Fundamental issues in biometric performance testing: A modern statistical and philosophical framework for uncertainty assessment

James L. Wayman[♪], Antonio Possolo[♫], and Anthony J. Mansfield[□]
[♪]Office of Graduate Studies and Research, San José State University, San Jose, CA 95192-0025 USA
[♫]Chief, Information Technology Laboratory Statistical Engineering Division, NIST, Gaithersburg, MD 20899-8980
[□]National Physical Laboratory, Teddington, Middlesex, TW11 0LW UK

Abstract

In 1994, the leadership[1] of the US Biometric Consortium asked a series of questions to the automated human recognition ("biometrics") community revolving around the issues of repeatability and reproducibility of measurements in performance testing. Although we have made significant progress in our understanding, these issues have not been totally resolved. This paper discusses our current approach to repeatability and reproducibility within a broader context of scientific experimentation and NIST traditions in data evaluation and reporting. We discuss the Duhem-Quine Thesis on testing holism, Churchill Eisenhart's concept of "statistical control", NIST and ISO approaches to uncertainty in laboratory measurements, the current disconnect (lack of inductive relevance) between testing results and "performance" as assessed by system operators, and the need for statistical control and uncertainty assessment in our current biometrics test programs. We illustrate how measurement uncertainty is manifested in the context of technology, scenario and operational tests and advocate for moving beyond the calculation of "coverage" intervals as defined in the ISO/IEC "Guidelines for the Expression of Uncertainty in Measurement" to full application of the concepts of uncertainty assessment.

Introduction

In this paper, our central focus will be on "Technical Testing" in biometrics[2] – the reporting of recognition error (in the sense of an incorrect decision – either positive or negative – as to whether a biometric sample is from the same person as the biometric reference) and throughput rates – that has been the dominant NIST biometric testing paradigm since 1970[3]. We fully acknowledge, however, that other forms of testing, such as standards-compliance, usability, and reliability/availability/maintainability, are closely intertwined with technical testing and may, from an applications point of view, be ultimately more important. This paper is concerned with the measurement of technical performance of biometric systems – what we measure and report and the uncertainty inherent in these measurements. This places our endeavor firmly within the

[1] BC co-chairs Joseph P. Campbell and Lisa Alyea.

[2] We accept for the purposes of this paper the definition of biometrics from ISO/IEC JTC1 N3385, 16 Sept., 2009, as the "automated recognition of individuals based on their behavioural and biological characteristics", but recognize that the general scientific community defines the term more broadly.

[3] This conference also included a companion talk on "The Modern History of Biometric Testing in the US" documenting this test approach at NIST and elsewhere since 1970.

mission statements of NIST[4] and the Statistical Engineering Division of the Information Technology Laboratory[5]. Consequently, we want to bring to this discussion the rich history of assessment and reporting of measurement uncertainty developed by NIST (previously known as the National Bureau of Standards – NBS) and by the broader international statistical and metrological communities.

We will look at the conceptual development of holism in scientific testing in the Twentieth Century, the work on statistical control pioneered within the NBS Statistical Engineering Laboratory (SEL) in the 1960s, the contributions by NIST/NBS and international metrology communities to uncertainty measurement and reporting in the 1980s and 1990s, and then apply these historical concepts to technical testing in biometrics, following the NIST taxonomy of "technology", "scenario" and "operational testing".

Our primary remarks will be:

1. "Uncertainty", a broader concept than "error", is doubt about how well a test result represents the quantity it is said to measure. Uncertainty can exist even in the absence of error in the sense of "mistake".
2. A central source of uncertainty is definitional incompleteness in specifying all of the factors influencing the measurement.
3. What we actually measure is usually only a proxy for what we want to measure.
4. How we control, measure and report the values in a test must reflect how we expect those values to be used by others. In other words, our testing and reporting must take into account, and state, how we expect the results to be used.

We will apply these findings to "technology", "scenario" and "operational" biometric tests, as defined in [1] and explicated further below.

The Duhem-Quine Thesis and the Theory of Holism in Scientific Testing

If "biometrics" in our sense of the term is to be undertaken as science, we then by necessity inherit the historical and philosophical understandings of scientific testing in general. Although much of the history of science has been concerned with the methodology of scientific testing [2-4] and more recently even its anti-methodology [5], we want to focus our discussion on philosophical and statistical underpinnings of testing and reporting metrics. So rather than discuss, "How do we test?", in this paper we focus on "What do we measure and report?" and

[4] "to promote U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology in ways that enhance economic security and improve our quality of life."

[5] "SED seeks to contribute to research in information technology, to catalyze scientific and industrial experimentation, and to improve communication of research results by working collaboratively with, and developing effective statistical methods for, NIST scientists and our partners in industry."

"What does it mean?" When we make a measurement in biometrics, whether in the laboratory or the field, how should the results be understood?

In 1906, the French physicist and philosopher, Pierre Duhem wrote "An experiment in physics can never condemn an isolated hypothesis, only a whole theoretical group" [6]. Duhem's point here and throughout his work is that every experiment takes place within a context of theory, apparatus, and experimental conditions.

> "In sum, the physicist can never subject an isolated hypothesis to experimental test, but only a whole group of hypotheses; when the experiment is in disagreement with his predictions, what he learns is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed" [6].

Duhem was writing about physics and explicitly excluding other sciences such as physiology and chemistry, which he considered as less advanced and "still close to their origins". Writing nearly a half century later, without knowledge of Duhem's work[6] and from a completely different perspective[7], W.V.O. Quine wrote

> "The unit of empirical significance is the whole of science[8]….The totality of our so-called knowledge or beliefs…is a man-made fabric which impinges on experience only along the edges. Or, to change the figure, total science is like a field of force whose boundary conditions are experience. A conflict of experience along the periphery occasions readjustment in the interior of the field. Truth values have to be redistributed over some of our statements…But the total field is so underdetermined by its boundary conditions, experience, that there is much latitude of choice as to what statements to reevaluate in the light of any single contrary experience. No particular experiences are linked with any particular statements in the interior of the field except indirectly through considerations of equilibrium affecting the field as a whole." [7]

Quine's paper is considered one of the most influential papers on the philosophy of science in the Twentieth Century – "a modern classic" [8]. A number of more recent philosophers of science [9] have taken the Duhem and Quine theses together to create a "Theory of Holism" in scientific

[6] In "Two Dogmas in Retrospect", *Canadian Journal of Philosophy* 21(1991), Quine discusses his lack of acquaintance with the Duhem's writings during work on "Two Dogmas".

[7] One of Quine's two central theses in "Two Dogmas of Empiricism" was that the analytic/synthetic distinction of Kant is not supportable. No statement is truly "analytic" and therefore above reassessment based on empirical evidence. While making a huge impact, Quine did not deal a knock-out blow to the analytic/synthetic concept, which continues to survive.

[8] Quine later backed off of this extreme statement to adopt a more moderate view of the scope of the "unit of significance". See W. V. Quine, "Five Milestones of Empiricism" in Theories and Things, Harvard University Press, Cambridge, Massachusetts, 1981, p. 71.

testing. That theory says that the results of any scientific test reflect the totality of conditions of the test ("the unit of empirical significance"), including instrumentation, background assumptions, auxiliary hypotheses, and even the theories being tested themselves. So what we measure in any experiment is the totality of all the elements existing in both the physical and intellectual environment of the test and, further, the measurement results must be expressed using words and concepts that themselves may be subject to change as our understanding progresses.

So a "technical performance test" in biometrics measures everything involved in the test — both tangible and intangible. This includes not only the collection, analysis and computational equipment, but the human subjects and their attitudes/behaviors/beliefs, the physical environment of the test, including acoustic noise levels/ humidity/day of the week/season of the year, all colored by the underlying theories of what it is that biometric systems do. How we express, discuss and use these results will be evolving as our understanding of our science changes.

Churchill Eisenhart and the NBS/NIST Tradition of "Statistical Control" in Scientific Measurement

So if the results of a scientific experiment involving measurement express not only the object of measurement ("measurand") but also the effects of all experimental factors, how do we proceed to learn anything useful at all? Addressing this question fully would take us into the realms of the philosophy of science and epistemology in general — we will not pursue such lofty goal [10, 11]. One component of the answer, however, involves the concept of "statistical control", as pioneered by Walter Shewhart [12], which Churchill Eisenhart applied to measurement procedures. Churchill Eisenhart founded the Statistical Engineering Laboratory of NBS in 1947, to execute a mission that then NBS Director Edward U. Condon described as follows:

> "In these days when so much emphasis is properly being placed on economy of government research operations, it is important to take advantage of the substantial savings which can be effected by substituting sound mathematical analysis for costly experimentation. In science as well as in business, it pays to stop and figure things out in advance." [13]

So Eisenhart's remit, as is our own at the 2010 International Biometric Performance Conference, was to find ways of substituting brains for brawn in the collection, analysis and use of data. NIST statisticians have always been designers of experiments and interpreters of data, not simply hired number crunchers. Eisenhart's publication record on these topics while at NBS was truly impressive. In perhaps his most influential paper, he wrote "a measurement operation must have attained what is known in industrial quality control language as a state of statistical control . . . before it can be regarded in any logical sense as measuring anything at all."[9] [14]

For Eisenhart, a central requirement for any scientific measurement was that it be repeatable[10] and reproducible[11]. In calling on the language of industrial quality control, Eisenhart was appealing to the concept that variation can be categorized as stemming from, in the parlance of the time, "common causes" and "chance causes". Both types of variation limit the repeatability and reproducibility of scientific measurements. The goal of the experimenter then is to understand and control those causes. This is a state of "statistical control". The resulting measurements require an associated statement of uncertainty stemming from both chance causes and common causes, "and their relative importance in relation to the intended use of the reported value, as well as to other possible uses to which it may be put"[14]. In other words, when we make any experimental measurements we are required to consider and report the uncertainty in those measurements, and the resulting uncertainty assessment should express the contributions from all recognized sources of uncertainty.

In the context of Duhem-Quine, this means that the tangible factors from a test's "unit of empirical significance" impacting the observed values of the measurement must be known, controlled and reported. Those that cannot be known, controlled or reported, become part of the uncertainty in our measurement. Without a statement regarding the conditions pertaining in the "unit of empirical significance" being measured, we in fact have measured nothing at all. To learn anything useful from our test, we must attain a state of "statistical control" within that "unit".

An International Approach to Uncertainty in Laboratory Measurement

So if the unit of empirical significance is the whole of science (or at least some less expansive "web of belief"[12]), then how can we practically attain a state of "statistical control" over everything all at once so as to avoid measuring nothing at all? As we can't even list all the tangible factors influencing our measurements, we can't possibly attain "statistical control" over all of them in any set of experiments. And if we can't even state these factors, we seem relegated to measuring "nothing at all", unless we can find a way to acknowledge and account for potentially vast areas of unknown and uncontrolled influences on our measurements.

[9] Eisenhart was writing from within the NBS tradition and was influenced by his colleague R. B. Murphy, "On the Meaning of Precision and Accuracy", *Materials Research and Standards* **1,** 264-267 (1961).

[10] "measurement precision under a set of repeatability conditions of measurement", meaning "a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time" [15]

[11] "measurement precision under reproducibility conditions of measurement", meaning "conditions that includes different locations, operators, measuring systems, and replicate measurements on the same or similar objects" [15]

[12] As described in a very accessible work created for Quine's undergraduate philosophy courses: W. V. Quine and J. S. Ullian, The Web of Belief, McGraw-Hill, 1978.

In the 1960s, the Statistics Engineering Laboratory of the NBS grappled with this perplexing state of affairs, with Harry H. Ku taking a lead role in providing guidance for it [16, 17]. This, and a paper that Eisenhart[13] co-authored with Ron Collé and Harry Ku [18] laid the groundwork for the original 1993 ISO Guide to the Expression of Uncertainty in Measurement [19], and the NIST guidelines to evaluate and express measurement uncertainty [20].

The same needs were felt in other countries, and in 1977 the Working Group on the Statement of Uncertainties was convened by the *Bureau International des Poids et Mesures* (BIPM) in response to a request of the *Comité International des Poids et Mesures* (CIPM). In 1980, that group issued Recommendation INC-1 [21], which was subsequently approved by the CIPM.

> "The uncertainty in the result of a measurement generally consists of several components which may be grouped into two categories according to the way in which their numerical value is estimated:
>
> A. those which are evaluated by statistical methods,
>
> B. those which are evaluated by other means.
>
> There is not always a simple correspondence between the classification into categories A or B and the previously used classification into "random" and "systematic" uncertainties. The term "systematic uncertainty" can be misleading and should be avoided.
>
> Any detailed report of the uncertainty should consist of a complete list of the components, specifying for each the method used to obtain its numerical value." [21]

Note that the classification into Types A and B applies to the methods of evaluation, not to the sources of uncertainty themselves, a point that both the GUM and the VIM [15] emphasize. Evaluation by statistical methods means assessment of a component of uncertainty using statistical methods applied to replicated indications obtained during measurement. Other means of evaluation include information derived from authoritative publications, for example in the certificate of a certified reference material, or based on expert opinion. In general, Recommendation INC-1 maintained a conventional approach to overall reporting of uncertainty, stating

> " ...The components in category A are characterized by the estimated variances… (or the estimated "standard deviations"...) and the number of degrees of freedom…. The components in category B should be characterized by quantities … which may be considered as approximations to the corresponding variances, the existence of which is assumed. … The combined uncertainty should be characterized by the numerical value obtained by applying the usual method for the combination of variances. The combined

[13] See also C. Eisenhart, "Expression of Uncertainties in Final Results", *Science* 160 (3833), June 14, 1968

> uncertainty and its components should…be expressed in the form of “standard deviations”.” [21].

In 1992, the NIST Director, John W. Lyons, established a new NIST policy on the reporting of measurement uncertainty which essentially adopted INC-1, stating “… worldwide use (of INC-1) will allow measurements performed in different countries and in sectors as diverse as science, engineering, commerce, industry, and regulation to be more easily understood, interpreted, and compared.” [22].

To assist NIST in implementing this new policy, the Director directed the NIST Physics Laboratory to create a Technical Note, originally issued in 1993 [20]. In 1994, ISO/IEC issued Guide 98, “Guide to Expression of Uncertainty in Measurement”, which has been updated and maintained within the ISO/IEC process through to the current version [19]. NIST Technical Note 1297 and ISO/IEC Guide 98 are similar, but only the latter has undergone updating since the early 1990s. In the interest of brevity, we will discuss only Guide 98 here.

ISO/IEC Guide 98

The ISO/IEC “Guide to Expression of Uncertainty in Measurement” (commonly referred to as the GUM) was motivated by the need to establish a single, internationally accepted approach to reporting uncertainty in laboratory measurements. The GUM sets the framework for uncertainty assessment as follows:

> “The concept of *uncertainty* as a quantifiable attribute is relatively new in the history of measurement, although *error* and *error analysis* have long been a part of the practice of measurement science or metrology. It is now widely recognized that, when all of the known or suspected components of error have been evaluated and the appropriate corrections have been applied, there still remains an uncertainty about the correctness of the stated result, that is, a doubt about how well the result of the measurement represents the value of the quantity being measured.” [19]

The GUM uses the term “measurand” to denote the quantity being measured, explaining this as

> ‘The first step in making a measurement is to specify the measurand — the quantity to be measured… in principle, a measurand cannot be *completely* described without an infinite amount of information. Thus, to the extent that it leaves room for interpretation, incomplete definition of the measurand introduces into the uncertainty of the result of a measurement a component of uncertainty that may or may not be significant relative to the accuracy required of the measurement.
>
> Commonly, the definition of a measurand specifies certain physical states and conditions.

> EXAMPLE The velocity of sound in dry air of composition (mole fraction) N_2 = 0.780 8, O_2 = 0.209 5, Ar = 0.009 35, and CO_2 = 0.000 35 at the temperature T = 273.15 K and pressure p = 101 325 Pa." [19]

The GUM discusses the measurement of a measurand observed through a test as the "realized quantity"

> "Ideally, the quantity realized for measurement would be fully consistent with the definition of the measurand. Often, however, such a quantity cannot be realized and the measurement is performed on a quantity that is an approximation of the measurand." [19]

In this context "uncertainty" is to be understood generically in its natural language meaning of incomplete knowledge (of the measurand we know only as much as indications obtained during measurement reveal and add to any preexisting knowledge, considering that these indications are themselves affected by uncertainty). This does not suggest the presence of "error" in the sense of "mistake". More specifically, the VIM (2.26) [15] defines "measurement uncertainty" as a non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used (this obviously fails to cover cases where the measurand is not a scalar, but can be generalized to address them too).

The GUM summarizes several possible sources of uncertainty in measurement, which happen to be operational in all forms of biometric testing:

> a) incomplete definition of the measurand;
> b) imperfect realization of the definition of the measurand;
> c) non-representative sampling — the sample measured may not represent the defined measurand;
> d) inadequate knowledge of the effects of environmental conditions on the measurement or imperfect measurement of environmental conditions;
> e) personal bias in reading analogue instruments;
> f) finite instrument resolution or discrimination threshold;
> g) inexact values of measurement standards and reference materials;
> h) inexact values of constants and other parameters obtained from external sources and used in the data-reduction algorithm;
> i) approximations and assumptions incorporated in the measurement method and procedure;
> j) variations in repeated observations of the measurand under apparently identical conditions."[19]

These quotations above from the GUM encapsulate the primary points being made in this paper, which will be applied in the next section to measurement in biometrics.

1. "Uncertainty" is a broader concept than "error", as historically understood; it is the doubt (reflecting incompleteness of knowledge) about how well the test result represents the quantity measured (or being said to be measured). Uncertainty can exist even in the absence of error in the sense of "mistake".
2. A central source of uncertainty is definitional incompleteness in specifying the "unit of empirical significance" for the measurand – full specification of which would require a "infinite amount of information".
3. What we are measuring is often only a proxy for the measurand of real interest, even if fully defined, which adds yet another source of uncertainty in our measurement.

But the GUM also says

> "In practice, the required specification or definition of the measurand is dictated by the required accuracy of measurement. The measurand should be defined with sufficient completeness with respect to the required accuracy so that for all practical purposes associated with the measurement its value is unique."[19]

This echoes Eisenhart in telling us that the required level of definition and specification of the measurand is limited by the measurement accuracy required by the intended use of the data. If we have a target application in mind, we may be able to estimate the degree of statistical control required so that our measurements will be close enough for our intended use. So the required level of statistical control over the tangible elements in the unit of empirical significance is limited by the uses intended for the data. To our three points above we can now add a fourth:

4. How we control, measure and report the values in a test must reflect how we expect those values to be used by others. In other words, our testing and reporting must take into account, and state, how we expect the results to be used.

One of the GUM's most consequential prescriptions is the suggestion that all assessments of uncertainty components should be treated alike, irrespective of whether they will have been evaluated by Type A or Type B methods. And the common treatment is as if they were variances or standard deviations, with their natural extensions when the measurand is not a scalar.

This approach gives a highly flexible method for assessing all uncertainty components and using the resulting assessments. In particular, the GUM recognizes the possibility of "insight based on experience and general knowledge" producing valid assessments, and notes "that a Type B evaluation of standard uncertainty can be as reliable as a Type A evaluation, especially in a measurement situation where a Type A evaluation is based on a comparatively small number of statistically independent observations." [19].

Estimates of the contributions from all sources of uncertainty, irrespective of their intrinsic nature and of the method that will have been employed to assess them, must be combined into a "combined uncertainty" taking the form of a standard deviation, duly taking into account any correlations between those sources. The resulting combined uncertainty may then be used to produce an interval that, with stated "level of confidence", is supposed to include the measurand: the endpoints of this interval may be obtained by adding and subtracting a suitable multiple of the combined uncertainty to and from the measured value, or in other ways, for example, using the methods described in the Supplement 1 to the GUM [23].

The GUM (6.2.2) goes to some lengths to distinguish such intervals as it defines from conventional confidence intervals as introduced in [24][14] in the context of frequentist statistics. However, the expression "confidence interval" also is used in Bayesian statistics [25], but its interpretation, of course, differs in the two settings. The GUM also creates additional opportunity for confusion when (in C.2.30) it defines "statistical coverage interval" as an "an interval for which it can be stated with a given level of confidence that it contains at least a specified proportion of the population" — indeed, this sounds more like the definition of a tolerance interval than a definition of any other type of probabilistic interval [26].

Leon Gleser provides an insightful, critical analysis of the meanings implicit in the GUM, and of its statistical underpinnings [27].   In particular, he notes that "The ISO recommendation has been of concern to many statisticians because it appears to combine frequentist performance measures and indices of subjective distributions in a way that neither frequentists nor Bayesians can fully endorse". The revision of the GUM, underway in Working Group 1 of the Joint Committee for Guides in Metrology (Bureau International des Poids et Mesures, Sèvres, France), hopefully may remove such ambiguities while continuing to provide guidance that is widely applicable and is responsive to the needs of all fields concerned with measurement, including biometrics.

[14] Neyman's seminal paper defining "confidence intervals" in fact limits the analysis to two cases: " (ia) The statistician is concerned with a population, π, which for some reason or other cannot be studied exhaustively. It is only possible to draw a sample from this population which may be studied in detail and used to form an opinion as to the values of certain constants describing the properties of the population ,π…..(ib) Alternatively, the statistician may be concerned with certain experiments which, if repeated under apparently identical conditions, yield varying results." GUM is concerned with a much broader range of conditions, including experiments which cannot be repeated under identical conditions, as in biometrics.

What We Measure in Biometrics: Its meaning, statistical control and uncertainty

In biometric technical testing, it is standard [28] to report performance in terms of "error rates" in the recognition process — those errors being either false positive (an incorrect decision that a biometric sample and the biometric reference are from the same individual when they are not) or false negative (an incorrect decision that a biometric sample and a biometric reference are not from the same individual when they in fact are) — as well as the interrelated measures of "failure to enroll", and "failure to acquire" rates and throughput rates. Any of these five measurands can be set as independent variables, impacting all of the other measurements in a test.

The false positive (also called "false match") and false negative (also called "false non-match") error rates are generally compared at a variety of thresholds using the traditional ROC[15] or more recent NIST IAD-championed DET curves [29]. These errors can be determined to the extent that "ground truth" of the test data is known – that is, that we really know which test comparisons were of biometric characteristics from the same source and which comparisons were of characteristics from different sources.

The fact that our measurands are "error rates" raises linguistic difficulties as common literature on measurement uncertainty uses the term "error" to indicate the deviation of the measurand from its "true" value[16]. In this paper we will use "error rate" with the meaning that it has in biometric testing.

But what is it that we actually measure, what statistical controls are required and what is our measurement uncertainty? In 1999, NIST developed the concept of three levels of testing in biometrics: technology, scenario and operational [1]. This concept made its way quickly into the literature and standards in the field. In the next section, we will consider test uncertainty in technology, scenario and operational testing sequentially.

Statistical Control and Uncertainty in Technology Tests

A technology test uses a pre-collected database of biometric samples and reports the ROC or DET for each software package[17] against that database. So each software package is tested

[15] The traditional ROC plots the complement of the false non-match rate (the "hit rate") against the false match rate (called the "false alarm rate")

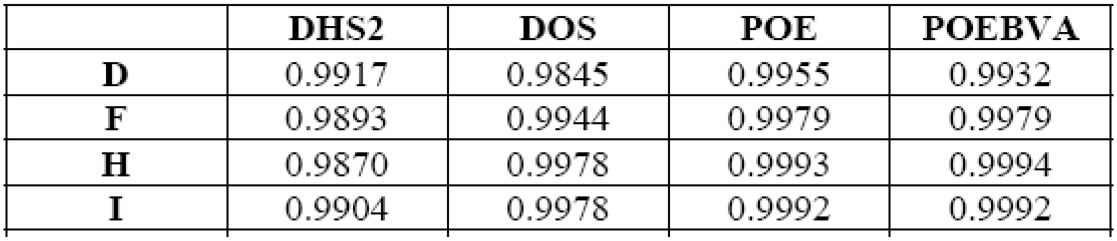
[16] VIM (2.16) defines "measurement error" as " measured quantity value minus a reference quantity value." [15]

[17] We are specifically avoiding the use of the word "algorithm" because a software package for biometric recognition will contain many different algorithms—for example, algorithms for signal detection, segmentation, feature extraction, model creation and comparison. Consequently, our technology test results will not tell us

against exactly the same data and our results tell us about the efficacy of the software with respect to the database. Examples include most of the NIST IAD biometric test programs, such as the Speaker Recognition Evaluation, the Multi-Biometric Grand Challenge, the Proprietary Fingerprint Template Testing, and MINEX programs.

If the software is deterministic, and there is a prescribed way to use the test data, then barring software or hardware limitations that may bias the results, and assuming pristine bookkeeping by the testing agency, the results of a technical test would be completely repeatable. Running a software package multiple times against the same database would, under these assumptions, result in the same performance measures. We can state number of false match and false non-match errors (or rate of errors per non-mated and mated comparisons, respectively) for each package when run on each database.

As an example, Table 1, from the NIST Proprietary Fingerprint Testing Program [30], shows results for True Accept Rate (TAR =1 – FNMR, where FNMR is the false non-match rate)[18] at a False Match Rate of 0.0001 for four packages (D, F, H and I) against each of four databases (DHS2, DOS, POE and POEBVA).



	DHS2	DOS	POE	POEBVA
D	0.9917	0.9845	0.9955	0.9932
F	0.9893	0.9944	0.9979	0.9979
H	0.9870	0.9978	0.9993	0.9994
I	0.9904	0.9978	0.9992	0.9992

Table 1 from NIST Proprietary Fingerprint Test Program[19] "TAR at FAR = 0.0001"[20]

Under the assumptions in the previous paragraph and allowing no changes to either the software or the database, these results would be completely repeatable within the same lab and reproducible across labs[21]. Table 1 shows that changes in either software or database result in unpredictable changes in reported results[22].

which of the algorithms in the package are functioning well and which are defective, but rather will reflect the function of the software package as a whole, to echo Duhem [6].

[18] NIST IAD currently does not report results in either ISO/IEC or INCITS-compliant formats.

[19] From the February 26, 2010 update at http://fingerprint.nist.gov/PFT/tables2f_121109.pdf

[20] Here, NIST's use of the term "FAR" (False Acceptance Rate) is to be interpreted as the false match rate.

[21] However, for reasons of data protection of the personal biometric data and to maintain the data as sequestered for testing purposes only, it is generally not possible to share databases across laboratories. An equivalent "reference" database of the same difficulty cannot be constructed by a second test lab, since the conditions under which the original data were collected cannot be replicated.

[22] For example, Software Package D performs better on database DHS2 than on DOS, while Software Package F performs worse. Therefore, we cannot predict even the direction of the impact on the performance measurements when making changes to algorithms or databases.

Even if completely repeatable and reproducible, these results will contain uncertainty in the statement of the measurand (comparison errors) resulting from the comparison process. The comparison process takes place in the context of a matching key (assumed "ground truth") — the answer sheet holding the truth values. The measured quantities reported in the test — the comparison errors as discrepancies between the algorithm decisions and the matching key — are really proxies for the stated measurands, which are the "error rates" of the algorithm against the database. An assumed equivalence between the comparison quantities measured and the stated measurands would require complete accuracy and statistical control of the matching key for the database. NIST has estimated the labeling errors in government databases to be 1 in 1000 [31]. It is quite common in the course of NIST evaluations for the matching keys to be updated as mistakes in the original are suspected, then confirmed through analysis methods independent of the recognition software.

GUM is not willing to give guidance on the issue of labeling errors, stating

> "Blunders in recording or analyzing data can introduce a significant unknown error in the result of a measurement. Large blunders can usually be identified by a proper review of the data; small ones could be masked by, or even appear as, random variations. Measures of uncertainty are not intended to account for such mistakes." [19]

The GUM's general probabilistic apparatus to measurement uncertainty, however, can be extended to include the consequences of labeling errors. Indeed, making either the most simplistic assumption (that labeling errors occur independently and with constant probability throughout all the items in a database), or assumptions that may more realistically capture the patterns on interdependence to be expected (illustrated below), statistical models can be built that produce assessments of uncertainty attributable to this particular cause, even if "by construction" the mislabelings are systematic, not random.

Also, considering labeling errors as one contribution to measurement uncertainty is consistent with the approach used by NIST in [31]. That approach was to refuse to report results below the measurand uncertainty level introduced by the suspected level of labeling errors. Figures in [31] intentionally cut off all graphs at measured values below 0.001 for either type of error, allowing the data line to disappear off the boundaries as the labeling error rate was given a subjective probability by the experimenters of exceeding the value of the measurand.

So the "unit of empirical significance" for a NIST technology test contains the software package, the database and the matching key[23] along with all the other factors in our "web of belief". If the

[23] Of course, each of these areas is a "unit of empirical significance" in itself, containing massive requirements for statistical controls. The oft-told story of the 1987 KING speech database collection illustrates how lack of statistical control in database collection can disrupt an entire government technical test program. About halfway through the KING collection, a minor equipment change was made that altered the spectral content of the speech recordings, creating a divide in the data that disrupted data analysis for years until the problem was understood.

only source of uncertainty is the labeling mistakes in the key, how can we estimate more quantitatively the impact on performance results? Often there is enough redundant information in a data collection for the experimental team to eliminate most labeling mistakes, but ultimately we can never know what residual mistakes remain and will be required to make a Type B assessment.

These labeling mistakes will be distributed in some unknown way over both incorrect mating and non-mating information. Incorrect non-mating information (to say two samples are from different sources when they are from the same source) will cause an accurate software package to appear to have a higher number of false matches. We'll call these "Type 1" mislabeling. Incorrect mating information (to say two samples are from the same source when they are not) will lead an accurate software package to appear to have a higher number of false non-match errors. We'll call this "Type 2" mislabeling.

The statistical impact of the two types of labeling mistakes will be different. Consider a database with n samples, of which there are actually m mated pairs to which the key subtracts k Type 1 mistakes and adds j Type 2 mistakes. The key will consider $m + j - k$ pairs to be mated. A perfect software package will fail to match the j pairs and be assessed a false non-match rate of $j/(m + j - k)$. It will match the k pairs and be assessed a false match rate of $k/[n(n - 1)/2 - (m + j - k)]$. An estimate of one labeling mistake per 1000 samples implies that $(j + k)/n = 0.001$, but there is no reason to suppose any partitioning of these mislabels over Types 1 and 2, so there is insufficient information to even guess at the quantitative bias introduced into FNMR and FMR except to say that reported values will be biased on the high side.

These mislabelings in the key will not be random, but rather systematic – for example, Type 2 mislabeling of twins as the same person or Type 1 mislabeling of different fingerprints from the same person as the same fingerprint. Therefore, the stronger (more accurate) software packages will be more likely to deliver results conflicting with the mistaken key, while weaker packages may agree with the key. So the impact of the labeling mistakes on measured performance error rates will not be uniform across all software packages and will force any meaningful coverage interval to expand primarily in one direction. The "true" value of measurands (i.e. the comparison error rates) are likely to be smaller than measured. Furthermore, uncertainty in the reporting of the FMR may require a change in the threshold at which the FNMR is reported. An FMR reported higher than the correct value because of labeling mistakes would further push the reported value of FNMR to be higher than the correct value at the stated FMR.

Returning to Table 1, we may be wrongly tempted to make broader statements about the results saying the Software Package D averaged a TAR ($1 - \text{FNMR}$) of 0.9912 over the interval of (0.9845, 0.9955) at a measured FMR of 0.0001, thus indicating some uncertainty owing to

See J. Godfrey, D. Graff, and A. Martin, "Public databases for speaker recognition and verification", *ESCA Workshop on Automatic Speaker Recognition, Identification and Verification*, 1994.

databases, but not to matching keys. Again, this uncertainty is asymmetric and cannot be expressed as *a*± *b* for any software package. The GUM's prescription of intervals of this form, however, cannot be divorced from the underlying assumptions that validate such prescription, one of whose central tenets is that the measured value is like an outcome of a Student's t random variable (possibly, but not necessarily, with a large number of degrees of freedom). The Supplement 1 to the GUM provides the means to tackle situations where such assumption of symmetry is not applicable [23].

Further, because we don't know the size of the various databases, such an average may overemphasize results from the smaller databases in the ensemble. The mean value might be wrongly taken to be the TAR which would be found if the databases were consolidated into a single test set. Consequently, calculation and reporting of the variance of the measures around the mean across all databases would be of no value. We must not be tempted to implicitly extend these results to include any possible databases. The uncertainty reported as above applies only to the tested databases under the assumption of no labeling errors, which we have shown introduces systematic, not random, variation in our measurements. In short, it would be highly misleading and scientifically wrong to make the overly generalized statement "Software Package D has a TAR of 0.9912 at FAR=0.0001". A more correct statement would be "Software Package D had a measured TAR at FAR=0.0001 of 0.9845 to 0.9955 across the various databases tested. Labeling errors could cause the true TAR to be as much as XX higher than measured" (where XX is a personal estimate of the experimenter).

Of course, the above paragraph could be repeated for Packages D, F, H and I for any one of the databases to give, for instance, a measure of database difficulty. But trying to make a general statement for all tested software packages across all tested databases would compound the problems already stated. We cannot use a technology test to give us performance metrics for a general technology (say automatic fingerprint recognition) independent of reference to a specific test database. Therefore, technology testing on artificial or simulated databases tells us only about the performance of a software package on that data. There is nothing in a technology test that can validate the simulated data as a proxy for the "real world", beyond a comparison to the real world data actually available. In other words, technology testing on simulated data cannot logically serve as a proxy for software performance over large, unseen, operational datasets.

But what about random uncertainty owing to small database size in a technology test? Such uncertainty stems from the random selection of a subset of data from a much larger database. The uncertainty then includes a component that is attributable to the corresponding sampling: if this should be random, then it is the object of Neyman's defining work [24]. One approach (used by NIST in their Face Recognition Vendor Test 2002 [32]) is to employ a "divide-and-conquer" strategy whereby a single database is randomly partitioned into subsets, and the variability of the performance (error rates) of the matching software among the different subsets be used as intimation of its operational acumen. This is similar to the idea of statistical "cross-validation", where models are developed based on one part of the data, another part being set aside, and their performance is assessed using the portion set aside. However, in technology testing, test data is generally not a randomly selected subset of a larger database. Rather, if a larger database of biometric samples does exist, a selection of the test database is made to systematically remove problems, such as “poor quality” or blank images[24]. Some of the issues of dataset selection are well considered in Appendix E of [31]. So lacking a larger database from which the test dataset was randomly selected, there are no associated random influences connected with the measures reported for the test database.

Of course, there could be the hypothetical extension of a database that might have been, but was not, collected. This would require the untenable assumption that “unit of empirical significance” would remain completely unchanged except for size and that all of the statistical controls on the smaller collection would be applied to the larger collection with no problems of scaling. We know from experience, however, that to acquire a dataset of 30,000 samples requires us to look for very different sources than to opportunistically acquire 300 samples. Therefore, the “hypothetical extension” argument does not create any larger database within the same unit of empirical significance as the real data, meaning that there is no established need for Type A random uncertainty calculation for technology tests. Values extrapolated from our test database to some hypothetically identical but larger database under the assumption of non-existing statistical controls would measure “nothing at all”.   Consequently, questions of the type commonly stemming from Table 1, “Are the measured differences in performance of the tested algorithms statistically significant given the size of the test?” is ill-posed, although the question “Are the measured differences significant given the expected occurrence of labeling errors?” will have merit requiring Type B analysis.

Is there any hope of inductively extending the results of our technical test more broadly to any other algorithms or databases?   A Type B systematic uncertainty evaluation after consideration of changes in the unit of empirical significance and statistical controls over its tangible elements might be of value, provided that the specifics of the changes could be given, but we should not sanctify such a “guesstimate” in an emperor’s cloak of imagined analytic rigor.

Statistical Control and Uncertainty in Scenario Tests

A “scenario test” moves us from the use of previously collected data to the process of data collection from human subjects.   The introduction of human data subjects directly into the test environment causes great changes in the unit of empirical significance and makes scenario testing a social science -- as much about the data subjects as about the technologies.

[24] The NIST Iris Comparison Evaluation is an example of a test set created by removing problem images selectively from a larger dataset. See Appendix A-1.1 of P.J. Phlips, et al, “FRVT 2006 and ICE 2006: Large-scale Results” *NISTIR 7408*, March, 2007

> “Scenario evaluations measure overall system performance for a prototype scenario that models an application domain… Scenario evaluations test complete biometric systems under conditions that model real-world applications. Because each system has its own data acquisition sensor, each system is tested with slightly different data. One scenario evaluation objective is to test combinations of sensors and algorithms.” [1]

In a scenario test, the experimenter controls to some extent the conditions of the collection and the test population, thereby exerting some influence on the statistical controls of the tangible elements within the unit of empirical significance. The goal is to model a system in an environment and with a population relevant to the evaluation, but what conditions should be included in the model?

In a scenario test, FNMR and FMR measurands are given as rates averaged over total transactions. The transactions often involve multiple data samples taken of multiple persons at multiple times. So influence quantities[25] extend to sampling conditions, persons sampled and time of sampling. These quantities are not repeatable across tests in the same lab or across labs, so measurands will be neither repeatable nor reproducible. We lack metrics for assessing the expected variability of these quantities between tests and models for converting that variability to uncertainty in measurands.

A recent NIST handbook [33] suggests that factors attributable to or impacting the data subjects which influence technical test measurements include, but are not limited to: data subject age, gender, experience, and height; the instructional materials presented the data subject; the placement, height and angle of the collection device; the temperature, humidity, lighting and noise of the collection location. For a scenario test, the values of these factors in the operational environment to be modeled may not be known or controllable by the experimenter.

In addition to this list of general, technology-independent, but human-related factors to be controlled, each specific recognition technology (iris, face, voice, fingerprint, hand, etc.) will have specific factors that must be within a state of statistical control. This list of factors is not well understood, although ample work in this area is continuing.

For example, recent analysis of iris [34] and face [35] recognition test results shows us that to report false match and false non-match performance metrics for such systems without reporting on the percentage of data subjects wearing contact lenses, the period of time between collection of the compared image sets, the commercial systems used in the collection process, pupil

[25] An Influence quantity is a “quantity that is not the measurand but that affects the result of the measurement” [19]

dilation, and lighting direction is to report "nothing at all"[26]. Our reported measurements cannot be expected to be repeatable or reproducible without knowledge and control of these factors.

But beyond the observable and easily quantifiable aspects of the collection conditions, such as lighting and acoustic noise, and of the data subjects, such as height, handedness and presence of glasses, the less tangible influence quantities of human attitudes, perceptions and behaviors of both the data subjects and supervising experimenters impact the measurands and thus the repeatability and reproducibility of scenario test results. The story of the "lost" 1993 scenario test from Sandia National Laboratory is illustrative. In the widely-referenced 1991 scenario test of several biometric access control devices [36], Sandia National Lab (SNL) determined the comparison error rates for hand geometry. At a threshold of 100 vendor-proprietary units, the device achieved an "equal error rate"[27] of 0.2% for both measurands. In the 1993 test [37], the "equal error rate" of 5% was achieved at the same threshold of 100[28]. The manufacturer asked for an explanation of the extreme increase in the measurand in the second test, given that the devices tested were nearly identical. An internal enquiry at SNL determined that one data subject had been conducting undocumented "experiments" with hand placement. So the test design was such that the behavior of one data subject was able to increase both measurands by a factor of 25 and prevent repeatability of the first test. The second test report was withdrawn and no copy was available within SNL until a "preliminary draft" was discovered in the archives of the National Physical Laboratory and a copy returned to SNL. The story illustrates that the uncertainty in our measurements in a scenario test extends to the attitudes and behaviors of the data subject.

In 2000, the National Physical Laboratory [38] tested a very similar hand geometry device from the same manufacturer, obtaining measurands of 0.7% (FAR) and 0.5% (FRR) at the threshold of 100 vendor-proprietary units. These results show that the test repeatability and reproducibility observed in technology tests are lost in scenario testing due to the loss of statistical control over a wide range of influence quantities.

An additional complication in a scenario test is that the software package may require modifications such that the comparison scores can be fully observed. This means that the software package being tested is itself only a proxy for the package that would be installed in the target operational environment with additional operational controls and interfaces. As in technology tests, complete statistical control over "ground truth" — accurate knowledge of which comparisons are mated and which are non-mated — may be difficult to attain.

[26] But this is not to say that our listing of relevant sources of uncertainty in this list is complete. So even if we report and control these factors, we may be still unable to repeat or reproduce these results and may indeed have measured nothing at all.

[27] "Equal error" or "cross-over" error rates are operationally meaningless metrics that allow us to report a single value for the two measurands. Comparing tests or devices in terms of "equal error rate", however, is to compare incommensurate metrics, so this custom introduces confusion into our community.

[28] That the equal error rates in the two tests occurred at the same threshold makes these results comparable.

The measurand in a scenario test usually refers to a system level decision of "recognized/not recognized". Due to the historical connection of some biometric technologies with access control applications, the terms "accept/reject" are often used in these meanings. So by convention, the measurands will be the rates of "false acceptance" (meaning that the data subject is recognized as someone else) and the "false rejection" (meaning the data subject is not recognized against her/his own reference biometric characteristics). The term "rate" indicates ratio of occurrences per number of opportunities for mistake. In the case of "false reject rate", it is the number of system decisions of "not recognized" over the number of attempts by enrolled data subjects to be recognized[29]. These "false acceptance/false rejection" rates will be related through inverse models[30] to the false match/non-match error rates of the software, but the relationship models will be complicated, meaning that the "false acceptance/ false rejection" rates will serve as complex proxies for the "false match" and "false non-match" rates. Or looking at it another way, the false match/ false non-match error rates measured in technology tests will be poor (highly uncertain) proxies for estimating, through usually complex models, the "false acceptance/ false rejection" rates of a scenario test.

Additionally, scenario testing will usually take place with a decision threshold in place that will supply feedback to both the experimenters and the data subjects, both of whom will generally adjust their behavior to produce favorable outcomes at the test threshold. Devices may also continue to collect samples from the data subject until a comparison score reaches the threshold. This latter condition would mean that "off-line" and "on-line" tests to determine a score distribution for non-mated comparisons would lead to different results and both would be dependent on the thresholds used in the scenario test. This implies that "impostor" testing should be done as part of the scenario, not with "off-line" testing with data collected during the test. Even so, test results shown on ROC or DET plots do not represent well the measurands at thresholds other than that used for the test. Consequently, scenario tests should be done at decision thresholds realistic to the target operational environment. Such thresholds are set in operation after a subjective analysis of the risk implied by each kind of error. Consequently, one important influence quantity in our unit of empirical significance will be the anticipated level of risk tolerance by the system operators. Additional influence quantities impacting scenario test measurands are given in [28, 39]. Our inability to apply concepts of statistical control to any or all of these factors will increase the level of uncertainty in our results and translate to loss of both repeatability and reproducibility.

Therefore, if the purpose of the scenario test was in fact to predict performance in the modeled operational system, we should accept that there is a high level of uncertainty in the measures. At our current level of understanding of the influence factors in both scenario and operational environments (owing to lack of both experience and data), we should endeavor as best we can to

[29] Our terminology here reflects an access control application.

[30] The forward model will give us the false acceptance/ false rejection rates as a function of the false match and false non-match rates, so determining the latter from the former will require inverse models.

develop Type B uncertainty assessments. Test data from scenario evaluations should not be used as input to mathematical models of operational environments that require high levels of certainty for validity.

Statistical Control and Uncertainty in Operational Tests

An operational test seeks to determine technical performance metrics in a real, application environment. The problems with such testing are legion and few operational tests have been performed. Of those performed, only a handful have been published [40, 41], owing to the sensitivity of system operators to what are perceived as security-related metrics. The lack of extensibility of technical and scenario test measurements to operational environments has been noted for as long as there has been third-party testing in biometrics [42].

NIST holds no illusions that the results of technical or scenario tests can be inductively extended to predict operational performance. In responding to USA-PATRIOT Act requirements that

> "…the National Institute of Standards and Technology (NIST)…shall… develop and certify a technology standard that can be used to verify the identity of persons applying for a United States visa or such persons seeking to enter the United States pursuant to a visa for the purposes of conducting background checks, confirming identity, and ensuring that a person has not received a visa under a different name or such person seeking to enter the United States pursuant to a visa…" [43]

NIST responded

> "For purpose of NIST PATRIOT Act certification this test certifies the accuracy of the participating systems on the datasets used in the test. This evaluation does not certify that any of the systems tested meet the requirements of any specific government application. This would require that factors not included in this test such as image quality, dataset size, cost, and required response time be included."[31]

The NIST response clearly limits the applicability of the technology test results of [31] to the test itself, discouraging extrapolation to any operational environment.

Operational testing inherits all of the uncertainty of technology and scenario tests, but with added problems of statistical control over all population and environmental elements in the unit of empirical significance. Of particular difficulty in operational testing is determining the "ground truth" of any comparison. In an operational environment, the data subject will make some claim[31] (perhaps only implicitly) as to being or not being the source of a biometric characteristic in the database. We can assume that an acceptance of this claim, regardless of whether true or

[31] For linguistic consistency, the "claim" is always from the point of view of the data subject. For example, in a watchlist environment, a data subject may not realize that a biometric system is in operation. However, if presence on the watchlist is pejorative, we will say that the data subject makes an implicit claim not to be the source of any biometric characteristic in the database.

false, by an operational data subject will go unreported. When a software decision results in "rejection" of the claim of an operational data subject, the data subject may be informed of the decision by some applications, allowing redress. If we assume that the redress process always correctly classifies the original claim as "true" or "false", we can get an estimate of the "system false rejection" rate. Of course, this assumption is incorrect, leading to measurement uncertainty.

But those with an "accepted" claim, whether that claim is true or false, will not seek redress. Therefore in an operational system, we will have a way of estimating the "system false rejection" rate, but not the "system false acceptance" rate.

But there is another hidden problem here. We use the same terms, "false acceptance rate" and "false rejection rate", to describe the measurands in both scenario and operational tests, but these measurands are actually very different across the two types of tests. In the scenario test, we seek to control the presentation of the biometric characteristic by the data subject. This control can be through training or supervision or even the exclusion of data deemed not given in "good faith"[32]. We base our measurands on data acquired and even edited according to such rules. In an operational test, we cannot actively control such presentations, nor can we make judgments on the "good faith" of the data subjects. In the case of "genuine" transactions (i.e., with a correct claim), we have no way of distinguishing poor faith attempts from confusion over the proper use of the equipment (assuming access control applications) or simply lack of motivation and knowledge (in applications beyond access control). Indeed in some cases, for example when the biometric characteristics are mostly obscured, rejecting the claim would be the appropriate action for the biometric recognition software, and so ought not be considered a "false rejection". Moreover, in scenario evaluation, "false acceptance" is defined over so-called "zero-effort" impostor transactions — i.e., where the impostor is making no special effort to be incorrectly recognized (other through making the incorrect claim). However, in an operational system the attacking impostors are most likely to be motivated to go beyond zero-effort impersonations.

For such reasons, "false acceptance" and "false rejection" in a scenario test are not directly comparable to "system false acceptance" and "system false rejection" in an operational test.

Consequently, the relationships between the measurands in an operational test (system false acceptance/ system false rejection) and those in either technology tests (false match/false non-match) or scenario tests (false acceptance/false rejection) will be difficult to assess. We can conclude that the three types of tests are measuring incommensurate quantities and therefore should not be at all surprised when the values for the same technologies vary widely and unpredictably over the three types of tests.

[32] See ISO/IEC 19795-1 for discussion on excluding data based on lack of "good faith" by the test subject.

Summary

If automated human recognition, or “biometrics” in our sense of the term, is to be considered as a science, then it must be subjected to the same requirements in testing and reporting as any other discipline within the scientific community. The international statistical and metrology communities, including the National Institute of Standards and Technology, has focused a century of effort into explicating scientific testing and reporting: what measurements are desired, what measurements can be made in a test, what measurements are actually made, how these various measurements are related and how the attendant uncertainties can be characterized and quantified. The International Organization for Standardization document, “Guidelines for the Expression of Uncertainty in Measurement” (GUM), was created to supply a framework for the consideration of measurement practices and the evaluation of uncertainties resulting from laboratory and field measurements. This document is as relevant to biometrics as to any other scientific field.

Technical testing in biometrics has historically focused on throughput and recognition error rates – the latter of two types: false positives (also called false matches – an incorrect decision that two biometric samples are from the same individual when they are not) and false negatives (also called false non-matches – an incorrect decision that two biometric samples are not from the same individual when they in fact are). These measurements have historically proved to be neither reproducible nor repeatable except in very limited cases of repeated execution of the same software package against a static database on the same equipment. Accordingly, “technology” test metrics have not aligned well with “scenario” test metrics, which have in turn failed to adequately predict field performance. All test results carry uncertainties related to measurand definitional incompleteness and systematic variation that have not been addressed by our biometrics community. In this paper, we applied the concepts developed by the statistical and metrology communities, as articulated in GUM, to explain why this is to be expected and how uncertainty assessment must be extended beyond the computational bounding of random variation owing to small test size.

We have shown that technology tests only measure proxies for comparison error rates, how scenario tests introduce uncontrolled and undocumented population and environmental variations, and how operational tests measure an overall system performance incommensurate with the software inadequacies measured in the other forms of testing.

Specifically, our primary remarks can be summarized as:

1. “Uncertainty”, a broader concept than “error”, is doubt about how well a test result represents the quantity it is said to measure. Uncertainty can exist even in the absence of error, in the sense of “mistake”.

2. A central source of uncertainty is definitional incompleteness in specifying all of the factors influencing the measurement.
3. What we actually measure is usually only a proxy for what we want to measure.
4. How we control, measure and report the values in a test must reflect how we expect those values to be used by others. In other words, our testing and reporting must take into account, and state, how we expect the results to be used.

Improving our ability to understand, predict and control the performance of biometric systems in operational environments will require more through study of the human and environmental influence variables, greater attention to definitional completeness in our measurands, and more experience in estimating the systematic uncertainties introduced by changes to the software, hardware, environmental and human factors encountered in the "real world".

References

[1] J. Philips, A. Martin, C. Wilson and M. Pryzbocki, "An Introduction to the Evaluation of Biometric Systems", *IEEE Computer*, Vol. 33, No. 2, February 2000

[2] Aristotle, Organon (compiled as such around 40BC)

[3] F. Bacon, Novum Organum ,1620

[4] J.S.Mill, System of Logic Ratiocinative and Inductive: Being a connected view of the principles of evidence and the methods of scientific investigation , 1843

[5] P. Feyerabend, Against Method , New Left Books, 1975

[6] P. Duhem, The Aim and Structure of Physical Theory, (English translation) Princeton University Press, 1954

[7] W.V. Quine, "Two Dogmas of Empiricism", *Philosophical Review* 60, 1951

[8] M. Curd and J.A. Cover (eds.), Philosophy of Science: The Central Issues , W.W. Norton, 1998

[9] D. Gilles, "The Duhem Thesis and the Quine Thesis", in M. Curd and J.A.Cover, ibid.

[10] S. Roush, Tracking Truth: Knowledge, Evidence, and Science, Oxford University Press, 2005

[11] Paul K. Moser, ed., The Oxford Handbook of Epistemology, Oxford University Press, 2005

[12] W. A. Shewhart, Statistical Method from the Viewpoint of Quality Control, The Graduate School, U.S. Department of Agriculture, Washington, DC, 1939

[13] I. Olkin, "A Conversation with Churchill Eisenhart", *Statistical Science* 7(4), 1992

[14] C. Eisenhart, "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems" , *J. Res. Natl. Bur. Stand.* **67C,** 161-187, 1963

[15] ISO/IEC, International vocabulary of metrology — Basic and general concepts and associated terms (VIM), ISO/IEC Guide 99, 2007

[16] H. H. Ku, "Expressions of Imprecision, Systematic Error, and Uncertainty Associated with a Reported Value", *Measurements & Data,* **2** (4), 72-77, 1968

[17] H. H. Ku, (ed.), Precision Measurement and Calibration — Statistical Concepts and Procedures, *Nat. Bur. Stand. Spec. Pub. 300*, Vol. 1,US Govt. Printing Office, Washington, 1969

[18] Churchill Eisenhart, Harry H. Ku, and R. Colle´, *Expression of the Uncertainties of Final Measurement Results: Reprints,* NBS Special Publication 644, National Bureau of Standards, Washington, DC, 1983

[19] Corrected and reprinted as ISO/IEC, "Guidelines for the Expression of Uncertainty in Measurement", ISO/IEC Guide 98, 1995, available for download at www.bipm.org/en/publications/guides/gum.html

[20] B. Taylor and C. Kuyatt, Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, *NIST Technical Note 1297*, National Institute of Standards and Technology, Gaithersburg, MD, 1994

[21] Joint Committee for Guides in Metrology (JCGM), "GUM 1995 with minor corrections", JCGM 100:2008, September, 2008

[22] "Statements of Uncertainty Associated With Measurement Results," Appendix E, NIST Technical Communications Program, Subchapter 4.09 of the Administrative Manual

[23] Joint Committee for Guides in Metrology, "Evaluation of measurement data — Supplement 1" to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method, JCGM 101:2008, BIPM, 2008

[24] J. Neyman, "Outline of a Theory of Statistical Estimation Based on the Classical Theory of Probability", *Philosophical Transactions of the Royal Society of London* A, 236**,** 333–380, 1937

[25] Mark J. Schervish, Theory of Statistics, Springer, 1995

[26] NIST/SEMATECH, e-Handbook of Statistical Methods, , National Institute of Standards and Technology, Gaithersburg, Maryland, 2010, available at www.itl.nist.gov/div898/handbook/

[27] L. J. Gleser, "Assessing Uncertainty in Measurement", *Statistical Science*, volume 13, pages 277-290, August, 1998

[28] A. Mansfield (ed.) "Information technology -- Biometric performance testing and reporting -- Part 1: Principles and framework", ISO/IEC 19795-1:2006, Geneva (2006)

[29] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The DET curve in assessment of detection task performance", in *Proc. EuroSpeech*, pp. 1895-1898, 1997, available at *www*.itl.nist.gov/iad/mig//publications/storage_paper/det.pdf

[30] C. Watson, C. Wilson, M. Indovina, and B. Cochran, " Two Finger Matching With Vendor SDK Matchers", *NISTIR 7249*, July 2005

[31] C. Wilson, et al, "Fingerprint Vendor Technology Evaluation 2003 – Summary of Results and Analysis Report", *NISTIR 7123*, June 2004

[32] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report", *NISTIR 6965*, March 2003

[33] M. Theofanos, B. Stanton, and C.A. Wolfson, Usability and Biometrics: Ensuring successful biometric systems, NIST, June 11, 2008, available at http://zing.ncsl.nist.gov/biousa/docs/Usability_and_Biometrics_final2.pdf

[34] K. Bowyer, S. Baker, A. Hentz, K. Hollingsworth, T. Peters and P. Flynn, "Factors That Degrade the Match Distribution In Iris Biometrics", *Identity in the Information Society* (to appear)

[35] B. Draper, J.R. Beveridge, Y.M. Lui, D. Bolme, and G. Givens, "Quantifying How Lighting and Focus Affect Face Recognition Performance", Colorado State University, (to appear)

[36] Holmes, J., Wright, L., and Maxwell, R., "A Performance Evaluation of Biometric Identification Devices", Sandia National Laboratories Report SAND91-0276, 1991

[37] J.R. Rodriguez, F. Bouchier and M. Ruehie, "Performance Evaluation of Biometric Identification Devices", Preliminary Draft, Sandia National Laboratories Report SAND93-1930, 1993

[38] A. Mansfield, G. Kelly, D. Chandler, and J. Kane, "Biometric product testing final report", National Physical Laboratory, 2000, available at www.cesg.gov.uk/policy_technologies/biometrics/media/biometrictestreportpt1.pdf

[39] A. Mansfield and J. Wayman, "Best Practices of Testing and Reporting Biometric Device Performance: Version 2.01", CESG/Office of the e-Envoy, 2002

[40] J.L. Wayman, "Evaluation of the INSPASS Hand Geometry Data", in J.L. Wayman (ed.) National Biometric Test Center Collected Works: 1997-2000, San Jose State University, August, 2000, on-line at www.engr.sjsu.edu/biometrics/nbtccw.pdf

[41] J.L. Wayman, "Biometrics at the Sydney Airport: Evaluation of the SmartGate Trial", *Proceedings of the Kyoto University Biometrics Symposium*, Kyoto, Japan, Jan. 16, 2006

[42] A. Fejfar and J. Myers, "The Testing of 3 Automatic ID Verification Techniques for Entry Control", *2nd International Conf. on Crime Countermeasures*, Oxford, 25-29 July, 1977

[43] "Uniting And Strengthening America By Providing Appropriate Tools Required To Intercept And Obstruct Terrorism (USA Patriot) Act of 2001", Public Law 107-56, Sec. 403(c)