified in favor of the acceptance of the technology "were experts affiliated with the development of the device" at that specific lab. Id. at 280. The Supreme Court contrasted the limited testimony of interested experts presented in Windmere with the much more robust record of impartial experts presented in Romano v. Kimmelman, 96 N.J. 66 (1984). In Romano, the Supreme Court held that certain breathalyzer models were admissible under N.J.R.E. 702. In the admissibility hearing for that evidence, the State presented seven expert witnesses, at least four of whom had no connection to the device or the State police who wanted to use the device. Windmere, 105 N.J. at 382; see also id. at 380 (quoting People v. Kelly, 549 P.2d 1240, 1249 (Cal. 1976)) (quoting the California Supreme Court's opinion rejecting the testimony of an expert who "has virtually built his career on the reliability of the technique" and warning that a person "too closely identified with the endorsement" of a new technique does not aid the Court in establishing its acceptance as a "more detached and neutral observer" might be able to do). The articles written by the very people whose livelihood and professional reputation depend on STRmix being purchased by other laboratories are not independent enough to be weighed heavily.

Not only are these publications largely not independent, they are devoid of the kind of detail and data that would allow for an independent assessment of STRmix's reliability. Peer-reviewed publications also do not assess the reliability of the implementation of STRmix as a piece of software nor could they; peer-review and publication are not acceptable methods of

demonstrating the reliability of software. In short, STRmix has not been subjected to the kind of rigorous review that would demonstrate its reliability, let alone STRmix v.2.5.11 and v.2.8.0 specifically.

**5. The State has not demonstrated general acceptance in the relevant fields.**

The defense does not dispute that many laboratories use STRmix and that shows some degree of acceptance by part of forensic DNA community. However, law enforcement using forensic techniques that have not been fully validated is a hallmark of the misuse of forensic science, not proof of the reliability of the technique. And, as discussed above, the acceptance by the developers and makers of STRmix is inherently biased.

Reliability has to be empirically demonstrated, not assumed. The information available about STRmix does not demonstrate its reliability. The National Institute of Standards and Technology, an organization created by the federal government in 1901 that has the mission to advance science and technology, thinks the publications and information about STRmix that is available in the public domain is insufficient to assess the reliability of any given PGS. Mixture Review at 90 (explaining that the information in these publications is insufficiently detailed for a meaningful evaluation of PGS performance), id. at 93 (explaining that internal validation summaries do not provide the information "necessary to assess the degree of reliability and performance under potentially similar case scenarios"); id. at 104 (explaining that because aggregate graphs, as opposed to data, are provided in publications and validation summaries an independent reviewer cannot meaningfully assess the claims made in the studies). The information NIST reviewed constitutes almost all of the information put forth by the State in this case. Dr. Karl Reich, an expert in DNA who does not solely work for law enforcement, also raises concerns about STRmix in his supplemental report. (Reich Report, Aug. 2024 (D-15 at 8-

28) From decisions about analytical thresholds to addressing stutter, it's far from clear that the decisions made by the software and in the laboratory are generally accepted to be reliable.

But even users and supporters of STRmix understand that its use cannot be unlimited, and that it is only generally accepted to be reliable within some limits. It is well-established in the field that the use of any probabilistic genotyping system, including STRmix, can only be demonstrated to be reliable through validation testing on sufficient samples of the kind of complexity present in casework samples. Finding of Fact F.8. That is why laboratories conduct internal validation studies. That is why neither Bode nor NJSP would run an 5-person mixture in its laboratory. That is why neither laboratory would try to analyze a sample with 1 picogram of DNA. This is why NIST cautions that "[t]he degree of reliability or trustworthiness of a given PGS system in a given case is dependent upon the number of instances in which that system has been tested with samples that are judged to be of similar complexity as the casework sample, the performance of the method among those instances, and how the characteristics (e.g., number of contributors, DNA amounts, level of degradation) of the ground-truth known samples compared to those of the sample in the case at hand." Mixture Interpretation at 76. And this is why, as found above, every single DNA analyst in this case agreed that limits on the reliable use of STRmix must be established by assessing how STRmix performs in a given laboratory across samples of varying complexity. Finding of Fact F.8.d. The only person who disagreed was John Buckleton, the person who has crafted his own professional career around STRmix's use and acceptance. While his belief in the unfettered utility of his own software is consistent, he is not the voice of the community, but the voice of the person most entwined with STRmix.

Critically, STRmix is not generally accepted to be reliable in the software engineering community. Three different, unrelated software engineers have made clear that STRmix cannot

be considered reliable by the standards in that field. Section E. The State does not put forth a single software engineer to argue otherwise. As Pickett established, when there is more than one relevant scientific community, the opinions of both matters. Pickett, 466 N.J. Super. at 302. In a situation such as this one, where the functioning of a technique is dependent on both fields, a total failure to meet the standards of one cannot make up for the evidence of some acceptance in the other.

**6. Very few courts have actually looked at the reliability of STRmix in a thorough and nuanced way.**

Out-of-state cases are of course not binding on this Court. But they are only as persuasive as they are well-reasoned and thorough. Unfortunately, almost all of the cases that address the admissibility of STRmix are neither well-reasoned nor thorough. And even the more thorough cases support Mr. Caneiro's as-applied challenge.

**a. Court opinions do not provide guidance when they do not substantively engage with the arguments presented in this case.**

There is no binding authority on the reliability of STRmix to govern this case. No New Jersey court has come to a conclusion about the reliability of STRmix after a full hearing that included source code review, there has been no published appellate decision reviewing such a conclusion, and these specific versions of STRmix used by these specific labs have not been examined. Nothing other courts have done is binding in New Jersey, but these other decisions are not even persuasive if they are not built upon a sufficiently thorough foundation. In shifting to Daubert, our Supreme Court made very clear that it was not going to reflexively sweep in reliability decisions from other jurisdictions:

> As we did in Accutane, however, we decline to embrace the full body of Daubert case law as applied by state and federal courts. The Daubert factors will help guide trial courts as they perform their important role as gatekeepers. But Daubert's non-exhaustive

> list of factors does not limit trial judges in their assessment of reliability. The same is true for caselaw from other jurisdictions, which can be persuasive but is not controlling.
>
> Olenowski I, 253 N.J. at 154.

Those cases are only as good as the underlying litigation and the reasoning of the decisions. As now-Justice Fasciale explained in Pickett, determinations of reliability of other PGS systems were not even persuasive when they "entailed no scrutiny of computer science or source code." Pickett, 466 N.J. Super. at 314. Repeating the less-than-fully informed or less-than-fully reasoned opinions of other jurisdictions, "without further scrutiny, creat[es] an authority house of cards." Id. at 316 (internal quotation marks omitted). That is a mistake that our courts do not make, rejecting reflexive reliance on other out-of-state caselaw to find many forensic fields unreliable. See e.g., State v. J.L.G., 234 N.J. 265, 288 (2018) (holding that Child Sexual Assault Accommodation Syndrome was not demonstrated to be reliable enough to be admitted, despite the fact that 40 other states and the District of Columbia allow CSAAS testimony); State v. Doriguzzi, 334 N.J. Super. 530, 546 (2000) ("Reliance upon other courts' opinions can be problematic: Unless the question of general acceptance has been thoroughly and thoughtfully litigated in the previous cases, . . . reliance on judicial practice is a hollow ritual."

**b. Courts often reflexively accept forensic science as reliable without sufficient examination of the method. New Jersey courts do not.**

The saga of the Forensic Statistical Tool (FST) is a great illustration of the dangers of blind acceptance of novel forensic science and of New Jersey courts' refusal to fall into that trap. The New York Office of the Chief Medical Examiner (OCME) developed its own in-house PGS known as FST. People v. Williams, 147 N.E.3d 1131, 1150 (N.Y. 2020) (DiFiore, J., concurring). It was used in casework beginning in 2011. After years of use and OCME's consistent refusal to share the inner workings of FST with courts or defense counsel, a court ordered source-code

review of the FST. United States v. Kevin Johnson, 15-CR-565 (VEC) (S.D.N.Y. July 6, 2016). That review was conducted by Mr. Adams. Notably, OCME sent him the source code on a CD or flash drive, without any of the restrictions imposed by ESR in this case. (14T 13-1 to 25) Mr. Adams found an undisclosed modification that changed FST's behavior from its original validation study, an error that was introduced while OCME tried to fix another error. (14T 15-1 to 18-24) These software faults "proved crucial to identification of significant errors, albeit not before compromised test results had already been used in many prosecutions." State v. Rochat, 470 N.J. Super. 392, 438 (App. Div. 2022) (internal quotation marks omitted).

Despite these flaws, court after court admitted FST evidence for two reasons: (1) because it had been validated by the DNA subcommittee of the New York Commission of Forensic Science and (2) because other courts had said it was reliable. It took until 2020 for the highest court in New York to put an end to these recursive, substance-free judicial approves of FST. State v. Williams, 147 N.E.3d 1131 (N.Y. 2020). FST is no longer used in New York. (14T 102-18 to 103-22)

The Williams New York Court of Appeals held that the lower court erred in relying on these rationales to deny the defendant a Frye hearing on FST. In coming to this conclusion, the Court explained that it was error to rely on the approval of the Subcommittee as sufficient to demonstrate the approval of the scientific community at large. Ibid. As the Court of Appeals noted, the Subcommittee's "insular endorsement is no substitute for the scrutiny of the relevant scientific community. To rely solely on the Subcommittee's approval as dispositive of the general acceptance would be supplant the courts' obligation to ensure, under Frye, that scientific techniques and methods are sufficiently reliable to be admitted into evidence in a criminal proceeding." Id. at 1142.

The Court also explained the mere fact that other courts have repeated the mistake of admitting FST without actually establishing its reliability does not make it any less a mistake. Relying on the poorly reasoned decisions of other courts is not a sufficient substitute for assessing the reliability of a scientific method. As the court explained, "[t]he repetition of a single, questionable judicial determination does not strengthen or add validity to such ruling, and it defies logic that an error, because it is oft-repeated, somehow is made right." Williams, 147 N.E.3d at 1140.

Our Appellate Division agreed, holding in Rochat that the State failed to demonstrate the reliability, and therefore admissibility, of FST, which had been used in a New Jersey criminal case. State v. Rochat, 470 N.J. Super. 392, 437 (App. Div. 2022), cert. denied, 252 N.J. 79 (2022). Our Court held that the analytically lax New York cases was insufficient to demonstrate the reliability of FST, that the testimony of two people who had worked at OCME was insufficient to demonstrate the reliability of FST, and the approval of New York's Commission of Forensic Science was insufficient to demonstrate the reliability of FST when no other members of the scientific community had reviewed it. Rochat, 470 N.J. Super. at 440-41. The State had failed to meet its burden to clearly establish FST's reliability. Id. at 441. This Court should not make these same mistakes in assessing whether the State has met its burden to demonstrate the reliability of the versions of STRmix at issue here.

**c. No judicial opinions persuasively demonstrate the reliability of STRmix.**

Much like court cases rubber stamping the use of FST, most cases approving of STRmix are not substantive or rigorous. Almost all of the cases cited by the State in its pre-hearing brief did not involve a source code review or consideration of the views of software engineering community, did not include a defense expert, and relied largely on the admissibility of STRmix in other courts. See United States v. Williams, 2023 WL 5155252 (D. Minn. 2023) (no

consideration of software engineering or code review, no hearing, no challenge presented about whether the samples analyzed were outside the bounds of the validation study, and a one-page opinion); United States v. Jaber, 2023 WL 8254358 (M.D Fla. 2023) (no challenge to STRmix's foundational reliability, no challenge presented about whether the samples analyzed were outside the bounds of the validation study, no consideration of software engineering); United States v. Washington, 2020 WL 3265142 (D. Neb. 2020) (no consideration of software engineering, does not appear to have had any witnesses testify at a hearing, the sole evidence presented by the defendant was a "news article" by NIST, and significant reliance on the fact that other courts have found STRmix reliable); People v. Blash, 2018 WL 4062322 (V.I. Super. 2018) (no substantive consideration of software engineering, no defense witness, no challenge presented about whether the samples analyzed were outside the bounds of the validation study, and significant reliance on the fact that other courts have found STRmix reliable); People v. Bullard-Daniel, 54 Misc. 3d 177, 42 N.Y.S.3d 714, 724-25 (N.Y. Co. Ct. 2016) (at a Frye hearing, the court considered only the views of forensic DNA analysts, the defendant did not present an expert in forensic DNA analysis, no challenge presented as to whether testing exceeded the limits of the validation study); United States v. Christensen, No. 17-CR-20037-JES-JEH, 2019 U.S. Dist. LEXIS 24623, 2019 WL 651500, at *2 (C.D. Ill. Feb. 15, 2019) (no substantive consideration of software engineering or code review, no defense witness, no challenge presented about whether the samples analyzed were outside the bounds of the validation study); United States v. Russell, CR-14-2563 MCA, 2018 WL 7286831, at *5 (D.N.M. Jan. 10, 2018) (no consideration of software engineering or code review, no defense witness familiar with STRmix, no challenge presented about whether the samples analyzed were outside the bounds of the validation study); People v. Smith, No. 340845, 2018 Mich. App. LEXIS 3274 (Ct. App. Oct. 9,

2018) (no challenge to reliability of STRmix results); People v. Seepersad, 101 N.Y.S.3d 701 (N.Y. Sup. Ct. 2018) (no challenge to the reliability of STRmix results); United States v. Pettway, No. 12-CR-103S (1), (2), 2016 U.S. Dist. LEXIS 145976, 2016 WL 6134493 (2016?) (no hearing held, no consideration of software engineering or code review, no defense witness, no challenge presented about whether the samples analyzed were outside the bounds of the validation study).

Those opinions are simply not good enough to take any guidance from.

**d. United States v. Lewis**

The two main opinions relied on by the State in support of the idea that there is meaningful judicial acceptance of STRmix are not only not binding, but they in large part poorly reasoning and in some part distinguishable in important ways. United States v. Lewis, 442 F. Supp. 3d 1122 (D. Minn. 2020), relies on large part on Dr. Thompson's report as special master in that case. Special Master's Report on the Scientific Foundations of STRmix, United States v. Kenneth Davon Lewis, available at https://bpb-us-e2.wpmucdn.com/faculty.sites.uci.edu/dist/0/594/files/2021/08/Special-Master-Report-10-31-19.pdf .

Dr. Thompson correctly found that STRmix "has not undergone the stringent verification and validation process specified by IEEE standards for safety critical systems." Id. at 37. He even correctly acknowledged that the standards of the software engineering field cannot be dismissed by the opinion of forensic scientists: "While it is tempting to conclude that forensic scientists are in a better position to set standards for forensic science than software engineers, it is not clear that this is true when the standards apply to complex software systems. Software engineers may well be better positioned to understand what might go wrong with such a system and how to minimize those risks." Id. at 37. But he took a wrong turn when he concluded that the

"concerns about the possibility of undetected points of failure are . . . somewhat hypothetical and must be weighed against evidence that this program has worked well in validation studies." Ibid. For all the reasons explained by the software engineering experts in this case, dismissing software engineering standards because no errors have been detected is a mistake, especially given the limits of the reviews that occur in criminal cases and because the software engineering documentation presented is simply insufficient.

Dr. Thompson also inappropriately dismissed the importance of defining STRmix's error rate, conclude that although in the studies "there were indeed many instances in which non-contributors were assigned LRs above one," most of "these misleading STRmix results were only weakly incriminating." Id. at 34. That's not good enough. This Court needs to be able to make an independent assessment of the reliability of STRmix by assessing its error rate. Being told that it exists and is indeterminate but it's usually not a big mistake is insufficient, especially since there is no reason to believe a jury would not take "low" LRs seriously.

Dr. Thompson's report supports the as-applied challenge in this case. Dr. Thompson agreed that the internal validation of the laboratory in that case did not establish the boundaries of validity of STRmix, a concern that is "well-founded and deserves to be taken seriously." Id. at 40. However, in that case, the sample analyzed did not present a "difficult case" in Dr. Thompson's opinion, because the defendant was the major contributor and his alleles across 22 loci were detected in the mixture. Id. at 40-41.

As an initial matter, that means that Dr. Thompson and the defendant both agree that the validation study should establish the boundaries of the reliable use of STRmix in any given laboratory. Where Dr. Thompson went wrong was determining that a specific sample was within

bounds that do not exist. Without those boundaries being established, it is inappropriate for a court to conjecture that a specific sample would be within those boundaries.

But in Lewis, the samples tested were "similar in all relevant respects to the bulk of mixtures that were successfully analyzed" by the 31-laboratory study and the FBI study. Id. at 41. This is beside the point, because no other internal validation study can substitute for a specific laboratory's study, even a study that aggregates many other validation studies. Bode's people have to be able to use STRmix properly and its equipment has to be able to handle it for any casework that comes out of Bode to be reliable. But in any event, the State has presented no information that the samples in this case are sufficiently similar to samples studied in those case that STRmix did reliably analyze. But Lewis, and Dr. Thompson's comments since Lewis, do support the proposition that a laboratory's internal validation limits its reliable use of STRmix and that there is grave concern with the reliability of STRmix in particularly challenging samples.

**e. United States v. Gissantaner**

United States v. Gissantaner, 990 F.3d 457 (6th Cir. 2021), the other case the State relies on, is simply not an example of nuanced judicial scrutiny of a scientific record. The court inappropriately dismissed the importance of independent authorship of scientific articles. Id. at 465. It did not consider that peer-review is less rigorous when the authors and their friends are on the boards of the journals publishing these articles. Ibid. It considered only aggregate rates of false positives and false negatives, not considering the importance of disaggregating them across sample types and not considering that the LR itself could be incorrect in magnitude even if correctly inclusionary or inclusionary. Ibid. See Fact Finding E.3. Unfortunately, the opinion is just poorly reasoned. Moreover, the opinion lends support to defendant's as-applied challenge:

the internal validation in that case did include samples as complex as the ones analyzed in that case. Id. at 467.

**7. The State has failed to demonstrate the foundational reliability of STRmix v.2.5.11 and v2.8.0.**

The developmental validation proffered here is akin to a test of an airbag where the manufacturer says, "we crashed this car against a lot of different kinds of walls at different speeds and the airbags performed as expected, which was good." How do we know the right speeds were tested? Are they the speeds cars actually travel at in the regions this car will be driven? Are the walls representative of the objects a car might crash into in these regions? Are the manufacturer's expectations the right ones? What about the expectations of the drivers and passengers and the legal system and the insurers? And how can we know how well the airbags worked? Does "good" mean there were zero injuries? Zero serious injuries? When is an injury serious? And how many crash tests would you want before you felt comfortable getting into a car?

Not only is the DNA testing inadequate, but the reliability of safety critical software is not established through this kind of testing only. It is established through IV&V. That is true across industries; it is a well-established standard that simply was not followed by STRmix's developers. Before cars with an airbag system are deployed on the road, software governing airbag deployment is subject to rigorous IV&V. Crash testing is not enough. The software testing is necessary too—and it has to be rigorous and it has to be independent. Both kinds of testing are necessary to say the risk of error is sufficiently reduced for the cars to be on the road or someone

to go to prison for the rest of their lives. STRmix does not meet those standards. The evidence must be excluded.

**B. The Likelihood Ratios produced by STRmix are not appropriate to analyze mixtures that contain related contributors and are inadmissible.**

As both Dr. Coble and Mr. Inman explained, the LR produced by STRmix is inappropriate for mixtures in which contributors are related. Finding of Fact F.8.e.ii. Even if the calculations STRmix runs to reach the genotype weights are reliable, the final output—the LR—are not. This is a bar to admissibility because the final opinion, the LR, does not reliably stem from the underlying data. As the Supreme Court of the United States explained in Daubert, this is a an issue of "fit." Fit arises because the "scientific validity for one purpose is not necessarily scientific validity for other, unrelated purposes." Daubert v. Merrell Dow Pharm., Inc., 509 U.S. 579, 591 (1993). For instance, a test that is validated to discriminate between drivers above and below a statutory BAC limit was found not to be validated as a measure of impairment, and is therefore not admissible for that purpose. State v. Lasworth, 42 P.3d 844, 848 (N.M. Ct. App. 2001). See also Commonwealth v. Davis, 168 N.E.3d 294, 304 (Mass. 2021) (holding that even where GPS monitor was known to be reliable for determining wearer's location, its data was inadmissible to prove wearer's speed or movements without showing of reliability for that purpose), Similarly, even if STRmix's LRs are reliable as an interpretation of mixtures with unrelated contributors, the undisputed evidence adduced at the hearing is that it is not reliable in a different context: the interpretation of mixtures with related contributors. Therefore, an LR produced by STRmix for such a sample is not reliable and never admissible in court.

**C. The State has failed to demonstrate that STRmix was used reliably in this case.**

Regardless of whether STRmix v2.5.11 or 2.8.0 are reliable as a general matter, the State also has to prove that the samples in this case were analyzed reliably. Every technique, no matter how reliable, can be misapplied. Therefore, this Court's job is also to make "case-specific determinations about the expert evidence—such as whether the witness has sufficient expertise, whether the evidence can assist the trier of fact in that case, and whether the relevant theory or technique can properly be applied to the facts in issue." Olenowski II, 255 N.J. at 581. In other words. to pass muster under N.J.R.E. 702, the technique must be valid as applied, "mean[ing] that the method has been reliably applied in practice." PCAST Report at 5. That means that the examiner "must have been shown to be capable of reliably applying the method and must actually have done so." Id. at 6 (emphasis in original). Although some concerns with expert evidence go to weight and not to admissibility, when an expert has failed to follow minimum standards for the reliable application of a technique, that is an issue that goes to admissibility, not weight. When faced with a proffer of expert testimony, as here, it is the trial court's duty to act as a gatekeeper and ensure that "expert witnesses demonstrate that they have reliably applied [their] methodology." Olenowski II, 255 N.J. at 616.

The emphasis is on the reliable application of the methodology, not whether the results may be otherwise considered reliable. If a well-done fingerprint analysis concluded someone's fingerprint was on the murder weapon, a psychic's opinion wouldn't be admissible just because they agreed.

**1. Internal validation studies establish the limits of what is admissible in court.**

As the FBI QAS, SWGDAM, ANSI/ASB 18, and every expert that testified other than Dr. Buckleton agreed, the internal validation of a laboratory should establish the limits of what a laboratory reliably analyzes. Finding of Fact F.8.d. Therefore, as a matter of law, whatever was not sufficiently tested by a laboratory in its internal validation or was tested and not the results

were not reliable, is not admissible in court. This is a hard limit that goes to admissibly and not to weight.

Olenowski requires the reliable application of a reliable technology. By their own standards and statements, internal validations are a mandatory way laboratories demonstrate that any given application of a technology could be reliable (assuming the individual analyst applied it reliably in this case). See also Mixture Interpretation at 16 ("When assessing the degree of reliability of DNA mixture results for a specific case, the assessor (e.g., an expert user of the results) needs to have access to validation data from known samples that are similar in complexity to the sample in the case."). In contrast, when a judgment call could fall within the boundaries of a validation study, that would go to weight and not admissibility—the decision to run an arguably 2-person mixture as a 3-person mixture by a laboratory that internally validated up to 4-person mixtures would be a matter for cross-examination, not admissibility. But when a laboratory reports a result on a sample that it has not demonstrated it can reliably analyze, that goes to admissibility. For instance, all the analysts who testified at the hearing said they would not analyze a 5-person mixture because the internal validation went only up to 4. What cannot be reliably analyzed cannot be reported. With no logical or scientific distinction presented as to why number of contributors tested by the study is a hard limit and mixture portion or template amount is not, all of these factors must be treated as a hard limit.

As discussed above, just because errors are inevitable as a matter of DNA science does not mean laboratories cannot recognize where the risks of error are high and should not mitigate the risk of error in their own casework and reporting. According to Dr. Buckleton's own writing and others who use STRmix, STRmix does worse under predictable circumstances: the more complex the DNA, the higher the chance of error. For instance, the FBI internal validation

summary discussed above noted that "[i]n all cases where non-contributor comparisons generated HPD LR>1," that is, in all cases false positives were reported, "the results were consistent with scientific expectations given the mixture quality and complexity (e.g. degradation, allelic dropout) and number of contributors." Moretti et al. at 143. The solution then, is clear: not to analyze or report all DNA profiles and accept predictable error when more complex samples are analyzed, but to determine at which level of complexity the rates of error are high and not analyze those kinds of samples. The point of an internal validation is to determine which kinds of samples are just too risky to analyze. To offer another analogy, if the maker of a scale said that it was proven to be reliable up to 300 pounds but would predictably, through no fault of its own, become less reliable after that, the answer is not to weigh things over 300 pounds. The answer is not to weigh things over 300 pounds and simply accept that there will be error. And if an attempt was made to report that that scale weighed something and reported that it weighed 500 pounds, that conclusion would not be reliable enough to be admitted in court.

Other courts have appropriately held that when the State fails to demonstrate that "STRmix performed an analysis that it was validated to perform[,]" the results are inadmissible. Order Granting Motion to Exclude DNA Evidence, United States v. Francisco Ortiz, Case No.: 21-CR-2503-GPC at 15 (S.D. Cal. June 10, 2024) (D-1213). In that case, the court excluded the results of a possible 6-person mixture because neither the developmental validation, nor the FBI internal validation, nor the relevant laboratory's internal validation contained any analysis of six-person mixtures. Id. at 17. The court rejected the State's argument that the question of how many contributors there were went to weight not admissibility, when the laboratory had validated certain 5-person mixtures and not 6-person mixtures, because validation is a necessary prerequisite to admissibility. Id. at 14-18. See also Memorandum and Order, Walker v. New

York, No. 14-cv-680(NRM)(PK) (E.D.N.Y. September 16, 2024) (excluding the results of a different PGS system and technique for analysis of low-template DNA where there was no evidence in the validation studies that demonstrated that system and technique could yield reliable results on samples containing less than 25 picograms of DNA) (D-1215).

The analysis of the samples in this case is not supported by the internal validation studies for two reasons. First, the internal validation study summaries do not contain enough content to assess the reliability of STRmix in these laboratories on any samples. Second, insofar as one would consider the limits of the testing in the studies to be the limits of reliable analysis, the samples in this case are beyond those limits.

**2. The State has not established the appropriate limits in this case due to validation summaries that are conclusory and insufficiently detailed. Therefore, no evidence from either lab is admissible.**

The internal validation summaries do not provide any information about the error rates when using STRmix in NJSP and Bode over different kinds of samples. Finding of Fact H.6. What we have instead is generic statements from the developers of STRmix, from Bode, and from NJSP acknowledging that errors occur, but they personally think STRmix is reliable. That is insufficient because as the gatekeeper this Court has to determine the reliability of the software. It can't just take the word of the people who make it, sell it, buy it, and use it to prosecute people.

None of the error rates discussed in any other validation can stand in for error rates for STRmix error rates for Bode or NJSP because each laboratory is different. Finding of Fact F.8.d. The State has provided none of these error rates for any of the relevant versions of the software or software as used in these specific labs. Without this information, STRmix cannot be used.

The fact that Dr. Reich agreed that the high LRs for the major contributors comported with his opinion that the major contributors would be included in the sample through manual

analysis does not change the requirement that a validation study provide sufficient information to assess the reliability of the use of STRmix in a laboratory access all sample types for two reasons.

First, there is no "right" LR and the LR is being admitted: "The correctness of that LR is important and there's no way to know that." (14T 163-15 to 164-20) (emphasis added) If Bode or NJSP want to use traditional DNA analysis and report an RMP that is manually verifiable according to its standard protocols, it would be welcome to try to do that. The potential for an RMP result that also reaches a strongly inclusionary result does not mean any given LR is admissible.

Second, just because another method could reach a similar answer does not mean that the answer reached in this case is the product of a reliable methodology reliably applied. Reliable application is a question of procedure, not of an independent assessment of the outcome. As Dr. Reich testified to, STRmix is not needed to analyze the major contributors in most of the samples Bode analyzed. (12T 212-7 to 14) Again, Bode or NJSP is welcome to try to analyze these samples using traditional DNA. The RMPs that result may very well be strongly inclusionary. That does not mean that these STRmix results are admissible.

3. **In the alternative, testimony about STRmix results is admissible only if the sample tested falls within the range of samples tested in a laboratory' internal validation study. Because in this case the samples tested are more complex than those in the validation summary, the results of those tests are inadmissible.**

The internal validation summaries, as explained above, are insufficient to demonstrate that the laboratories can use STRmix reliably at all. The summaries demonstrate that STRmix was tested on certain kinds of samples, but not what the results were. Taking a test and passing a test are two different things.

There are limits to what any technology can do. Those limits have to be known and adhered to. No one would (credibly) argue that a laboratory that has tested only two-person mixtures could test twenty-two-person mixtures. But scientific standards demand that those limits come from a laboratory's demonstration of what it can reliably analyze, not an individual analyst's feelings about what she can reliably analyze. But even assuming that the results of the samples tested passed a reliability test—an assumption that the State, which bears the burden of proving reliability cannot benefit from—the samples analyzed in this case are beyond any of the tested limits.

**a. All of the results are inadmissible because they all involve analysis of mixtures that involve related people.**

Neither validation summary demonstrated that either laboratory can reliably analyze samples that consist of related contributors or can reliably analyze a sample when the true contributor might be related to the person of interest. Finding of Fact H.1.a.v and H.3. In fact, neither laboratory tested any such samples at all. All of the potential contributors analyzed in this case are related to at least two other potential contributors, and most of them are related to three, and the mixtures are hypothesized to be of related people. The State has not established that either STRmix in general, or Bode or NJSP as they applied STRmix in the case, can reliably handle so many related contributors or that it did so according to standards generally accepted in the scientific community.

Analysts should not have even run these samples because they know there is no evidence that STRmix can reliably interpret them, even if STRmix is working perfectly. "[I]t is the old garbage in, garbage out[.]" (16T 71-7)

Analyzing related contributors is the hardest part of DNA analysis. That is true whether using traditional or PGS systems. The risk of error is high. The only responsible way to move

forward is to test how often a laboratory gets it wrong when relatives are considered in their validation study, the conditions under which the laboratory gets it wrong, and to make sure that the laboratory does not attempt to run samples where the risk of unreliability is too big. Neither Bode nor NJSP did that. They did not demonstrate that they can reliably analyze samples potentially made up of multiple related people. Every single sample analyzed by both laboratories in this case must be excluded.

**b. Samples E02b1, E03b1, E04a1, E06a1, and E07a1 are inadmissible because they are samples at or below the limits of what was tested by Bode in its internal validation summary and are comprised of related people.**

Bode analyzed five samples that in addition to being mixtures comprised of related people were at or below what was tested in the validation summary. Findings of Fact H.1.a.i. and H.1.a.ii. There is an insufficient basis to believe Bode can reliably analyze these samples.

It is true that Bode's contributor ratios differed from STRmix's in its validation study; in some cases, the disparity was enormous. But uncertainty in the actual percentage or amount of DNA deposited by a given contributor means Bode needs to be more conservative in what results it reports, not less. Moreover, there is a greater concern and a need for more data because there is more uncertainty in the template amount and contributor ratio measurement at low template. 12T 91-12 to 92-15 (when testing samples "at the very limit of what Bode did" in its validation study, there is more cause of concern "because of the variability in the measurement and these are not measured twice and cut once. They're just measured once. We don't know whether that's an accurate template from the quantification but the quant is variable. And so we don't know what the actual amount of DNA is. Are we just under? Are we under by twice, three times? Are we above it and we're not aware of it? We don't have a validation study that goes beyond this so that we're more sure that this sample conforms.")

To be clear, analysts do not know what mixture proportion or template amount STRmix will produce before it runs a sample through STRmix. But just because it runs a sample does not mean it has to report an outcome. When STRmix produces an output below what the laboratory has demonstrated it can reliably analyze, the analyst can and should simply report that the sample was not suitable for STRmix analysis in its laboratory. That is the truth.

**4. The State has not established that these analysts can or did reliably use STRmix.**

The State has not established that either Ms. Reed or Ms. Schlenker can or did reliably use STRmix in this case. It is not that they don't have the expertise on paper, it's that their report and testimony in this case cast grave doubt in their ability to reliably apply STRmix to this case.[2] Ms. Reed's use of STRmix is particularly concerning given Ms. Reed's inappropriate visual exclusions and lack of documentation about her work. The vague and subjective SOPs do not help channel her discretion, Finding of Fact H.2., and the case-specific information she had before she began her analysis raise a significant specter of cognitive bias. Neither Ms. Reed nor Ms. Schlenker demonstrated a thorough understanding of how STRmix works or how the diagnostics operate. (5T 99-6 to 105-2; 9T 101-13 to 19) Ms. Schlenker testimony that she does not know what "ground truth" means in DNA (9T 113-1 to 7), a term Dr. Coble testified that everyone analyst should understand (10T 32-10 to 21), is also quite concerning. Without any proficiency testing, let alone proficiency testing on samples similarly complex to the ones in this case, there is no basis to believe that either analyst reliably applied STRmix in this case.

**5. The results are reported in a manner incompatible with the reliable use of STRmix.**

[2] Mr. Caneiro reserves the right to challenge whether these analysts possess sufficient expertise to render an opinion, factor (2) under Olenowski, if this motion is denied.

The results reported by these laboratories is incompatible with the reliable use of STRmix for four reasons. These results cannot be admitted

**a. Bode should be using familial or unified likelihood ratios to consider whether the real contributor could be related to the person of interest.**

Bode did not validate and did not use the familial or unified LRs. Finding of Fact H.1.a.vi. That is the appropriate LR to use when the hypothesis is that a related person is the true contributor to a mixture (e.g. rather than Paul Caneiro the true contributor was his brother or niece or nephew). Finding of Fact F.8.e.ii. That the true contributor is unrelated to any of the Caneiros is also a relevant hypothesis. But it is not the only one. The inability to render an LR that relates to an important hypothesis renders Bode's results inadmissible.

**b. The verbal scale is misleading.**

As found above, Finding of Fact F.11., the verbal scale is at worst baseless and at best misleading. Its use would violate N.J.R.E. 403.

**c. New Jersey State Police's use of STRmix to render a source attribution is inappropriate.**

NJSP should not be converting a genotype probability from STRmix into a definite statement of source attribution. NJSP used STRmix to begin its analysis of 6-1-4-1. It used STRmix to deconvolute the sample. Then, for loci that STRmix was 99% sure about its estimate for the alleles, NJSP generated a chart. NJSP then compared that chart manually to [redacted] and declared that [redacted] is "[redacted] is identified as the source of the STR DNA profile of Unknown Male." This statement is inappropriate for two reasons.

First, this violates well-accepted standards within the forensic science community for how experts should discuss their findings. For instance, Department of Justice guidance to forensic scientists provides that experts "shall not assert likelihood ratio of any magnitude provides an absolute identification or source attribution of a known individual to an evidentiary

sample." Department of Justice, Uniform Language For Testimony And Reports For Forensic Autosomal DNA Examinations Using Probabilistic Genotyping Systems, at 4 (Sept. 2019) (D-1214). It is well-established that "[i]n the legal context, an opinion on source should be made by considering all the evidence in the case as well as the consequences of the decision. This task is the responsibility of the factfinder, not the DNA analyst." Human Factors at 78. STRmix can report how likely it believes it is that specific alleles are represented at each locus. What to make of that is up to the jury, not the analyst.

Second, STRmix does not deal in absolutes. It deals in probabilities. It cannot tell you with certainty that any specific genotype is the genotype in the sample. It can give its confidence that that is true. Dr. Buckleton explained this at the hearing: "the key" to probabilistic genotyping "lies in the word probabilistic. So instead of saying yes or no, say yes, this. This is a genotype from this mixture or no, which you can think of is naught or one. It assigns a probability to the evidence given that genotype." (6T 16-1 to 9) Dr. Buckleton also specifically tested that he "wish[ed] people wouldn't use" the genotypes that STRmix believes has a 99% of being the genotype of the contributor to create a static profile. (7T 84-11 to 21) It is misleading, in violation of N.J.R.E. 702, to use STRmix to determine what the most likely genotype is and then declare it as a fact and not a probability.

**6. The STRmix results are inadmissible in this case.**

All techniques have their limits. There is no technology that is appropriate to use at all times in all situations. Yet Bode and NJSP have refused to find those outer limits to guide their use of STRmix in their laboratories, despite the consensus in the scientific community that establishing those limits is paramount. Moreover, insofar as those limits can be inferred by the validation summaries or SOPs, the use of STRmix in this case exceeded those limits.

"Forensic scientists need to know 'when to punt'—that is, when to decline the opportunity to move forward with questionable or problematic evidence." Thompson, Uncertainty in Probabilistic Genotyping at 14. The scientists in this case did not know when to decline the opportunity to analyze problematic evidence in this case and did not decline. The result is unreliable DNA evidence that must be excluded.

## CONCLUSION

"Properly exercised, the gatekeeping function prevents the jury's exposure to unsound science through the compelling voice of an expert . . . . Difficult as it may be, the gatekeeping role must be rigorous." In re Accutane Litig., 234 N.J. 340, 346, 390 (2018). The State seeks to admit the most compelling form of evidence at a criminal trial—DNA evidence—that was generated by a piece of software that has not been tested against the most fundamental software engineering standards, for which no error rates are provided, and whose use is pushed by the group of people who developed it as well as by the laboratories that have paid for it. That is not enough to establish admissibility.

Moreover, there is no evidence that the laboratories in this case can reliably analyze the samples in this case. The DNA samples at issue in this case have the classic hallmarks of complex DNA: multiple related contributors, low quantities of DNA, and a low percentage of DNA contributed by the minor. Whatever the abilities of STRmix are, it cannot be assumed that, as used in these laboratories, it can handle these kinds of samples. To the contrary, the State bears the burden of clearly establishing reliability; it gets no benefit of the doubt. The State has failed to make that demonstration. The evidence must be excluded.

Respectfully submitted,

TAMAR Y. LERER
Public Defender
Attorney for Defendant-Appellant
BY: _____
Deputy Public Defender
Attorney ID No. 063222014