Foundations of uncertainty in evaluation of nominal properties

Luca Mari[a*], Claudio Narduzzi[b*], Gunnar Nordin[c*], Stefanie Trapmann[d*]
[a] Università Cattaneo – LIUC, Castellanza, Italy
[b] Università degli Studi di Padova, Padova, Italy
[c] Equalis, Uppsala, Sweden
[d] European Commission, Joint Research Centre, Geel, Belgium

Abstract

Measurement uncertainty is a key component of metrology but, as it is defined, it does not apply to nominal properties. The possibility to define, evaluate, and express the uncertainty in the examination of nominal properties is then a critical prerequisite for a harmonized treatment of nominal properties in metrology. The assumption at the basis of this paper is that examination uncertainty can be understood in analogy with and as a generalization of measurement uncertainty. To this aim a foundational framework is introduced, grounded on a generic concept of evaluation uncertainty that applies equally to quantitative and non-quantitative evaluations. Based on this, a concept of examination uncertainty is presented and some examples of mathematical functions of examination uncertainty are proposed.

Keywords: nominal property; examination; uncertainty

1. Introduction

Measurement uncertainty is a well-established tool for modelling and reporting the quality of measurement results. Not surprisingly, then, widening the scope of metrology to new fields arises the question whether measurement uncertainty, or an appropriately generalized version of it, can be applied also to such new fields. The evaluation of non-quantitative properties – which is usually not considered a measurement (a notable exception is [Possolo 2015: p.12]) – is one of these fields, where properties are called "categorical". A basic distinction is drawn: "many categorical variables have only two categories [and] are called *binary* variables. When a categorical variable has more than two categories, we distinguish between two types of categorical scales. Variables having categories without a natural ordering are [...] called *nominal* variables. [...] Many categorical variables do have ordered categories. Such variables are [...] called *ordinal* variables." [Agresti 2013: p.2] (other traditional classifications are presented in [Stevens 1946] and [Carnap 1966]).

The result of evaluations of nominal properties is affected by uncertainty as well, and currently there is no consensus on how to evaluate and express such uncertainty. The fact that often the results of nominal property evaluations are reported as single values – for example when the result of the examination of the blood type of an individual is A in the ABO system – only hides the problem, by implicitly conveying the misleading message that such results are in any case certain. Rather, a properly generalized *evaluation uncertainty* should be applicable also in the case of nominal property evaluations, a critical condition for adopting the principles and methods of metrology: "irrespective of whether the assignment of value to nominal properties should or should not be called measurement, the need is both clear and present for methods to evaluate the uncertainty associated with such assignments" [Possolo 2014: p.S231].

In the path toward a metrological treatment of nominal properties, a critical step is indeed the modelling and quantification of evaluation uncertainty, whose applications are becoming more and more important. For example, there is an increasing request for reference materials certified for nominal properties, such materials being needed for the quality control of examinations, in applications such as disease control (anti-microbial resistance), food fraud (fish speciation, food adulteration), doping control (chemical structure). As a consequence, reference material producers, seeking to fulfil the general requirements for reference materials producers [ISO 2016], assign nominal property values with uncertainties or certify the reference material for another quantitative property while a non-certified nominal property will be used by the users of the reference material. Furthermore, evaluation uncertainty would be useful, for example, "in determinations of identity of substance (is it salicylamide or is it aspirin?), and in forensic studies, including matching hairs and fibres, comparing bite marks and shoe prints, examining firearm tool-marks, and in serological studies"

[*] The authors are members of the Joint Committee for Guides in Metrology (JCGM) Working Group 2 (VIM). The opinion expressed in this paper does not necessarily represent the view of this Working Group.

[Possolo 2014: p.S231] (indeed, serological studies include a vast number of non-quantitative methods: the mentioned case of blood typing is just an example of them).

While approaches for the evaluation of measurement uncertainty are accessible and highly standardised, a general treatment of examination uncertainty is still to be developed, and in many cases only an indication about the reliability of the assigned nominal property value is given. Even though some papers in the field of metrology have already touched on the subject (e.g., [Ellison et al 1998], [Watanabe 2005], [Possolo 2014], [Mari 2017], [Possolo, Iyerb 2017]), also in reference to specific applications (e.g., [Possolo 2015: E6], [Trapmann et al 2017]), examination uncertainty still requires a foundational study. In such a situation a framework like the one we propose in this paper seems to be useful, in which measurement uncertainty is generalized to an evaluation uncertainty, thus also applicable for nominal property evaluations.

The terminology in the field of nominal properties is still not standardised, and in particular the term "examination" is sometimes used for both quantitative and qualitative evaluations [ISO 2012, 3.7]. Here it is used in compliance with the *Vocabulary on nominal properties*, a recent IFCC-IUPAC Recommendation, whose phrasing explicitly mirrors the definition that the JCGM *International Vocabulary of Metrology* (VIM) gives of 'measurement': 'examination' is defined as "process of experimentally obtaining one or more nominal property values that can reasonably be attributed to a nominal property" [Nordin et al 2018: 2.6]. The adoption of the principles and methods of metrology would make examination results accountable in their trustworthiness [Mari 2017].

This paper aims to contribute to this purpose, based on the assumption that *examination uncertainty can be understood in analogy with measurement uncertainty*. Hence, what in the last decades has been developed about measurement uncertainty – as presented in particular in the JCGM *Guide to the Expression of Uncertainty in Measurement* (GUM) documents [JCGM 2008] – can be taken as a rich, structured set of lessons learned toward the application of metrological concepts to examination. We are aware that measurement uncertainty still arises controversial issues, as discussed for example by [Thompson 2012], [Ye et al 2016], and [Grégis 2019]. In the trade-off between drawing a parallel with JCGM documents and complete freedom of exploration, we opted for the former, plausibly at the price of inheriting some flaws that future editions of the VIM and the GUM might fix. Our justification is that several communities from different scientific fields are interested and involved in the matter: we believe that, at this stage of the development process, the VIM and the GUM provide a precious common ground for mutual understanding and a sufficiently appropriate point to start from.

The paper is structured as follows. Section 2 proposes a conceptual framework in which measurement uncertainty is interpreted as a specific case of evaluation uncertainty. This provides the context for introducing in Section 3 an example of nominal property examination, on which Section 4 develops a concept of examination uncertainty, then applied in Section 5 to some specific cases of mathematical functions which evaluate examination uncertainty. Section 6 further broaden the scope of the framework, by showing examination uncertainty may be coupled with a second parameter – which could be called "examination confidence" or "examination reliability" – to convey more complete information on the quality of examination results.

2. Uncertainty of measurement as a quantifiable attribute and its specifications

In its most generic meaning, 'uncertainty of measurement' "as a quantifiable attribute" [JCGM 2008: 0.2] is an informal concept, related to the "doubt about the validity of the result of a measurement" [JCGM 2008: 2.2.1] due to the "dispersion of the values being attributed to a measurand, based on the information used" [JCGM 2012: 2.26]. More operatively, this general concept of uncertainty could be understood as *doubt about which values should be reported in the measurement result*. In order to make it mathematically tractable, such a "quantifiable attribute" needs then to be quantified. While more general approaches have been proposed (see Ferrero, Salicone 2006), the GUM framework assumes that probability distributions are the basic mathematical tool for this quantification (the GUM maintains a distinction between statistical distributions and probability distributions, as referred to the way they are generated (see [JCGM 2008: 2.2.3]); since also the former fulfil Kolmogorov axioms, the term "probability distribution", and "distribution" for short, can be used for referring to both).

The definition of *measurement uncertainty*, adapted from the GUM, that the VIM gives [JCGM 2012: 2.26],

> "non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used"

is “an operational one that focuses on the measurement result and its evaluated uncertainty” [JCGM 2008: 2.2.4]. Though still generic, what is defined here is a mathematical entity: measurement uncertainty as a parameter, and more generally an index, of a distribution (as defined in [ISO 2006: 2.9], a parameter is an “index of a family of distributions”; in reference to this definition, measurement uncertainty – and then examination uncertainty – may be then more generally considered as an index, so as to encompass the (non-parametric) cases in which the distribution does not belong to any identified family, the usual situation in examinations). In reference to the condition that “when reporting the result of a measurement of a physical quantity, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its reliability” [JCGM 2008: 0.1], the underlying hypothesis is that a “quantitative indication of the quality of the result” is effectively provided by an index of the distribution that, explicitly or implicitly, summarizes the information acquired on the measurand. On this basis some specific indexes can be chosen, as is the case of *standard measurement uncertainty* [JCGM 2012: 2.30], defined as

> “measurement uncertainty expressed as a standard deviation”

Reporting the information on the quality of a measurement result in terms of standard uncertainty might not always be adequate. The concept ‘measurement uncertainty’ is not exhausted by standard uncertainty, which is just one of the several possible examples of measurement uncertainty. The GUM framework provides in fact a richer array of options to characterise uncertainty, and in particular *expanded measurement uncertainty* [JCGM 2008: 3.3.7]:

> “To meet the needs of some industrial and commercial applications, as well as requirements in the areas of health and safety, an expanded uncertainty U is obtained by multiplying the combined standard uncertainty u_c by a coverage factor k. The intended purpose of U is to provide *an interval about the result of a measurement that may be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand*.”

Expanded uncertainty [JCGM 2012: 2.35] involves a different, and often more useful, attitude towards “dispersion of the values”, as witnessed by the fact that measuring instrument specifications are typically given in terms of expanded uncertainty and level of confidence. Other examples of measurement uncertainties which are not standard uncertainties include half-width of an interval having a stated level of confidence (as in JCGM 2008: 2.2.3 Note 1) and interquartile range (i.e., the difference between the third and the first quartile of the distribution).

There is a critical point to remark here: measurement uncertainty is unavoidably related to what is a measurement result, and the very idea of *what is a measurement result* has been changing in the last decades. The GUM – originally published in 1993 – adopted the traditional approach of considering that the measurement result is *the value* attributed to the measurand (see in particular JCGM 2008: 4.1.4), as in the second edition of the VIM, which defined ‘result of a measurement’ as “value attributed to a measurand, obtained by measurement” [ISO 1993: 3.1]. Under this assumption, the acknowledgement that generally we cannot be certain about one value to attribute to a measurand – so that “a complete statement of the result of a measurement includes information about the uncertainty of measurement” [ISO 1993: 3.1 Note 2] – leads to the concept ‘uncertainty *of* the measurement result’. Uncertainty is then uncertainty of a given value, in its role of summarizing the available information about the measurand. In this perspective an appropriate index of a distribution provides a summary of such information, but of course “measurement uncertainty *is described* fully and quantitatively by a probability distribution on the set of values of the measurand” (emphasis added) [Possolo 2015]. Along the same path Thompson [2012] is even more direct in stating that “the uncertainty of a result *is* the density function (or mass function) that best describes the probability of possible values of the measurand” (emphasis added).

One of the key changes introduced in the third edition of the VIM [Mari 2015] is a new, more encompassing definition of ‘measurement result’ as a “set of quantity values being attributed to a measurand together with any other available relevant information” [JCGM 2012: 2.9]. Here measurement results are intended in a generalized sense, as what results from a measurement, and therefore including the information related to measurement uncertainty. The “relevant information” about the set of values is “such that some may be more representative of the measurand than others” so that “this may be expressed in the form of a probability density function” [JCGM 2012: 2.9 Note 1]. But if the measurement result is (or at least may be) the distribution, not one value, then what is measurement uncertainty needs to be reconsidered, where the concept is now ‘uncertainty *within* the measurement result’, given that the “available relevant information” which is a component of a measurement result generally includes measurement uncertainty. This justifies the

position, of both the GUM and the VIM, that measurement uncertainty is an index of a distribution. From this distribution one representative value may be chosen – what the VIM calls "measured value", the "quantity value representing a measurement result" [JCGM 2012: 2.10] – so that the uncertainty *within* the measurement result may be interpreted as the uncertainty *of* the chosen measured value.

On this basis two other points are worth to emphasize.

First. If measurement uncertainty is an index that provides summary information on the dispersion of the values in a measurement result, and if the underlying probability distribution is known, then measurement uncertainty is usually sufficient to derive the confidence level for the chosen set of measured values.

Second. In these distributions no conditioning elements, such as true values, are explicitly included (whereas, as usual, an implicit conditioning element is the available knowledge). Hence, measurement uncertainty *may* be intended as the uncertainty about the true value of the measurand, as in [Possolo 2015: p.6], but the concept is compatible also with true-value-agnostic positions, that may interpret it – as mentioned above – as doubt about which values should be reported in the measurement result: "uncertainty of measurement is the doubt that exists about the result of any measurement" [Bell 1999: p.1]. This rules out from our consideration any modelling that assumes an explicit dependence of the distribution on one or more conditioning values, be them true values or just generic reference values (an example of this more specific approach in the case of ordinal properties is in [Bashkansky, Gadrich 2010], which assumes that "in order to understand how to evaluate the uncertainty of an ordinal measurement result, one needs to know the likelihood that a measured level *j* is received whereas the true level is *i*").

Since "the word "uncertainty" means doubt, and thus in its broadest sense "uncertainty of measurement" means doubt about the validity of the result of a measurement", the GUM acknowledges that "because of the lack of different words for this general concept of uncertainty and the specific quantities that provide quantitative measures of the concept, for example, the standard deviation, it is necessary to use the word "uncertainty" in these two different senses." [JCGM 2008: 2.2.1]. This interpretation may be refined by understanding uncertainty in/of measurement as admitting three layers of specification:

L1. *a generic concept*: uncertainty as a quantifiable attribute;

L2. *a mathematical concept*: uncertainty as a generic quantitative attribute, corresponding to a yet unspecified index of a distribution;

L3. *several specific mathematical concepts*: uncertainty as a given quantitative attribute, related to a specified index of a distribution.

This is the framework that we aim to apply, through an appropriate generalization, to examination. While pursuing this, we also consider concepts that are "applicable to evaluating and expressing the uncertainty associated with the conceptual design and theoretical analysis of experiments, methods of measurement, and complex components and systems" [JCGM 2008: 1.3], which is a key feature of the GUM.

3. Background and example

A simple representative example may be useful to help understand the analysis that follows. First the *examinand* has to be defined, i.e., the "property intended to be examined" according to the *Vocabulary on nominal properties* [Nordin et al 2018: 2.7], in intentional analogy with measurement and the VIM, where the property to which the measurement result is attributed is called the "measurand", i.e., a "quantity intended to be measured" [JCGM 2012: 2.3].

Let us assume that a person or a technological system has to determine the character written in a given ink pattern, a task called "optical character recognition" (OCR) in the context of Information Technology. The implementation of an OCR system is a complex task (see, e.g., Mori et al, 1999), due to the fact that:

– character recognition is a multi-faceted task, where a variety of factors may come into play; for example, even disregarding hand-writing recognition, character shapes still vary depending on size, font family, font type, and attributes (e.g., bold and italic);

– typical groupings of characters often occur, which may affect character recognition performances in a language-dependent way;

– the same shape may correspond to different characters in different alphabets, as in the case of Latin 'P' and Greek/Cyrillic 'P', thus showing the need of a reference to the context in the examination.

In spite of this, the input-output description of the process (i.e., its black box model) is analogous to a generic description of a measurement process:

– the input is an object considered with respect to a given empirical property, in this case ink on paper considered with respect to the shape of the ink pattern;

– the output is information about that property, in this case the recognized and assigned character.

Given our aims, we shall not "open the black box" here, as we focus only on the outcome of the recognition process (incidentally, this implies that we shall not discuss examination uncertainty budgets; for an in-depth analysis of the OCR techniques see, for example, [Mori et al, 1999]).

Like measurement, OCR can be abstractly intended as an evaluation, i.e., a value attribution [Mari 2013], of the shape of any given ink pattern, where the possible characters ("a", "b", "c", and so on), as elements of a predefined set, are the possible values. Hence the shape of a given ink pattern could be reported by OCR as the character "a" of a given alphabet, called an "examined value" [Nordin et al 2018: 3.5], much like the reported length of a given rod is, e.g., 1.234 m, called a "measured value" [JCGM 2012: 2.10].

In the same way that measurement results can be compared if they are traceable to the same unit, comparison among examination results requires reference to the same *classification system*, i.e., to the same set of reference values, each of them being the identifier of a class of equivalent nominal properties (a major difference between measurement and examination is that for the latter nothing like the International System of Quantities (ISQ) [JCGM 2012: 1.6] and the International System of Units (SI) [BIPM 2014] exist: examinations typically refer to established classification systems which in some cases may change due to knowledge improvement). We refer in the following to a basic Latin alphabet and, for simplicity, assume as the classification system the set of 26 lowercase and the 26 uppercase characters, disregarding punctuation marks. We also do not consider the distinctions among other features such as font types (e.g., Arial and Times), attributes (e.g., italic and bold), or size, that would be part of the detailed examination model for the evaluation of uncertainty. While conceptually simpler to discuss, such feature-independence may make OCR a more complex task in practice. Of course, character recognition might involve and exploit order or metric information, e.g., lengths of stems, radii of curvature of rings, suitably defined distances among patterns, etc., but this remains "within the black box".

What differentiates examination from measurement, as shown by the OCR example, is that the examinand cannot be compared to a unit, and a property-related ordering cannot be defined: claims that the shape of an ink pattern is greater than another, or that one shape is, say, double than another have no empirical significance in the context of OCR. Rather, outcomes of the process are purely classificatory, and the values attributed to the shapes are themselves elements of a set with no algebraic structure. Hence, shape is not a measurable property in the sense of the VIM [JCGM 2012: 2.1].

Despite these differences, the fundamental idea of metrology – that a measurement must provide not only one or more values to be attributed to the measurand, but also some quantitative information about the quality of this attribution – applies also to examination. This seems to be indeed a key condition for the application of metrological concepts/approaches/understanding to examinations, that triggers the next step in our analogical consideration of examination with respect to measurement. Let us paraphrase the statement of [JCGM 2008: 0.1] (changes are underlined):

> When reporting the result of an examination of a nominal property, it is obligatory that some quantitative indication of the quality of the result be given so that those who use it can assess its reliability. Without such an indication, examination results cannot be compared, either among themselves or with reference values given in a specification or standard. It is therefore necessary that there be a readily implemented, easily understood, and generally accepted procedure for characterizing the quality of a result of an examination, that is, for evaluating and expressing its *uncertainty*.

4. From measurement uncertainty to examination uncertainty

Once the set of possible values for the nominal property under consideration has been chosen, the core idea of examination uncertainty is not different from the one at the basis of measurement uncertainty. An examination process may be formally described as a mapping from a set of nominal properties of objects to the chosen set of values, or to a more complex structure derived from it. In our OCR example, each shape s in a set S is expected to be recognized as a character c in the chosen reference set of characters. Ideally, the process maps each shape to a character, and therefore it is formalized by a recognition function $\varrho: S \rightarrow C$,

$c=\varrho(s)$, i.e., shape of $s = c$ in the set C (where, as mentioned above, the mapping could result from the composition of multiple intermediate mappings, which might involve extracting and then exploiting measurable features). This induces on S a partition into subsets S_j, where two shapes s and s' belong to the same subset, and therefore are recognized as equivalent, $s \approx s'$, if and only if they are mapped to the same character, $\varrho(s)=\varrho(s')$.

It is a fact that the recognition of some shapes may be less than ideal for a variety of reasons. In the OCR example introduced above, these might include insufficient scanner pixel resolution, blurred edges of the printed or written character, and so on. In these cases it might happen that a given shape is not recognized at all, or that it is mapped to more than one character, possibly each of them with an associated probability of recognition. As a consequence, the recognition function becomes more complex, its range being the set of the subsets of C, possibly extended with one element for all cases of non-recognition, or even the set of the probability distributions on C: it is then legitimate to ask what uncertainty can be attributed to the examination result in these cases, and how it can be reported.

In discussing a definition of 'examination uncertainty', we shall undertake to fulfil the following three conditions:

(1) in analogy with 'measurement uncertainty' as discussed in Section 2, 'examination uncertainty' is structured as a three-layer concept: a generic concept of uncertainty (L1) is quantified into a quantitative attribute of examination uncertainty (L2), that may correspond to several possible specific indexes of a distribution (L3);

(2) it generalizes measurement uncertainty, so as to maintain the well-known conceptual and formal hierarchical structure of invariant conditions among statistics, those that Stevens called "permissible statistics" [Stevens 1946] and such that, for example, the mean can be computed only for numerical distributions but the mode can be computed for both categorical and numerical distributions;

(3) it is general enough to include cases in which an examination result is reported as either a probability distribution of values or a set of values, somehow extracted from the distribution and possibly reduced to a single value, in analogy with measurement results: "If the measurement uncertainty is considered to be negligible for some purpose, the measurement result may be expressed as a single measured quantity value. In many fields, this is the common way of expressing a measurement result." [JCGM 2012: 2.9 Note 2] (measurement results reported as single values may be interpreted as implicitly conveying an uncertainty through their number of significant digits; due to the absence of algebraic structure, nothing similar applies to examination results).

Like in the case of measurement, the generic concept of uncertainty in nominal examination (L1) has to do with the quality of the examination result, the information provided by the quantitative specifications (L2 and L3) depending on the way the result is reported:

– if a distribution is reported as the result, an index of examination uncertainty allows us to compare different results, and therefore different distributions, so as to establish their relative uncertainty;

– if instead the result is a single value or a set of values chosen from the distribution, an index of examination uncertainty gives us some information about the specificity of the choice: the greater the uncertainty, the less the specificity; in this case a properly generalized version of the concept 'statistical coverage interval' – "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" [JCGM 2008: C.2.30] – may be adopted to refer to the set, where the lack of algebraic structure of the set of values requires "interval" to be substituted with "subset". With the same substitution, the concept 'confidence coefficient', or 'confidence level' – "the value $(1 - \alpha)$ of the probability associated with a confidence interval or a statistical coverage interval" [JCGM 2008: C.2.29] – is appropriate also for examination results. In fact, the non-parametric nature of the distributions defined over sets of values of nominal properties makes the reference to confidence levels even more important than in the case of measurement: while the confidence level of a coverage interval obtained from a parametric distribution may be typically computed, in the case of examination results the choice of the coverage subset is generally not sufficient to know the confidence level, which needs then to be explicitly specified.

This shows that examination uncertainty may be conceived in analogy to measurement uncertainty also at the quantitative level (L2). However, while measurement uncertainty refers to "dispersion of values", it is arguable whether the concept 'dispersion' applies to nominal properties. For quantities, being more or less dispersed usually means 'being more or less distant, from each other or from a given point'. The reference to a distance, and therefore to a structure in the set of values, suggests that applying the concept 'dispersion' to

nominal properties might be inappropriate (in fact, the *Vocabulary on nominal properties* uses ‘dispersion’ also in the case of nominal properties – for example when noting that “examination precision can be expressed numerically by measures of dispersion of examined values” [Nordin et al 2018: 3.12] – thus showing that the subject is not settled down). Alternative terms to “dispersion” might be “discrepancy”, “diversity”, or “variation”. Our task here is to propose a conceptual framework, not a set of definitions: as a placeholder, we adopt here “variation” (as in [Wilcox 1967], but without the adjective “qualitative”, which could be misleading), to be intended in the sense of ‘variation internal to the distribution’. In reference to our OCR example, we may say then that the result R_1 = {“a”, “d”, “o”} has a greater variation (or discrepancy, or diversity, or, possibly, dispersion) than R_2 = {“a”, “d”}. By a slightly modified version of [JCGM 2012: 2.26], a simple draft definition is then (changes are underlined):

examination uncertainty: index characterizing the variation of the values being attributed to an examinand, based on the information used

The comparison between measurement and examination that we have drawn so far can be summarized as in Table 1.

Table 1: Side by side comparison of a measurement and examination, including the treatment of measurement uncertainty and examination uncertainty.

In the case of measurement,	**In the case of examination**,
the length of a given rod	the shape of a given ink pattern
is an example of a *measurand*.	is an example of an *examinand*.
A *measured value* for the measurand is, e.g.,	An *examined value* for the examinand is, e.g.,
1.234 m.	“a” in the set of Latin characters.
In presence of *measurement uncertainty*	In presence of *examination uncertainty*
a *measurement result* is, e.g.,	an *examination result* is, e.g.,
the interval of values 1.2340±0.0005 m,	the subset of values {“a”, “d”, “o”} in the set of Latin characters,
possibly with an associated level of confidence,	possibly with an associated level of confidence,
and more generally it is a probability distribution defined on the set of possible values.	and more generally it is a probability distribution defined on the set of possible values.

This summary may be then framed in the context of the proposed three layers of specification, as in Table 2.

Table 2: Side by side comparison of a measurement uncertainty and examination uncertainty, in reference to the proposed three layers of specification.

Layer	**Uncertainty of an evaluation as**	**In the case of measurement:**	**In the case of examination:**
L1	a quantifiable attribute of the evaluation result	the same (a quantifiable attribute of the measurement result)	the same (a quantifiable attribute of the examination result)
L2	an index of the distribution that is the evaluation result	the same (an index of the distribution that is the measurement result)	the same (an index of the distribution that is the examination result)
L3	a specific index of the distribution	e.g., standard deviation, interquartile range	e.g., f_1, f_2, as defined in section 5

On this basis specific indexes can be defined (L3) to provide quantitative information on the uncertainty of examination results (ordinal properties can be interpreted as nominal properties for which an ordering is also meaningful [Agresti 2013]: hence what follows also applies to ordinal properties, although more specific statistical techniques, e.g., indexes based on distribution percentiles, can be exploited for dealing with ordinal properties).

5. Some specific cases of examination uncertainty

Let us consider the general case in which the result of examining a nominal property s is a probability mass function $R = \{(c_j, p_j)\}$ defined on a set of values $C = \{c_j\}$, where p_j is the probability that the nominal property s is examined to be the value c_j. For OCR, this corresponds to a distribution such as

R = {(“a”, 0.80), (“d”, 0.15), (“o”, 0.05), (any other character in the set C, 0.00)}

that might be generated by repeating the recognition of the same shape s and obtaining, for example, the character c_1 = “a” 8 times out of 10 and so on, or by either a person or an OCR system attributing a probability of recognition of the shape, in this case $p_1 = 0.8$ to “a” and so on (how such probabilities can be obtained depends of course on the specific application; for example, in the case of DNA sequencing Possolo shows how the probabilities of the nucleobases at any given location can be computed from “quality scores” [2015: E6], and a technique for estimating the probabilities of correct identification of chemical substances is presented by Stein [1994]). An explicitly probabilistic information might not be available, and in this case the person or the OCR system could report a subset of C as the result, e.g., R_1 = {“a”, “d”, “o”}: this can be interpreted as the implicit assumption of a uniform distribution, {(“a”, 0.33), (“d”, 0.33), (“o”, 0.33), (any other character in the set C, 0.00)}. In the case of examination results that are subsets, their comparison by variation is trivial. For example, in reference to the results R_2 = {“a”, “d”} and R_3 = {“a”}, where the subscript denotes different examinations of the same shape, as a general (L1) description, we may conclude that there is more variation in R_1 than in R_2, and that in R_3 there is no variation at all, so that according to R_3 the given shape is claimed to have been recognized as the character “a” with certainty.

The idea of a variation index may be then formalized by a set function (L2), and the statistical literature provides a variety of indexes that may be considered for this purpose. After [Wilcox 1967], a simple choice, applicable to examination results formalized as subsets, is a function whose minimum value is 0, in the case of no uncertainty (i.e., if $\#R = 1$, where $\#R$ is the cardinality of R), and monotonically increasing with the cardinality of the subset. If a normalization condition is added, so as to obtain the value 1 in the case of complete uncertainty (i.e., if $\#R = \#C$), a first specific (L3) instance of examination uncertainty is the function:

$$f_1(R) =_{\text{def}} (\#R - 1) / (\#C - 1)$$

where the term $\#C - 1$ is a normalization factor, such that uncertainty ranges from 0 to 1 (another option might be to divide by $\#C$, so that uncertainty ranges from 0 to the probability that can be associated to the set of "wrong" values, still assuming all elements in C are equiprobable. As mentioned above, we are assuming a set C of 26 lowercase and 26 uppercase characters in a basic Latin alphabet, with one more element to account for all other possibly unidentified shapes, so that $\#C = 53$). Hence $f_1(R_1) = 2/52$, $f_1(R_2) = 1/52$, and $f_1(R_3) = 0$.

Another and more general option to evaluate examination uncertainty is based on information entropy, as already suggested by [Possolo 2015]: “for nominal (or, categorical) properties, the entropy of the corresponding probability distribution is one of several possible summary descriptions of measurement uncertainty” (where, interestingly, measurement is assumed to encompass also the evaluation of nominal properties). For an examination result $R = \{(c_j, p_j)\}$, entropy $H(R)$ is defined as

$$H(R) =_{\text{def}} -\Sigma_j\, p_j \log(p_j)$$

with

$$0 \le H(R) \le \log_b(\#C)$$

the largest value being obtained when all values have equal probability, i.e., when $\{p_j\}$ is again a uniform distribution. Any logarithm basis b can be considered, and it can be observed that the inequality above yields

$$b^0 = 1 \le b^H \le \#C$$

As a specific (L3) index $H(R)$ could be taken, but another interesting candidate is

$$f_2(R) =_{\text{def}} b^{H(R)} - 1$$

or its normalized version $(b^{H(R)} - 1) / (\#C - 1)$. Both H and f_2 associate smaller uncertainty with smaller values, as expected of an uncertainty index, and do indeed characterize variation, and both provide meaningful information about the uncertainty in the examination result. The latter, being independent of b and comparable to $\#C$, might be more directly understood and could perhaps be preferred. Indeed, if R contains a single value with probability 1, then $f_2(R) = 0$, i.e., there is no uncertainty. At the opposite end, if the examination result R includes all values in C with equal probability, all $\#C - 1$ values but the first one contribute to uncertainty, which is indeed maximal, $f_2(R) = \#C - 1$. For all intermediate examination results, $b^{H(R)}$ can be non-integer, given that the values in R are not merely counted by $f_2(R)$, but suitably weighted according to their probability. Then, $b^{H(R)}$ can be interpreted as an equivalent number of possible values, and $b^{H(R)} - 1$ provides the equivalent number of values contributing to the examination uncertainty.

The fact that multiple indexes of examination uncertainty can be defined is not surprising, given that the same happens for measurement uncertainty, of which standard deviation and interquartile range are two examples: each index provides different information about, and has different uses related to, uncertainty. A discussion of specific indexes, their suitability and relevance is beyond the scope of this paper, but may be the subject of future works, once steady foundations have been laid.

6. Examination uncertainty and examination confidence

We have already quoted the VIM in its consideration that "if the measurement uncertainty is considered to be negligible for some purpose, the measurement result may be expressed as a single measured quantity value", with the acknowledgement that "in many fields, this is the common way of expressing a measurement result" [JCGM 2012: 2.9 Note 2]. The same happens in examinations of nominal properties, whose results are in fact usually single values. For an examination the probability distribution that formalizes the available information is defined on a finite set of values: this offers a further option for assessing the quality of the results, by simply reporting the probability that the chosen value has in the distribution. For example, starting again from the distribution

R = {("a", 0.80), ("d", 0.15), ("o", 0.05), (any other character in the set C, 0.00)}

the examination result could be

shape s = "a" in the set C, with probability 0.80

This probability is not an index of variation, in the sense proposed above, but an index *of correct classification* (and therefore its complement, 0.20 = 1 – 0.80, is an index of mis-classification), assessed on the basis of the available information. This index, which could be called maybe "examination confidence" or "examination reliability", is complementary, not alternative, to examination uncertainty, and over uncertainty it has the advantage of being simple to understand and of offering positive (confidence, reliability), instead of negative (un-certainty) information. Furthermore, this is immediately generalized to the case in which more than one value is reported, so that for example

shape s = "a" or "d" in the set C, with probability 0.95 (=0.80+0.15)

A simple draft definition might be then:

examination confidence / examination reliability: probability of the values being attributed to an examinand, based on the information used

This complementarity is well highlighted in [Bell 1999: p.1]: "Since there is always a margin of doubt about any measurement, we need to ask 'How big is the margin?' and 'How bad is the doubt?' Thus, two numbers are really needed in order to quantify an uncertainty. One is the width of the margin, or *interval*. The other is a *confidence level*, and states how sure we are that the 'true value' is within that margin.". Basically the same applies to examination then, the only difference being that in this case the margin of doubt must be interpreted as related not to the width of an interval but to the cardinality of a subset. Hence, if a single value is reported, the confidence level remains the only relevant index.

7. Conclusions

The possibility to evaluate and express uncertainty in examination of nominal properties is a critical prerequisite for a harmonized treatment of nominal properties in metrology, and a foundational framework like the one we have proposed here contributes to this endeavour.

– Exactly as in the case of measurement, it starts (layer L1) from a concept of uncertainty as a quantifiable attribute related to the doubt about the values to be reported in an evaluation result.

– Under the hypothesis that results are probability distributions, it establishes (layer L2) that uncertainty is quantified as an index of the distribution.

– This provides the context for defining (layer L3) specific indexes, each quantifying an aspect of the complex concept of uncertainty.

In reference to the concept of evaluation as value attribution, such that both measurements and examinations are evaluations, L1 and L2 establish a common ground, on which measurement-specific and examination-specific indexes are defined at the layer L3. This suggested structure seems to be a useful component of a conceptual, mathematical, and operative framework for embedding the evaluations of nominal properties in metrology.

Acknowledgements

Though this paper is the authors' work and they solely have the responsibility for what it is proposed here, they gratefully acknowledge that some of these ideas emerged from JCGM/WG2, in the endeavour toward the development of the next edition of the *International Vocabulary of Metrology*.

References

A. Agresti. Categorical data analysis, Hoboken: Wiley, 3rd edition, 2013.

E. Bashkansky, T. Gadrich, Some metrological aspects of ordinal measurements, Accreditation and Quality Assurance, 15, 6, 331-336, 2010.

S. Bell, A beginner's guide to uncertainty of measurement – Measurement Good Practice Guide No. 11 (Issue 2), Teddington: National Physical Laboratory, 1999.

BIPM, SI Brochure: the International System of Units (SI), 8th ed. (2006 edition with updates), 2014, www.bipm.org/en/publications/si-brochure.

R. Carnap, Philosophical foundation of physics, New York: Basic Books, 1966.

S.L.R. Ellison, S. Gregory, W.A. Hardcastle, Quantifying uncertainty in qualitative analysis, Analyst, 123, 1155-1161, 1998.

A. Ferrero, S. Salicone, Fully comprehensive mathematical approach to the expression of uncertainty in measurement, IEEE Trans. Instr. Meas., 55, 3, 706-712, 2006.

F. Grégis, On the meaning of measurement uncertainty, Measurement, 133, 41-46, 2019.

ISO, International vocabulary of basic and general terms in metrology (VIM, 2nd ed, Guide 99:1993), published in the name of BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, and OIML, Geneva: International Organization for Standardization, 1993.

ISO, Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability (ISO 3534-1:2006), Geneva: International Organization for Standardization, 2006.

ISO, Medical laboratories – Requirements for quality and competence (ISO 15189:2012), Geneva: International Organization for Standardization, 2012.

ISO, General requirements for the competence of reference material producers (ISO 17034:2016), Geneva: International Organization for Standardization, 2016.

JCGM, Evaluation of measurement data – Guide to the Expression of Uncertainty in Measurement (GUM) (JCGM 100:2008: 1995 Edition with Minor Corrections), Sèvres: Joint Committee for Guides in Metrology, 2008, www.bipm.org/en/publications/guides/gum.html.

JCGM, 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), Sèvres: Joint Committee for Guides in Metrology, 2008, www.bipm.org/en/publications/guides/gum.html.

JCGM, International Vocabulary of Metrology – Basic and general concepts and associated terms (VIM, 3rd ed, JCGM 200:2012: 2008 Edition with Minor Corrections), Sèvres: Joint Committee for Guides in Metrology, 2012, www.bipm.org/en/publications/guides/vim.html.

L. Mari, A quest for the definition of measurement, Measurement, 46, 2889-2895, 2013.

L. Mari, Evolution of 30 years of the International Vocabulary of Metrology (VIM), Metrologia, 52, R1–R10, 2015.

L. Mari, Toward a harmonized treatment of nominal properties in metrology, Metrologia, 54, 784-795, 2017.

S. Mori, H. Nishida, H. Yamada, Optical Character Recognition, Hoboken: Wiley, 1999.

G. Nordin, R. Dybkaer, U. Forsum, X. Fuentes-Arderiu, F. Pontet, Vocabulary on nominal property, examination, and related concepts for clinical laboratory sciences (IFCC-IUPAC Recommendations 2017), Pure Appl. Chem., 1-23, 2018, https://doi.org/10.1515/pac-2011-0613.

A. Possolo Statistical models and computation to evaluate measurement uncertainty, Metrologia, 51, S228-S236, 2014.

A. Possolo, Simple guide for evaluating and expressing the uncertainty of NIST measurement results (NIST Technical Note 1900), 2015, http://dx.doi.org/10.6028/NIST.TN.1900.

A. Possolo, H.K. Iyerb, Concepts and tools for the evaluation of measurement uncertainty, Rev. Sci. Instrum., 88, 011301, 1-33, 2017.

S.E. Stein, Estimating probabilities of correct identification from results of mass spectral library searches, J. Am Soc Mass Spectrom, 5, 316-323, 1994.

S.S. Stevens, On the theory of scales of measurement, Science, 103, 677-680, 1946.

M. Thompson, What exactly is uncertainty?, Accreditation and Quality Assurance, 17, 93-94, 2012.

S. Trapmann, A. Botha, T.P.J. Linsinger, S. Mac Curtain, H. Emons, The new International Standard ISO 17034: general requirements for the competence of reference material producers, Accreditation and Quality Assurance, 22, 381-387, 2017.

H. Watanabe, Coarse-grained information in formal theory of measurement, Measurement, 38, 295-302, 2005.

A.R. Wilcox, Indices of qualitative variation (ORNL-TM-1919), Oak Ridge: Oak Ridge National Lab., 1967.

X. Ye, X. Xiao, J. Shi, M. Ling, The new concepts of measurement error theory, Measurement, 83, 96-105, 2016.