Deadline for Comments: 8/26/2019

ASB Standard 062: Standard for Topography Comparison Software for Toolmark Analysis, First Edition

Note: a specific Proposed Resolution must accompany each comment or it cannot be considered.

#	Section	Type of Comment (E-Editorial, T-Technical)	Comments	Proposed Resolution	Final Resolution
1	3	T	Section 3 (Definitions) fails to include the concept of "specificity", i.e., the system's ability to detect a true exclusion. When a system "fails to exclude" two toolmarks that arise from different sources, whether because it incorrectly "identified" or the result was "inconclusive", this impacts the specificity of the system. Specificity is a key component of validation. Relatedly, there is no discussion of the concepts of "inconclusive" or an "inconclusive range" in the definition section or anywhere else in the document. The document as a whole fails to discuss in subsequent sections how to assess system performance at excluding a source and how to address test results where no conclusion is reached but the answer is known.	Definitions should be developed for "specificity" (which includes the concept of "failure to exclude") and "inconclusive" and the document s+E21	Accept with Modification: We added "specificity' to section 3 and section 4.3.3.1. Many statistical measures exist and cannot all be included in this document.
2	4.2.1.1	T	This section includes the phrase "degree of certainty". To our understanding, a measure of uncertainty may be derived from these systems, but "degree of certainty" is not. Statistical measures of confidence are a separate concept, and terms should not be used loosely.	Delete reference to "degree of certainty".	Reject: The document uses the term certainty as a general term to cover multiple statistical models (e.g."probability" or "confidence" or "likelihood"). Also in this section this term is provided as an example as to what the numerical measurement could be.
3	4.2.2.2	T	Section 4.2.2.2 (Use of Category 0 Software) states that "one may only state that the pair of cartridge cases ranked in the top Y of the database search (where Y may be a number, e.g., 10, 20). If asked about the significance of this statement the only acceptable response shall be that there is no statistical confidence established for any match results for a rank-scores only, non-statistically validated scoring function." There are two concerns with this statement. First, to avoid misleading the audience the examiner should not have to be asked before stating the critical limitation to this result -- "that there is no statistical confidence established for any match results for a rank-scores only, non-statistically validated scoring function." Second the term "match" should not be used in connection with a "result". To the lay audience match implies a match to the exclusion of all others. "Results" needs no modification and the addition of "match" suggests an unsupported level of certainty.	The two sentences should be combined and redrafted as follows: "One may only state that the pair of cartridge cases ranked in the top Y of the database search (where Y may be a number, e.g., 10, 20) and one must explain that there is no statistical confidence established for any results for a rank-scores only, non-statistically validated scoring function."	Accept with Modification: The term "match" was removed in section 4.2.2.2, 4.2.3.1, 4.2.3.3 and 4.2.4.2. This document is not intended to guide the attorneys about their questions.
4	4.2.3.3	T	Section 4.2.3.3 (Use of Category 1 Software) states that "one may state that a pair of cartridge cases has a match score of X, which is very high. If asked about the significance of this statement for software at Category 1, the only acceptable response shall be that there is no statistical confidence established for any match results for a non-statistically validated scoring function." There are three concerns with this statement. First, to avoid misleading the audience the examiner should not have to be asked before stating the critical limitation to this finding -- "that there is no statistical confidence established for any match results for non-statistically validated scoring function." Second the term "match" should not be used in connection with any "score" or "result". As explained above to the lay audience this implies a match to the exclusion of all others or an unsupported level of certainty. Third, the qualifier "very high" is subjective, has no statistical basis, could be used inconsistently, and has the potential to mislead the fact-finder, or be unintentionally misused by the actors in the legal system. Therefore, the score should stand on its own.	The two sentences should be combined and redrafted as follows: "...one may state that a pair of cartridge cases has a score of X but must explain that there is no statistical confidence established for any results for a non-statistically validated scoring function."	Accept with Modification: The terms "very high" and "match" were removed in section 4.2.3.3. This document is not intended to guide the attorneys about their questions.
5	4.2.4.2	T	Section 4.2.4.2 (Use of Category 2 Software) states "the study demonstrated that a match score of X indicates a false match probability of Y." As with the sections above match should not be used when describing a score or result as it implies that it is a match to the exclusion of all others.	Redraft the sentence as follows: "The study demonstrated that a score of X indicates a false match probability of Y."	Accept
6	4.3.1	T	What is "an organization with the appropriate knowledge and/or expertise" for developmental validation? The absence of specifics renders this standard a non-standard and yet this is a critical aspect of sound developmental validation	Define appropriate knowledge and expertise and submit the definition for public comment.	Reject:The document uses standard definitions. Appropriate knowledge and /or expertise means that the organization has the ability to satisfactorily perform the function being requested. This expertise is application specific.
7	4.3.1	T	The suggestion that software upgrades that can affect scoring function functionality (such as a major version upgrade) only "may" require additional validation is problematic. Absent rigorous testing one cannot assert with confidence that software upgrades will not have unintended consequences.	Clarify that software upgrades shall be subject to developmental and deployment validation.	Acept with Modication: The sentence was edited to read "Software upgrades that affect scoring function functionality shall require deployment validation."
8	4.3.3.1	T	This section does not explicitly call for a requirement for peer-reviewed publication in a journal widely available to the larger scientific community. Peer-review and the subsequent larger post-publication review by the scientific community is the hallmark of sound science. In the absence of a regulatory body, normal market forces, and the multi-disciplinary nature (e.g. statistics, computer science, research design) of the field of forensic science, rigorous and transparent developmental validation is an essential component of the development of accurate and reliable comparison software.	Include an additional requirement for peer reviewed publication in a widely available journal.	Reject: All journals are widely available online. The document states that validation shall follow the referenced implementation document 063 which spells out scale/scope and requirements. For example that developmental validation shall be described in a peer-reviewed publication.
9	4.3.3.1	T	While a broad selection of makes and models and substrate types is important it is equally important that the system (1) be tested on closely related items (e.g. consecutively manufactured items), and (2) be tested on lower-quality toolmarks (e.g. damaged or poorly marked bullets or casings), in order to fully understand the limits of the technology. Further, the results should be reported separately (e.g. an FPR for closely related items as well as an FPR for items produced by different manufactures).	Include language requiring testing of closely related items and lower-quality toolmarks, and for reporting separately on different classes of testing. Additionally, include reference to these in Annex A.	Accept with Modication: The current text states the the study shall be robust. You are allowed to develop a study with fragments, damaged, or consecutively manufactured. Additional sentence was added "The validation set should therefore reflect the range of toolmark types expected to be seen in casework." It is not possible for the document to spell out every scenario.
10	4.3.3.1	T	As discussed above (discussion of section 3), the concept of specificity (to include "inconclusive" results) must be addressed in connection with validation (both developmental and deployment).	Include requirement to assess specificity during validation and address these results in protocol development (e.g. by establishing a validated "inconclusive range").	Accept with Modification: We added computation of "specificity" to the last sentence of 4.3.3.1
11	4.3.3.1	T	This section includes the phrase "degree of certainty". To our understanding, a measure of uncertainty may be derived from these systems, but "degree of certainty" is not. Statistical measures of confidence are a separate concept, and terms should not be used loosely.	Delete reference to "degree of certainty" (if appropriate, replace with "statistical confidence level").	Reject: The document uses the term certainty as a general term to cover multiple statistical models (e.g."probability" or "confidence" or "likelihood").
12	4.3.3.2	T	As discussed above (discussion of section 3), the concept of specificity (to include "inconclusive" results) must be addressed in connection with validation (both developmental and deployment).	Include requirement to assess specificity during validation and address these results in protocol development (e.g. by establishing a validated "inconclusive range").	Reject: Specificity is not likely to be established during deployment validation as the study is too small. Deployment validation establishes that a developmentally validated technology is working as expected in-house.

#	Section	Type of Comment (E-Editorial, T-Technical)	Comments	Proposed Resolution	Final Resolution
13	4.5.1	T	Section 4.5.1 states that "Statistical models shall output a statistically grounded metric indicative of whether or not two toolmarks have a common origin or support for a common origin." The output of a statistical model provides support for one explanation (common origin) versus another (different origins); it does not actually indicate which explanation is true.	Redraft the sentence as follows: "Statistical models shall output a statistically grounded metric indicative of the level of support for a common origin."	Reject: The comment applies if the statistical model is a likelihood ratio; however, as written the text is consistent with multiple statistical models.
14	4.6.2	T	Section 4.6.2 states that "conclusions regarding common origin or statements of weight of evidence shall follow the standard operating procedure of the laboratory." This statement is inconsistent with the stated purpose of ASB Draft Standards 061, 062 and 063 as set forth in the forward to the documents – "The purpose of these standards is to ensure the production of reliable data and statistically based conclusions . . . ." Anything beyond the statistical statements included in this document should be addressed in a separate document that has been through the notice and comment process, not by reference to laboratories' protocols.	The referenced sentence should be deleted from this standard	Reject: This statement improves readability and states that lab SOPs should be followed and that the statistical models described in this standard can be used to support those conclusions.
15	4.6.2	T	This section includes the phrase "degree of certainty". To our understanding, a measure of uncertainty may be derived from these systems, but "degree of certainty" is not. Statistical measures of confidence are a separate concept, and terms should not be used loosely.	Delete reference to "degree of certainty".	Reject: The document uses the term certainty as a general term to cover multiple statistical models (e.g."probability" or "confidence" or "likelihood").
16			Suggest "the only acceptable response shall be that a statistical confidence is not established" or similar in place "the only acceptable response shall be that there is no statistical confidence established". The latter sounds potentially biasing.		Reject: Proposed change is equivalent to existing text.
17			My comments, as they have previously, apply to all three topography and comparison software related standards (61-63). While all three standards clearly represent an admirable first step towards greater objectivity in the field of firearms examination, one clearly earned through concerted and well-intentioned labor, a wealth of lingering concerns nevertheless prevents me from voting in favor of any of the three. My main points of contention (though not all encompassing of my objections) are as follows:		This comment was broken down into 6 sections as following and a resolution for each section has been provided.
17 (17.1)			(1) Software Engineering Standards: Outside of the forensic sphere, the Institute of Electrical and Electronics Engineers (IEEE) has developed a whole series of Standards meant to define best practices for the development of software systems, most relevant among them being IEEE, "Std. 2012-2016 Standard for System, Software, and Hardware Verification and Validation," (2016). And this body would be unwise to ignore the consensus guidelines of the world's largest and most prestigious software body. In fact, multiple groups who have addressed probabilistic genotyping systems in the DNA sphere have arrived at precisely that conclusion. See United Kingdom Forensic Science Regulator, "Guidance: DNA Mixture Interpretation Software: Validation, FSR-G-223," (2017); M.D. Coble et al., "DNA Commission of the International Society for Forensic Genetics: Recommendations on the validation of software programs performing biostatistical calculations for forensic genetics applications," 25 For. Sci. Int'l Genetics 191, 192 (2016); Nathaniel Adams et al., "Letter to the Editor- Appropriate Standards for Verification and Validation of Probabilistic Genotyping Systems," 63 J. For. Sci. 339 (2018). Yet, despite the significant problems caused by the DNA community's late adoption of best practices for software, Standards 61-63 make no effort to track IEEE's guidance.		Reject (1): The IEEE standards are intended for developers while these standards are intended for end users. The software described in these documents can be empirically tested on real-world data. Software performance is important and the document describes three stages of validation testing to ensure that the software meets the needs of the end user.
17 (17.2)			(2) Independence: One of the gaps between IEEE and Standards 61-63 actually brings me to my second concern, that of independent validation. Developers are by definition self-interested and until these Standards require full developmental style validation occur by independent researchers they fall short of what the legal and forensic communities deserve.		Reject (2): This falls outside of the scope of this document as these requirements are for the end user and not the developer. The end user must validate the technology prior to use. The strength of the validation is a function of the scale/scope and quality of the validation completed. Larger more independent validations are more substantial than smaller or less independent validations.
17 (17.3)			(3) Laboratory Validation: To the extent laboratory implementation validation might plug this gap, however, Standards 61-63 do not lay 0out rigorous enough requirements, most centrally because they offer little in the way of the qualifications required of those performing such work. Are we really to believe that a science bachelors degree suffices for lab personnel necessarily to have the background in statistics, computing, and experimental design necessary to catch flaws in these new systems?		Reject (3): The strength of a validation is affected by several factors, these include scale/scope as well as who conducts the validation. The documents state that development validation shall be peer reviewed and publicly available which means that one or more peers would provide an independent check. The document states that development and deployment validations shall also be evaluated by a technical reviewer. The document states that validation must be performed by someone with appropriate knowledge or expertise. Laboratory personnel are self-motivated to find someone appropriate. Otherwise the laboratory may have problems down the road.
17 (17.4)			(4) Transparency: Whatever a company might be able to claim in the private sphere in terms of patents, trade secrets, etc… such proprietary incentives are a poor fit for the criminal justice system. While it might* go to far to demand open source programs, these standards do not go as far even as has the DNA community. It should not be enough to publish the scientific principles of a system. At minimum all algorithms used must be public and peer-reviewed as well. Every major probabilistic genotyping system has managed at least that much, and the firearms examination community cannot afford to do less.		Reject (4): This falls outside of the scope of this document as these requirements are for the end user and not the developer. The document states peer-reviewed publication of the scientific principles shall be required. As described above, firearm and toolmark examination algorithms can be tested empirically using large real-world test sets that are representative of actual casework. The strength of the validation is a function of the scale/scope and quality of the validation completed.
17 (17.5)			(5) Reporting Language: Standard 62's treatment of reporting results is disturbing and likely to cause real difficulty in the courts. In allows for use of the word match when describing Category 0 scores in the false assurance that the fairly confusing statement about statistical significance that follows will cure the prejudice associated with the word match. And it allows analysts to testify to "high" match scores in Category 1 as if such statements will not be interpreted incorrectly by jurors. Even under category 3 it does not require any explanation to jurors of the underlying probability and statistical principles associated with the program (is it Bayseian, frequentist, do they need to take into account a prior probability or base rates???)		Accept with Modification (5): The term "match" was removed in sections 4.2.2.2, 4.2.3.1, 4.2.3.3 and 4.2.4.2. The terms "very high" and "match" were removed in section 4.2.3.3. Regarding explanation to jurors of underlying principles, the document is not intended to guide attorney questioning.

#	Section	Type of Comment (E-Editorial, T-Technical)	Comments	Proposed Resolution	Final Resolution
17 (17.6)			(6) Validation Samples: Especially considering the massive role left to labs to validate and implement the systems described in Standards 61-63, those documents provide grossly insufficient guidance as to the types of samples necessary to particular validation studies opening these technologies up to abuse. While it may not be possible to exhaustively list all possible sample variations, these standards can at least set a floor or minimum and describe what they are doing as such. That would allow researchers to do more but would not allow labs or developers to get away with doing less (ie failing to run any known subclass samples).		Accept with Modification (6). The current text states the the study shall be robust. You are allowed to develop a study with fragments, damaged, or consecutively manufactured. An additional sentence was added "The validation set should therefore reflect the range of toolmark types expected to be seen in casework." It is not possible for the document to spell out every scenario. An example validation study is described in Annex A.
18			As with my previous comments, I am writing with respect to the three standards, and as before I believe these standards are a very good, thoughtful start but need more fleshing out to give guidance to practitioners. 1) I agree with Richard about the failure to cite to or address the IEEE standards. Frankly I don't know enough about contents of the IEEE validation standard to know what the precise discrepancies are, but I believe it should be treated as a normative reference (sec. 2) for all three standards. 2) Who should conduct validation: a. For developmental validation, the documents say it should be conducted "by an organization with appropriate knowledge and/or expertise." What does this mean? What constitutes "appropriate" knowledge and/or expertise? This should be more precisely defined. Also, studies by organizations/instituti-ons independent of the developer should be required. b. For deployment validation, the docs say it should be conducted by someone with minimum of a bachelor's degree w/ a science major. If the requirements of deployment validation were set out in great detail, I might not be concerned with this fairly low bar. However, this is not the case. Instead, four one-line "aspects [that] shall be documented" are set forth, and otherwise the persons/entities conducting the study are given unfettered discretion on study design. As I've said before, I think the validation aspect of these documents needs to be bulked up significantly to give practitioners meaningful guidance. For example, there is no mention at all of concepts of sensitivity or specificity, or of the idea of designated inconclusive ranges—all concepts essential to the actual deployment of this technology/software. If additional clarifying detail regarding the design of validation studies is not going to be added to this document, I would at very least make it clear these documents do not provide a formula for adequate validation, and require the validator to consult with a statistician or someone with expertise in study design. 3) Sample sets for validation: As I mentioned in my last set of comments, I don't think these documents provide sufficiently specific guidance about what kinds of samples have to be included at minimum to make the studies adequate. Certainly, the example given in Annex A should include samples that test the limits of the system, e.g. damaged and poorly marked ammunition, toolmarks left by consecutively manufactured firearms, etc. 4) Preservation and disclosure of data: I'd like to see some language in these documents emphasizing the importance of independent review of data, and requiring validators (whether involved with developmental, deployment, or performance checks) to maintain data and make it available upon request. 5) Reporting and testimony: Throughout these documents, I'd like to eliminate the word "match", which has an unavoidable connotation of absolute source attribution. For example, I would suggest substituting "match score" with "similarity score" or simply "score". Further, I have real concerns with analysts being permitted to report that a similarity score is "very high" using Category 1 software given "there is no statistical confidence established" for these scores. Finally, I'd like the limitations of these methodologies (e.g. the fact that "there is no statistical confidence established" for certain results; measures of uncertainty) be a required component of reporting and testimony.		Reject (1): Please see above resolution of comment # 16 Part 1. Reject (2a): Appropriate expertise is application specific. Reject (2b): The quality of the validation is related to the quality of those conducting the validation and the study design; the comment that a formula is not presented is already covered, the need to seek out those with appropriate expertise is explicitly mentioned. Discussion of specificity has been added (see comment #1), note that not all statistical measures can be mentioned in a single standard. Reject (3): The current text states the the study shall be robust. You are allowed to develop a study with fragments, damaged, or consecutively manufactured. An additional sentence was added "The validation set should therefore reflect the range of toolmark types expected to be seen in casework." It is not possible for the document to spell out every scenario. An example validation study is described in Annex A. Reject (4) Standard mentions documentation and peer review; documentation preservation should follow lab protocols for all other laboratory equipment; no need to restate here. Accept with Modification (5): The term match and high have been removed from relevant sections. See comment #16 part 5.
19			In light of the objections raised by Richard Gutierrez, I would urge further study regarding the discrepancies between these Standards and the Standards of the IEEE. I also do not believe that quantification of results can be justified in the abstract scientifically since they would have to be evaluated on a case by case basis, requiring the showing of a sufficiently robust comparison database as well as proficiency in testing. This is not something that can be delegated. Furthermore, comparison analysis is a study, at best, resulting in degrees of uncertainty. This is true of all science. The degree of uncertainty is not something to be conveyed in mathematical terms to a jury since, 1) there is insufficient data to justify it in FATM analysis foundationally or as applied in a given case; and , even if that were established, 2) mathematically expressing an opinion as the weight of evidence is inconsistent with the jury's heuristic means of metaphorically weighing "all" the evidence and comparing that metaphorical weight to a metaphorical standard, e.g., proof beyond a reasonable doubt.		Reject: Discrepancies between this standard and IEEE are adressed above in previous comment resolutions. The rest of this comment is too general and outside of the specific scope of this document. This standard advances the science of this discipline with respect to quantitative methods.