Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

The Statistical Evaluation of Categorical Measurements: "Simple Scales, but Treacherous Complexity Underneath"

Jeroen de Mast [a], Thomas Akkerhuis [a] & Tashi Erdmann [a]
[a] Institute for Business and Industrial Statistics of the University of Amsterdam, Amsterdam, The Netherlands
Published online: 11 Dec 2013.

To cite this article: Jeroen de Mast, Thomas Akkerhuis & Tashi Erdmann (2014) The Statistical Evaluation of Categorical Measurements: "Simple Scales, but Treacherous Complexity Underneath", Quality Engineering, 26:1, 16-32, DOI: 10.1080/08982112.2013.846062

To link to this article: http://dx.doi.org/10.1080/08982112.2013.846062

Quality Engineering, 26:16–32, 2014

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2013.846062

Taylor & Francis
Taylor & Francis Group

The Statistical Evaluation of Categorical Measurements: "Simple Scales, but Treacherous Complexity Underneath"

Jeroen de Mast,
Thomas Akkerhuis,
Tashi Erdmann

Institute for Business and Industrial Statistics of the University of Amsterdam, Amsterdam, The Netherlands

ABSTRACT The statistical evaluation of measurements on categorical scales is hampered by hiatuses in insight and conceptualization. Categorical scales have a simple mathematical structure. The underlying empirical reality, however, that they aim to reflect usually has a very complex structure. This complexity induces intricate challenges for the statistical evaluation of the performance of categorical measurement systems. Most current techniques deal ineffectively with these challenges, relying on simplistic conditional independence assumptions and careless sampling strategies. Moreover, they typically evaluate measurement systems in terms of concepts not clearly related to a notion of measurement error. This article proposes an approach for modeling the behavior of categorical measurements based on characteristic curves. The approach is intended to facilitate the development of more effective techniques. It is applied in a case study that illustrates what the authors believe is a realistic degree of complexity.

KEYWORDS binary measurement, categorical data, gauge capability, latent variable modeling, pass/fail inspection, repeatability, reproducibility

INTRODUCTION

The importance of the validity of measurements is generally acknowledged, and techniques for assessing the error of measurement systems are therefore an important subject of research in statistics. The evaluation of measurement systems that produce results on numerical scales is, we believe, a reasonably mature science (see, e.g., Automotive Industry Action Group [AIAG] 2003; Hunter 1980; International Organization for Standardization 1995). This does not mean that all problems have been solved. However, the evaluation of categorical measurements struggles with much more fundamental hiatuses in understanding and modeling, to the extent that we have only a slight hesitancy in claiming that a substantial part of such evaluations are misguided.[1] This article will focus

Article presented at the First Stu Hunter Research Conference in Heemskerk, Netherlands, March 2013.

Address correspondence to Jeroen de Mast, Institute for Business and Industrial Statistics of the University of Amsterdam, Plantage Muidergracht 12, Amsterdam 1018 TV, The Netherlands. E-mail: j.demast@uva.nl

[1]We are not aware of a systematic study that underpins this claim, but weak evidence at the least is provided by our findings in a review of evaluation studies in top medical journals; see Erdmann et al. (in press).

on the statistical evaluation of measurements on categorical scales, with special emphasis on binary measurement.

As we see it, overly simplistic modeling has been a substantial impediment to the development of effective and reliable techniques for the evaluation of categorical measurement systems. The purpose of this exposition is to show how the models underlying currently recommended techniques fail to capture the complexity of structures underlying categorical measurements, and we wish to put forward an alternative approach for statistical modeling and associated principles. In line with the philosophy of the Stu Hunter Research Conference, where this article was presented, the article focuses less on novel technical contributions but instead aims to explore where the field is now and make the case for a certain direction that, in the view of the authors, is a fruitful way to make progress.

To ensure that the reader has an idea what a statistical evaluation of a measurement system may comprise, and to set the stage for the discussion, we briefly describe a so-called gauge repeatability and reproducibility (GR&R) study (Montgomery and Runger 1993; Vardeman and Van Valkenburg 1999). The function of such an experiment is to estimate the standard deviation of random measurement error and two of its components: repeatability and reproducibility. In a typical setup, 10 items are measured twice by three appraisers. Assuming that the items' properties do not change during the experiment and are not affected by it, the variation across the six measurement results for a single item can be interpreted as random measurement error. The results are analyzed as realizations of a two-way random effects analysis of variance model. The residual error in this analysis is interpreted as repeatability, the variance of random measurement errors when items are measured under identical conditions and by a single appraiser. The variance components associated with the appraiser factor and the Appraiser × Item interaction effect are interpreted as reproducibility, the additional measurement variation induced by the variability of conditions and differences among appraisers. The sum of the repeatability and reproducibility components is the total measurement variance $\sigma^2_{R\&R}$. Precision is sometimes defined as the $\pm k\sigma_{R\&R}$ margins, where typically $k = 2.575$ (99% margin) or $k = 3.000$.

WHAT WENT WRONG: THE KAPPA STATISTIC

Perhaps the most popular concept for the evaluation of categorical measurements is that of agreement and the associated κ (kappa) statistic. The statistic originated in psychometrics and medicine, and seminal papers include Cohen (1960), Fleiss (1971), Conger (1980), and Kraemer et al. (2002), but the literature on the subject is extensive. The approach is generally used in engineering as well (De Mast and Van Wieringen 2007) and offered in Six Sigma courses as attribute GR&R. The statistic is included in Minitab and recommended in the AIAG's (2003) *Measurement System Analysis: Reference Manual*. Given the continuing popularity of agreement studies for the evaluation of categorical measurements, we cannot ignore this line of thinking, and it actually makes a good opening section, because it allows us to illustrate much of what in our view has gone wrong in this endeavor.

If appraisals are on a scale consisting of unordered categories, we speak of nominal measurement, and examples include the determination of failure modes of rejected products in quality control in industry or the diagnosis of patients by a radiologist on the basis of an X-ray image into a set of disorder types. We will consider an example in a call center of a bank, where calls are categorized by agents on a five-point scale. The aim of the categorizations is that management wishes to know for what reasons (and especially in what proportions) callers contact the company. Thus, the empirical property of the calls that the categorizations are meant to reflect (the measurand) is the intent of a caller. Five types of intent are discerned:

1. Service request: customer instructs bank to perform a banking service.
2. Inquiry: customer requests information.
3. Error: customer reports a mistake.
4. Error and complaint: customer reports a mistake and expresses dissatisfaction.
5. Complaint: customer expresses dissatisfaction.

Instructions for the call center agents are not much more elaborate than these five definitions. In particular, there are no operational guidelines that tell agents how to discern, say, a service request from an inquiry.

To evaluate the reliability of the categorizations made by the call center agents, one could set up an experiment that resembles the design of a GR&R study. Real evaluations would involve substantial numbers of calls and appraisers, but for simplicity let us assume an experiment in which two agents categorize the same 15 calls (recorded or transcribed) once. The raw results and a cross-tabulation are shown in Figure 1. The first two calls are rated the same category by both appraisers, and this is called *agreement*. For the third call, the appraisers are in disagreement. The traditional analysis is based on a cross-tabulation, in which the observed frequencies are compared to the frequencies expected when appraisers rate calls with the same marginal distribution as observed but totally independent from each other (chance ratings). The observed proportion of agreement is $\widehat{P}_A = \frac{1}{15}\sum_{k=1}^{5} N_{kk} = \frac{11}{15} = 0.733$ (where N_{kk} is the number of calls rated k by both agents), and the expected agreement of chance ratings is $\widehat{P}_{A|chance} = \frac{1}{15^2}\sum_{k=1}^{5} N_{k\cdot}N_{\cdot k} = \frac{3.93}{15} = 0.262$, where $N_{k\cdot}$ and $N_{\cdot k}$ are the row and column marginals. The $\hat{\kappa}$ statistic is $\widehat{P}_A$ normalized such that $\hat{\kappa} = 1$ for perfect agreement and $\hat{\kappa} = 0$ for chance ratings:

$$\hat{\kappa} = \frac{\widehat{P}_A - \hat{P}_{A|chance}}{1 - \hat{P}_{A|chance}} = 0.639.$$

Alternative definitions have been proposed for the case of more than two appraisers and different situations (e.g., Conger 1980; Fleiss 1971; and many others). Despite the simplicity of the statistic itself, the extensiveness of the literature on the subject reflects that its behavior is actually poorly understood, and its interpretation is controversial. For example, in the left data set in Figure 2, there is disagreement in one out of 15 cases, whereas in the right data set there is disagreement in 4 out of 15 cases. The reader may be surprised that $\hat{\kappa} = 0.643$ for both data sets and may wonder what information this number conveys. The traditional interpretation of the statistic is as a proportion of agreement corrected for agreement by chance, but on closer inspection (Erdmann et al. in press), chance and its tie to the marginal distribution turns out to be a problematic concept. The behavior of $\hat{\kappa}$ in some cases is described in the literature as paradoxical (e.g., Feinstein and Cicchetti 1990).

It is characteristic that κ is defined as a sample statistic only. This makes it difficult to assess its merits as an estimator for a population parameter, and this practice obscures model assumptions. In the rare expositions where κ is presented on the basis of a population model (Kraemer et al. 2002), the modeling is tied to the concepts of classification and cross-tabulation (where the diagonal cell counts N_{kk} estimate the agreement probabilities $P(Y_1 = Y_2 = k)$, and the marginal counts $N_{k\cdot}$ estimate the marginal distribution $P(Y = k)$).

In a number of papers, we have taken a different approach in interpreting the kappa statistic (De Mast 2007; Erdmann et al. in press). We refrained from taking the cross-tabulation analogy as a point of departure, because this conceptualization sees appraisals as classification, rather than as measurement.

Call	Appr-1	Appr-2
1	3	3
2	1	1
3	1	2
4	3	5
5	5	5
6	4	4
7	3	3
8	2	4
9	3	3
10	3	3
11	2	4
12	3	3
13	3	3
14	5	5
15	2	2

Appraiser 2 \ Appraiser 1	1	2	3	4	5	All
1	1 0.13	1 0.27	0 0.80	0 0.40	0 0.40	2.00 2.00
2	0 0.20	1 0.40	0 1.20	2 0.60	0 0.60	3.00 3.00
3	0 0.47	0 0.93	6 2.80	0 1.40	1 1.40	7.00 7.00
4	0 0.07	0 0.13	0 0.40	1 0.20	0 0.20	1.00 1.00
5	0 0.13	0 0.27	0 0.80	0 0.40	2 0.40	2.00 2.00
All	1 1.00	2 2.00	6 6.00	3 3.00	3 3.00	15.00 15.00

FIGURE 1 **Fictitious data set in raw and cross-tabulation format. (Color figure available online.)**

Call	Appr-1	Appr-2
1	1	1
2	1	1
3	1	1
4	1	1
5	2	2
6	1	1
7	1	1
8	1	1
9	1	1
10	1	1
11	1	1
12	1	1
13	3	1
14	1	1
15	1	1

Call	Appr-1	Appr-2
1	1	1
2	4	2
3	1	1
4	1	1
5	4	4
6	2	3
7	2	2
8	3	3
9	4	4
10	2	4
11	2	2
12	3	3
13	3	3
14	3	2
15	1	1

FIGURE 2 **Two fictitious data sets. In both cases, $\hat{\kappa} = 0.643$. (Color figure available online.)**

The crucial difference between classification and measurement is that measurement is a special form of classification, aimed to reflect an empirical property (the measurand) of the items being measured. Including this measurand in the modeling allows one to separate assumptions about the measurand (which is a characteristic of the population of calls) from the behavior of measurement errors (a characteristic of the classification procedure). This gives the following model.

We assume an unordered scale $\{0, 1, \ldots, a-1\}$ with a categories and denote items (calls, in the example) by the subscripts $i = 1, 2, \ldots, n$ and appraisers (or repeated appraisals) by $j = 1, 2, \ldots m$. The true state of an item (the intention of a call) is $X_i \in \{0, 1, \ldots, a-1\}$, with probability distribution $p(k) = P(X_i = k)$ (accounting for variability of the measurand). The appraisal result of item i by appraiser j is denoted Y_{ij} and conditional on the item's true state $\{X_i = k\}$, the $Y_{i1}, \ldots, Y_{im}$ are assumed independent and identically distributed (i.i.d.) with $q(l|k) = P(Y_{ij} = l|X_i = k)$ (accounting for measurement variability). The unconditional distribution is $q(l) = P\left(Y_{ij} = l\right) = \sum p(k)\, q(l|k)$ (marginal distribution). The probability of agreement is defined as

$$P_A = P(Y_{i1} = Y_{i2}) = \sum\nolimits_{k=0}^{a-1} \sum\nolimits_{l=0}^{a-1} p(k) q^2(l|k).$$

De Mast and Van Wieringen (2007) showed that the traditional sample statistic $\widehat{P}_A$ is an unbiased estimator of this probability. Trying to find a probability $P_{A|chance}$ that mirrors the traditional definition in sample statistics, we propose $P_{A|chance} = P(Z_{i1} = Z_{i2})$, with Z_{ij} the chance ratings done by an uninformative classification procedure. The traditional definition of κ amounts to the assumption that the distribution of chance ratings equals the marginal distribution of the classification procedure under study when applied to the items population under study; that is, $P\left(Z_{ij} = l\right) = q(l) = \sum p(k) q(l|k)$, and this gives $P_{A|chance} = \sum q^2(l)$ and

$$\kappa = \frac{P_A - P_{A|chance}}{1 - P_{A|chance}} \qquad [1]$$

Personally, we are not satisfied with this definition of a parameter for evaluating the validity of categorical measurements and, in particular, we find the concept of chance ratings too ambiguous to provide a well-defined zero point for the probability of agreement. In addition, we do not see why chance ratings would happen to have the same probability distribution as the marginal distribution of the classification procedure under study. But whatever one's view on this matter, our analysis reveals a number of strong ramifications of the traditional definitions that are underappreciated in the literature, and that make the interpretation of agreement studies precarious.

First, our analysis (De Mast 2007) allows a more effective interpretation of κ, which explains many of the paradoxes. Rewriting the terms in [1] we have

$$\kappa = 1 - \frac{1 - \sum (p(l) \sum q^2(k|l))}{1 - \sum q^2(k)} = 1 - \frac{\Delta^G_{Y|X}}{\Delta^G_Y}, \qquad [2]$$

where $\Delta^G_X = 1 - \sum p^2(k)$ is the Gini dispersion of a discrete variable X. The form on the right in Eq. [2], with Δ a measure of dispersion, is the generic form of measures of (predictive) association, and thus we have shown that κ can be interpreted as a measure of association between repeated ratings of an item. Replacing the Gini dispersion in [2] with $\Delta^E_X = -\sum p(k) \log p(k)$ (the entropy), we find Theil's uncertainty coefficient, which is thus a direct cousin of κ. And with $\Delta^V_X = \sigma^2_X$ (the variance of a continuous variable X), the right-hand side of Eq. [2] reduces to the intraclass correlation coefficient. Interpreting κ as a measure of intraclass association, much of its paradoxical behavior makes sense (Erdmann et al. in press).

Second, the normalization based on $P_{A|chance} = \sum q^2(l)$ depends on the population of items (because the $q(l)$ depend on the $p(k)$) and, consequently, a classification procedure's κ is meaningless in other item populations than the one from which the items in the study were randomly sampled. Likewise, the normalization is based on the marginal distribution of the rating procedure under study and therefore κ cannot be used to compare two different rating procedures. Namely, the chance correction for one rating procedure is different from the correction for the other and, therefore, the resulting κ values are on scales with different zero points. A review of actual agreement studies in top medical journals reveals that these two pitfalls are not generally recognized (Erdmann et al. in press).

Further, the κ statistic is sometimes used in industry to evaluate pass–fail inspections. In such cases, the true state of items is $X=0$ (defective) or $X=1$ (good), and the inspection result is $Y=0$ (fail) or $Y=1$ (pass). Because the defect rate $p(0)$ is typically extremely small, we have

$$P_A = p(0)\big(q^2(0|0)+q^2(1|0)\big)+p(1)\big(q^2(0|1)+q^2(1|1)\big)$$
$$\approx p(1)\big(q^2(0|1)+(1-q(0|1))^2\big),$$

where $q(0|1)$ is the probability of a false rejection. Consequently, P_A and κ evaluate pass–fail inspections almost exclusively in terms of the producer's risk (the probability of a false rejection) but ignore the consumer's risk (the probability of a false acceptance).

Finally, the lack of a population model in most expositions obscures an assumption that is crucial for the sample $\hat{\kappa}$ to be a meaningful estimator. Namely, conditional on the calls' true states X_i, the Y_{ij} are assumed independent, an assumption that amounts to the claim that besides $X\in\{0,\ 1,\ldots,\ a-1\}$ there are no other properties of the calls and environment that induce dependencies among the ratings Y. Implicitly or explicitly, most expositions make this assumption, but in many cases, such as the one at hand here, it is implausible. For example, In addition to the intent of a caller, the wording that he or she chose and the intonation will affect the probability distribution of Y. The ramifications of such violations of conditional independence assumptions have been studied thoroughly for two-point scales in De Mast et al. (2011), the main conclusion being that the bias in estimated parameters may be substantial if the sample of items is not representative.

Agreement studies and κ statistics have even more serious flaws when used to evaluate measurements on ordinal scales, such as judging the quality of soldered joints on a four-point scale {*A*(*reject*), *B*(*critical*), *C*(*acceptable*), *D*(*excellent*)}. In addition to other problems from which the κ statistic suffers, this practice has the problem that it ignores the order information in such ratings and treats an ordinal scale as a nominal scale. Suppose two appraisers judge the same five items as *A*, *C*, *A*, *C*, *B* and *B*, *D*, *B*, *D*, *C*, respectively. An agreement study would find zero agreement and conclude that the appraisers are even less consistent than chance ($\hat{\kappa}=-0.19$). An evaluation that incorporates order into the analysis, to the contrary, would find that the appraisers are in fact very consistent in ordering items relative to each other.

A popular method to improve agreement studies for ordinal ratings is the use of a weighted κ statistic. Instead of the proportion of agreement $\widehat{P}_A = \frac{1}{n}\sum_{k=0}^{a-1} N_{k,k}$, we have the degree of disagreement

$$\widehat{D} = \frac{1}{n}\sum\nolimits_{k1=0}^{a-1}\sum\nolimits_{k2=0}^{a-1} N_{k1,k2}w_{k1,k2},$$

with weights $w_{k1,\ k2}$ quantifying the severity of a disagreement between classes $k1$ and $k2$. For $w_{kk}=0$ and $w_{k1,k2}=1(k1\neq k2)$, we have $\hat{D}=1-\widehat{P}_A$. Weighted kappa is defined with similar modifications for the degree of disagreement expected for chance appraisals. The usual weighting scheme is quadratic: $w_{k1,\ k2}=(k1-k2)^2$. The scheme is motivated as follows (Fleiss and Cohen 1973): provided that the classes of the ordinal scale are in fact equidistant points on an interval scale, the weighted kappa based on quadratic weights approximates the intraclass correlation coefficient $Cor(Y_{i1},\ Y_{i2})$, which is a commonly used parameter for expressing the agreement of numerical measurements. Thus, the ordinal scale {*A*, *B*, *C*, *D*} is treated as an interval scale {1, 2, 3, 4}. But this makes one wonder: if the scale is in fact an interval scale, it should be called an interval scale (instead of ordinal), and the measurement procedure should be evaluated in terms of metrics suited for that type of scale. If the ordinal classes cannot be interpreted as equidistant points on an interval scale, then the chosen weights are arbitrary and make the analysis hard to interpret.

Note how the space needed to explain the κ statistic is substantially less than the space needed to explain at least some of its behavior and discuss at least some aspects of its interpretation. The discussion may give an impression of the extent to which this field struggles with misguided statistical modeling and conceptualizations of measurement. In particular,

- The statistical modeling is ineffective. The lack in most expositions of a population model obscures crucial model assumptions. The conceptualization is that of classification, rather than that of

measurement. And the model is not based on the independent drivers of the stochastic behavior, namely, $p(k)$ (the variability of the true state X in the population of items) and $q(l|k)$ (the variability of the random measurement error).

- The statistic in terms of which the measurement system is evaluated is not based on a notion of measurement error. Further, the statistics that are used for the evaluation are sometimes not appropriate for the type of scale and the structures that it can represent, especially for ordinal measurements, which are either treated as nominal or as numerical.
- For measurement on a categorical scale, the empirical properties that the categorizations intend to reflect usually have a much more complex structure than the scale's limited number of classes can capture, and this fact makes simplistic assumptions, such as the ubiquitous assumption about conditional independence, suspect.

We conclude that we statisticians need to develop a more fundamental understanding of measurement and the important concept of measurement error. This is the topic of the next section.

A CONCEPTUAL MODEL OF MEASUREMENT AND MEASUREMENT ERROR

Measurement is the assignment of symbols to items (or phenomena or substances or . . .) in such a way that mathematical relations among the symbols represent empirical relations among the items with respect to a property (the 'measurand') under study.

This definition comes from a branch of mathematics and philosophy of science called *measurement theory* (Hand 1996; Wallsten 1988), and similar definitions are generally accepted in psychometrics (Allen and Yen 1979; Lord and Novick 1968). We briefly discuss the theory on which it is based, because it helps us to understand in what way categorical appraisals can be seen as measurements, and it helps us to develop a useful notion of measurement error for categorical measurements.

Measurement is a mapping from a set of items to a measurement scale. The latter is a set of numbers or symbols, such as $\mathbb{R}^+$, $\{0, 1\}$, or $\{A, B, C, D\}$, equipped with algebraic structures and operators such as $\leq$(order) and $+$(addition). Measurement constitutes a homomorphic (that is, structure preserving) map. Let us say we are interested in a set of items $\mathscr{I}$ and one of their properties, namely, their mass. This empirical property induces various relational structures among the items, such as order. One can, for instance, compare two items A and $B \in \mathscr{I}$ directly using a balance (i.e., without measuring them) and establish that one of them (say, B) is heavier ($A \sqsubseteq B$). Even richer structures are created if we allow operations applied to the items. For example, let us denote by $A \circ B$ the operation of grouping two items together on one side of the balance, and $A \circ B \sqsubseteq C$ denotes the empirical fact that a balance with A and B on one side and C on the other tips down on the latter side. Thus, we have an empirical system $[\mathcal{I}, \sqsubseteq, \circ]$, consisting of a set of items and structures induced by empirical relations and operations.

By measuring the mass of the items, using a spring scale, for example, we map these empirical relations to a set of numbers equipped with mathematical structures. The measurement assigns to each measured item $A \in \mathscr{I}$ a value $M(A) \in \mathbb{R}^+$ and, thus, measuring is a map M: $\mathscr{I} \rightarrow \mathbb{R}^+$ (or another scale). The set of numbers $\mathbb{R}^+$ is equipped with mathematical relations, such as the order relation $\leq$, and the idea of measurement is that the mathematical order between measured mass values mirrors the empirical order given by the comparison of items using a balance (that is, if $A \sqsubseteq B$ then $M(A) \leq M(B)$). Moreover, $\mathbb{R}^+$ is equipped with algebraic operators such as $+$(addition), and the idea of measurement is that they have empirical counterparts such as the grouping of items together on one side of the balance ($A \circ B$) and that relations carry over ($M(A \circ B) = M(A) + M(B)$). Measurement, in short, takes empirical relations among items and operations applied to them and maps the resulting structure onto a set of numbers equipped with algebraic structures, thus establishing a homomorphism between $[\mathcal{I}, \sqsubseteq, \circ]$ and $[\mathbb{R}^+, \leq, +]$. This homomorphism ensures that mathematical statements such as $M(A) + M(B) \leq M(C)$ have empirical meaning.

Some measurement systems preserve more structure than others. An influential typology of measurement as to how much structure it preserves

is Stevens' (1946) discerning nominal measurement (preserves equivalence relations ($=$) only), ordinal measurement (preserves order relations ($\leq$)), interval measurement (preserves, in addition, distances between items and thus allows addition and subtraction), and ratio measurement (preserves, in addition, an empirical zero point and thus allows multiplication and division). Note that the popular distinction between continuous and discrete is useful for random variables but not for measurement data, because these are always discrete.

$\mathbb{R}^+$ is equipped with order ($\leq$), a norm or distance metric (e.g., 3 and 1 are twice as far apart as 2 and 1), and a zero point. When measuring mass in kilograms or pounds, each of these has an empirical counterpart (for example, $M(A) = 0$ corresponds to the total absence of mass). But when one measures temperature in degrees Celsius or Fahrenheit, one uses the same numerals in $\mathbb{R}^+$, but some mathematical relations and algebraic operators stop having empirical counterparts. In particular, the zero point of neither of these temperature scale has empirical meaning in the sense that it corresponds to the total absence of an empirical quantity. As a consequence, a statement such as: "30°C is two times 15°C" is true about the numbers themselves, but it is difficult to see what empirical meaning such statement has. Note that the Kelvin scale does have a zero point that corresponds to the total absence of the empirical quantity in question and, consequently, "30K is twice as warm as 15K" does have empirical meaning.

On an ordinal scale such as $\{A, B, C, D\}$, only the $\leq$ and $=$ relations have empirical meaning, but there is no distance metric, and operators such as addition and subtraction do not in general have empirical meaning. Recoding the scale using numerals (that is, $1 = A$, $2 = B$, $3 = C$, $4 = D$), one can apply mathematical operators to these numerals, but a statement such as: "the difference between C and A is twice as large as the difference between B and A" (since $3-1$ is two times $2-1$) is in general empirically meaningless.

The conception of measurement as a homomorphism shows that, in addition to the measurement of quantitative properties such as temperature and weight, ratings on a nominal scale, quality inspections on a binary scale, and diagnoses by radiologists or physicians on a nominal or ordinal scale can be conceived of as measurement. Thus, the given definition seems more general than typical definitions of measurement in metrology (e.g., Joint Committee for Guides in Metrology [JCGM] 2008; Kimothi 2002) that seem especially geared to quantitative measurements (JCGM 2008, p. 16, explicitly states that the word *measurement* does not apply to nominal properties).

In addition to the notion of measurement as a homomorphism, we need the concepts of measurand, true value, and measurement error. The measurand is the empirical property that the measurements aim to reflect. It can be a continuous property, such as the length of cables, but also a dichotomous or polytomous property or a more complex combination of properties, such as this one:

> Visual inspection (M: $\mathscr{I} \rightarrow \{Accept, Reject\}$) of products for scratches:
>
> - The inspections should yield *Reject* if a product has one or more scratches with a depth of at least 40 μm and a width of at least 100 μm, and *Accept* otherwise.

The definition of the measurand should not be confused with the operational definition of the measurement procedure. Often, a measurand is not assessed directly but via a known relationship with other, more readily observable properties. Examples include the determination of an item's mass by determining the compression of a spring in a scale, a pregnancy test that produces a positive result if the levels of certain chemical markers are beyond a threshold value, or a spam filter that classifies messages on the basis of the usage of certain indicative words resulting in a spam score. In these cases, the measurands are the mass of items, whether a woman is or is not pregnant, and the intention with which a message was sent. The compression of the spring, levels of chemical markers, or spam score are merely part of the operating procedure or algorithm.

The true value $T(A)$ of a measured item A is the value that should be assigned according to the measurand's definition (JCGM 2008), and it is a homomorphic map T: $\mathscr{I} \rightarrow \mathscr{S}$ (with $\mathscr{S}$ the measurement scale). It is generally acknowledged that the true value is more a construct than a concept that can be given operational meaning, and the JCGM (2008) acknowledges that it is not in general a unique value. For example, if we define the measurand to be "the

length of cables," we ignore the fact that cables do not have a unique length, because this depends on temperature and tension. A better definition of the measurand would be "the length of cables at 20°C and under a tension of 10N," but even this does not define a unique value, because length depends on more variables, and their infinite number cannot be captured in a finite definition.

Because the true value is usually unknowable on principle, a more pragmatic concept is the reference value $R(A)$, a value obtained from an authoritative measurement system that is accepted by convention to play the role of the true value (c.f. JCGM, 2008). In medicine, a similar concept is that of a gold standard (or criterion standard), an authoritative test for evaluating diagnostic and screening tests. Reference values and gold standards should be of a higher order of accuracy than the measurement system under study, but it is generally acknowledged in metrology and medicine that they are usually not perfect.

Actual measurements are subject to random measurement error and therefore are a stochastic map M: $\mathscr{I} \times \Omega \to \mathscr{S}$ (with Ω a probability space). Measurement error is the discrepancy between a measurement result $M(A)$ and the true value $T(A)$ or reference value $R(A)$. The numerical expression of measurement error depends on a scale's algebraic structures and operators. Binary and nominal scales are only equipped with the simplest of structures—the equivalence relation—and measurement error therefore takes the form of misclassification: $\{M(A) \neq R(A)\}$. A statistical evaluation will typically be in terms of a probability of misclassification. For ordinal scales, measurement error can refer to misclassification but, in addition, to whether pairs of items are ordered correctly ($\{M(A) \leq M(B),\ R(A) \geq R(B)\}$). For measurements on an interval scale, measurement error can be defined as the difference $M(A) - R(A)$ between a measurement result and the reference value, and a statistical evaluation can be in terms of the mean and standard deviation of this difference. For ratio scales, finally, one could even consider the relative measurement error $(M(A) - R(A))/R(A)$.

Based on this conceptual framework, in recent years we have presented methodologies for evaluating categorical measurements by statistical modeling of the measurement error. The next sections present this type of modeling, especially for binary measurements.

STATISTICAL MODELING OF BINARY MEASUREMENT

Examples of binary measurement include leak tests, visual inspections, inspections based on go–no go gauges, and diagnostic and screening tests in medicine. Until recently, the statistical evaluation of such tests has been impeded greatly by overly simplistic statistical modeling. In addition to the measurement result $Y=0$ (Reject) or $Y=1$ (Accept), traditional models featured a true state $X=0$ (Defective) or $X=1$ (Good). The measurements were evaluated in terms of misclassification probabilities, such as the false acceptance probability $FAP = P(Y=1|X=0)$ and the false rejection probability $FRP = P(Y=0|X=1)$ (depending on the situation, also called the false positive and false negative rate, or their complements, specificity and sensitivity). Implicitly or explicitly, statistical methods for estimating these misclassification probabilities from experiments hinge on a very important assumption, that of conditional independence. This assumption amounts to the claims that for parts $i=1,\ 2,\ldots$ and repeated appraisals $j=1,\ 2,\ldots$, the Y_{ij} are i.i.d. conditional on the events $\{X_i=0\}$ or $\{X_i=1\}$. And this in turn amounts to the assumption that the measurand dichotomizes the parts into two subpopulations that are homogeneous with respect to Y_{ij}:

$FAP = P(Y_{ij} = 1|X = 0)$ is identical for all defective items, and

$FRP = P(Y_{ij} = 0|X = 1)$ is identical for all good items.

This modeling is widespread in medicine and engineering alike (Boyles 2001; Danila et al. 2008; Van Wieringen and De Mast 2008, to mention just a few examples).

Recent research in engineering and medicine (e.g., Danila et al. 2012; De Mast et al. 2011; Irwig et al. 2002) has revealed that such modeling approaches are misguided in most situations. Although binary scales have a simple structure, the empirical phenomena with which they are homomorphic can

in fact have a very complex structure. The measurand is typically related to continuous properties, and misclassification probabilities are in addition affected by properties of the parts not related to the measurand that create intricate dependency structures among the Y. De Mast et al. (2011) demonstrated that statistical evaluations based on unwarranted conditional independence assumptions may lead to substantially biased estimates of misclassification probabilities.

In many cases, the measurand underlying the binary classification is related to a continuous property. Erdmann et al. (2013) discussed the inspection of car parts for the misalignment of a clip to a pad to which it should be attached. Here, the true value $X \in \{0, 1\}$ is determined by the misalignment $Z \in \mathbb{R}$, namely, $X = 1$ if $Z \leq USL$ and $X = 0$ otherwise, with USL the upper specification limit that demarcates the acceptable extent of misalignment. The conditional independence assumption implies that the rejection probability $P(Y_{ij} = 0)$ is a step function with values FRP and $1 - FAP$ (Figure 3). But it is much more plausible that the rejection probability, as a function of misalignment, is a continuous S-curve $q(z) = P(Y = 0|Z = z)$, where the rejection probability gradually changes from 0.0 to 1.0. Put differently, the rejection probability does not only depend on $X \in \{0, 1\}$ (*whether* the part is good or defective) but also on the *degree* of defectiveness or goodness.

The type of modeling that we have explored in recent years is based on the concept of the characteristic curve $q(z) = P(Y = 0|Z = z)$, where q could be the logistic function:

$$\log\left(\frac{q(z)}{1 - q(z)}\right) = \alpha(z - \delta) \qquad [3]$$

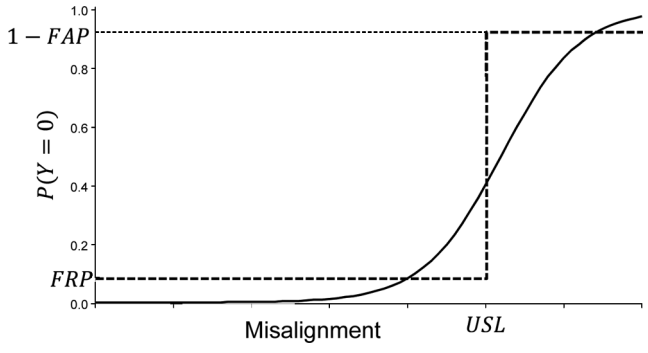

FIGURE 3 $P(Y = 0)$ as a function of misalignment (a) under conditional independence (step function) and (b) as a continuous characteristic curve.

(see Figure 4). The parameter δ is the point where $q(\delta) = 0.5$ and can be interpreted as a decision threshold: parts with a misalignment $z > \delta$ are likely to be rejected. The difference $\delta - USL$ could be interpreted as systematic measurement error. The parameter α determines the steepness of the curve (larger α corresponding to a steeper curve). The distribution of misalignment in the population of parts is denoted $F_Z(z) = P(Z \leq z)$, with associated density $f_Z(z)$.

For any specific part the misclassification probability now depends on the part's misalignment Z, but the *average* probabilities (weighted by f_Z) are

$$FAP = P(Y = 1|X = 0) = \frac{\int_{USL}^{\infty}(1 - q(z))f_Z(z)dz}{\int_{USL}^{\infty} f_Z(z)dz}, \qquad [4]$$

and

$$FRP = P(Y = 0|X = 1) = \frac{\int_{-\infty}^{USL} q(z)f_Z(z)dz}{\int_{-\infty}^{USL} f_Z(z)dz}. \qquad [5]$$

The traditional way of estimating FAP and FRP is as follows (see, for instance, the cross-tab method in AIAG 2003). One obtains a sample of n_1 good parts and has them each appraised once; the resulting number of rejected parts is $m_{0|1}$. Likewise, one takes a sample of n_0 defective parts and has them appraised, resulting in $m_{1|0}$ accepted parts. The FAP and FRP are estimated by the sample proportions $m_{0|1}/n_1$ and $m_{1|0}/n_0$. Equations [4] and [5] show, however, that it is crucial for this to work that the distribution of misalignment in the samples is representative for the population distribution F_Z. If, for example, parts with a misalignment Z near δ

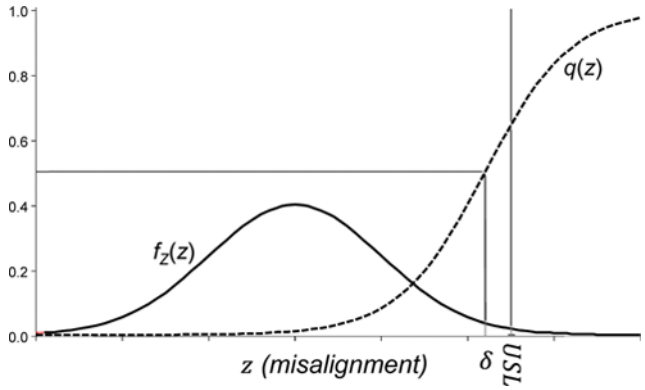

FIGURE 4 Density function f_z of misalignment and characteristic curve q. (Color figure available online.)

are overrepresented in the samples, *FAP* and *FRP* are overestimated (De Mast et al. 2011). Random samples from the populations of defective and good parts ensure that the sample distribution of misalignment is representative for the population distribution, but in practice it is not at all clear how such random samples can be obtained. In practice, one has the streams of rejected and accepted parts, but taking random samples from these and removing the incorrectly rejected and accepted parts does not result in representative samples and may lead to substantial bias (De Mast et al. 2011). Note, for example, that for low defect rates $P(X_i = 0)$ and realistic *FAP* and *FRP*, the stream of rejected parts consists mostly of good parts that have been rejected falsely (De Mast et al. 2011).

Alternatively, one could take a representative sample from the total population of parts, establish for each part the reference value (which gives n_0 defective and n_1 good parts), and next have them appraised by the inspection system under study (yielding $m_{0|1}$ and $m_{1|0}$). However, in a typical manufacturing process, the defect rate is very small and, consequently, the number n_0 of defectives in the sample will be zero or very small. This makes estimation of the *FAP* in particular precarious.

We advocate an approach in which the *FAP* and *FRP* are determined indirectly, by fitting the characteristic curve $q(z)$ and population distribution $F_Z(z)$ and substituting these in [4] and [5]. If reference (that is, X-) values for parts can be obtained, the characteristic curve can be fitted by logistic regression, and this is essentially what is done in AIAG's so-called analytic method (AIAG 2003). A representative sample of parts will do to fit $F_Z(z)$.

It is quite common, however, that reference values cannot be obtained, for lack of a higher order measurement system. Erdmann et al. (2013) demonstrated how to deal with such situations. By having a sample of parts inspected more than once by each of the appraisers, one can use techniques from latent variable modeling to fit $F_Z(z)$ and $q(z)$ simultaneously. Such approaches require sophisticated sampling strategies. The challenge is that defects are remote tail phenomena: given that δ and *USL* will typically be in the remote tail of f_Z, a representative sample from the total population of parts will contain zero or very few parts with misalignment values around and above δ, resulting in large standard errors for the estimated *FAP*. Taking a nonrepresentative sample with more parts with misalignment around δ, to the contrary, would result in a substantial bias in the estimation of the parameters of F_Z. Erdmann et al. (2013) proposed a method based on a combination of samples from various streams (total parts population, the stream of rejected parts, and a historical reject rate). The parameters F_Z and q are estimated by a maximum likelihood procedure, in which the bias induced by the nonrepresentativeness of some of the samples is corrected by calculating the likelihood contributions conditional on the source of the samples.

In these situations where reference values cannot be obtained, *USL* is typically ill defined, making the definitions of the *FAP* and *FRP* problematic. Note, however, that the *FAP* and *FRP* decompose into (Erdmann et al. 2013) the systematic error $\delta - USL$ and the random errors

$$IAP = P(Y = 1|Z > \delta)(\text{inconsistent acceptance probability}), \text{and}$$
$$IRP = P(Y = 0|Z \leq \delta)(\text{inconsistent rejection probability}).$$

The latter components are the probability that an appraiser's classification is inconsistent with his or her own decision threshold δ. They can be interpreted as the random measurement error, and they can be estimated instead of the *FAP* and *FRP* if the *USL* is not well defined.

This type of modeling sometimes requires new statistical techniques. Danila et al. (2010) have explored the incorporation of historical data in model fitting. In Erdmann et al. (2013) we described the technique of fitting the characteristic curve and measurand distribution simultaneously based on combinations of random samples, and we proposed techniques for model diagnostics. Erdmann and De Mast (2012) explored alternative and especially asymmetrical functions for the characteristic curve. De Mast and Van Wieringen (2010) studied the application of similar latent variable models for the evaluation of ordinal measurements, which required the development of a new optimization algorithm to determine maximum likelihood estimates and powerful model diagnostics. Assuming an ordinal scale $\{1, 2, \ldots, a\}$, and a continuous property Z

underlying the ordinal measurements, the model in question fits characteristic curves of the form

$$P(Y=k|Z=z)=q(k|z)=\frac{\exp\left(\sum_{m=1}^{k-1}\alpha(z-\delta_m)\right)}{\sum_{n=1}^{a}\exp\left(\sum_{m=1}^{n-1}\alpha(z-\delta_m)\right)},$$

where the $\delta_1,\ldots,\delta_{a-1}$ are the thresholds in between the scale's a categories, and the α parameter determines the steepness of the curves (similar to the δ and α in [3]). The model allows the calculation of probabilities of inconsistent classification $\sum_{k=1}^{a}P(Y_i\neq k|\delta_{k-1}<Z_i<\delta_k)P(\delta_{k-1}<Z_i<\delta_k)$ (comparable to the *IAP* and *IRP*) and probabilities of incorrect order $P(Y_{i1}>Y_{i2}|Z_{i1}\leq Z_{i2})$.

A CASE OF REALISTIC COMPLEXITY

Approaches such as agreement studies, based on sample statistics without a solid statistical modeling and not grounded in a notion of measurement error, are too simplistic, and the results are often meaningless. Estimation of *FAP* and *FRP* based on the conditional independence assumption is also precarious, because it may be difficult to obtain estimates with no or acceptably small bias. The approach based on univariate characteristic curves is, we think, applicable in some real cases. We have found, however, that most real cases are substantially more complex even than the car parts example in the previous section.

In the previous year, our group has been involved in assessment studies of a number of real binary measurement systems. We cannot present here the cases that we have been involved in but, instead, we constructed a fictitious but realistic example, concerning inspection of products for scratches. On the basis of this example we want to bring across what we think is a realistic level of complexity of such situations and offer an approach for coping with it.

The measurand in the example is a complex property. The true value is defined as follows (see also Figure 5):

- If a product has scratches with a width Z_1 of at least 100 µm and a depth Z_2 of at least 40 µm, it is nonconforming ($X=0$).
- If a product has no scratches wider than 60 µm and deeper than 30 µm, it is acceptable ($X=1$).

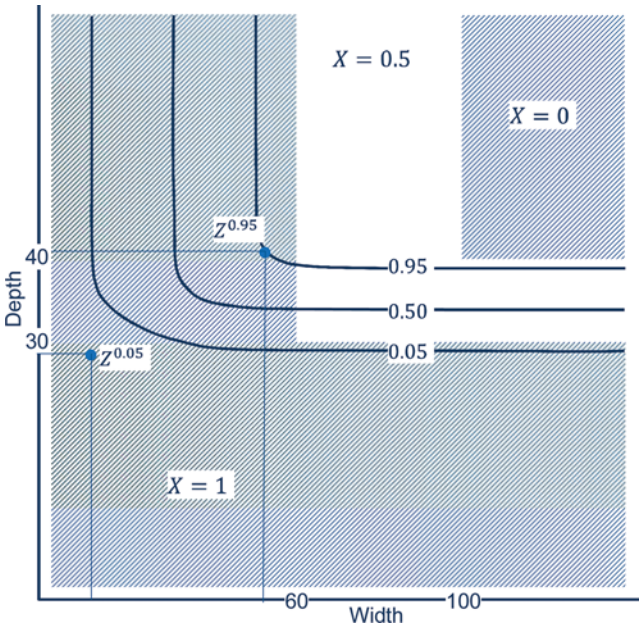

FIGURE 5 **Contour plot of fitted characteristic curve in scratch width (Z_1) and depth (Z_2) (fictitious results); contours for $q(z_1, z_2)=0.05$, 0.50, and 0.95 indicated by solid curves. (Color figure available online.)**

- Otherwise, it is marginal ($X=0.5$), meaning that either inspection result is acceptable.

We will assume (not very unrealistically) that there is at most one scratch on a product. The measurand is defined in terms of Z_1 and Z_2 but, in this example, the probability of rejection depends on more covariates:

Z_3 is the color of the part. The product is sold in 16 standard colors, and scratches are more easily spotted on light than on dark surfaces.
Z_4 is appraiser fatigue at the time when a part is measured.
Z_5 quantifies light conditions at the time when a part is measured.

The misclassification probabilities are affected by all of these covariates. Let

$$q(z_1,\ldots,z_5)=P(Y=0\,|\,\text{scratch};Z_1=z_1,\ldots,Z_5=z_5),\quad \text{for } z_1>0 \text{ and } z_2>0,$$

and

$$r(z_3,z_4,z_5)=P(Y=0\,|\,\text{noscratch};Z_3=z_3,\ldots,Z_5=z_5),$$

with $\lim_{z_1 \downarrow 0} q(z_1, \ldots, z_5) = \lim_{z_2 \downarrow 0} q(z_1, \ldots, z_5) = r(z_3, z_4, z_5)$. We have

$$FAP = \frac{\int_{z_1=40}^{\infty} \int_{z_2=100}^{\infty} \int_{z_3,z_4,z_5} (1 - q(z_1, \ldots, z_5)) f_{Z_1}(z_1) \cdots f_{Z_5}(z_5) dz_5 \cdots dz_1}{\int_{z_1=40}^{\infty} \int_{z_2=100}^{\infty} f_{Z_1}(z_1) f_{Z_2}(z_2) dz_2 dz_1}, \quad [6]$$

$$FRP = \frac{P(\text{no scratch}) \int_{z_3,z_4,z_5} r(z_3, \ldots, z_5) f_{z_3}(z_3) \cdots \quad f_{z_5}(z_5) dz_5 \cdots dz_3}{P(\text{no scratch}) + P(\text{scratch}) \int_{(z_1,z_2),\in L}^{\infty} f_{z_1}(z_1) f_{z_2}(z_2) dz_2 dz_1} + \frac{P(\text{scratch}) \int_{(z_1,z_2)\in L}^{\infty} \int_{z_3,z_4,z_5} q(z_1, \ldots z_5) f_{z_1}(z_1) \cdots f_{z_5}(z_5) dz_5 \cdots dz_1}{P(\text{no scratch}) + P(\text{scratch}) \int_{(z_1,z_2),\in L}^{\infty} f_{z_1}(z_1) f_{z_2}(z_2) dz_2 dz_1}, \quad [7]$$

where L is the L-shaped area $\{(z_1, z_2)$: $z_1 \leq 60$ or $z_2 \leq 30; z_1 > 0; z_2 > 0\}$. The reader will realize that the complexity of such situations is a far cry from the simplistic depiction of measurement results being independent conditionally on a dichotomous measurand. Note, however, that also this approach ultimately makes a conditional independence assumption: the Y are assumed independent conditional on X and the $Z_1, \ldots, Z_5$. The point is not that our modeling approach avoids conditional independence assumptions altogether but, rather, that more of the underlying complexity is accounted for in the statistical model.

Practically, fitting q, r, P(scratch), and the densities f_Z is unfeasible. Instead, we propose three simplification strategies. Although they reduce the number of arguments in q and r, they do require that the covariates have been identified beforehand.

1. Sidelining covariates: It may be possible to make some covariates irrelevant by ensuring in the measurement protocol that they are always constant during the inspections. This is in fact a strategy to improve the measurement procedure itself, but as a by-product it also simplifies an evaluation of its performance.
2. Averaging out by experimental randomization: Suppose that we are not interested in modeling the effect of, say, Z_5. The strategy is to ensure that during the experiment from which the characteristic curve is fitted, the values of Z_5 are representative for F_{Z_5} and that they are assigned randomly to the runs in the experiment. This ensures that we fit, in effect:

$$q_{\overline{z_5}}(z_1, \ldots, z_4) = \int_{-\infty}^{\infty} q(z_1, \ldots, z_4, z) f_{Z_5}(z) dz.$$

The approach typically requires thoughtful planning, and just sampling haphazardly is unlikely to achieve representativeness, with biased estimates as a consequence. Note that the extreme version of this strategy, averaging out over all $Z_1, \ldots, Z_5$, boils down to the traditional approach of estimating the average *FAP* and *FRP* from sample proportions. This extreme form lets go of the idea of determining *FAP* and *FRP* indirectly by first fitting the $f_Z(z)$ and $q(z)$ functions and computing *FAP* and *FRP* from there and, as stated earlier, only works under the doubtful assumption that one can obtain samples that are representative with respect to $f_{Z_1}, \ldots, f_{Z_5}$.
3. Worst-case evaluation: Instead of averaging out the effect of Z_5, we can fit characteristic curves with Z_5 fixed to the values $\breve{z}_5^{\mathrm{A0}}$ and $\breve{z}_5^{\mathrm{R0}}$ that maximize the *FAP* and *FRP*:

$$\breve{z}_5^{\mathrm{A0}} = \arg\max_z \int_{z_1,\ldots,z_4} (1 - q(z_1, \ldots, z_4, z)) f_{Z_1}(z_1) \cdots f_{Z_4}(z_4) dz_4 \cdots dz_1$$

(and analogously for $\breve{z}_5^{\mathrm{R0}}$). Instead of $q(z_1, \ldots, z_5)$ we fit $q_{\breve{z}_5^{\mathrm{A0}}}(z_1, \ldots, z_4) = q\left(z_1, \ldots, z_4, \breve{z}_5^{\mathrm{A0}}\right)$ and

$q_{\breve{z}_5^{R0}}(z_1, \ldots, z_4) = q\left(z_1, \ldots, z_4, \breve{z}_5^{R0}\right)$, and these curves represent worst-case bounds. In addition, this strategy requires that one is aware of a covariate and that its worst value can be guessed with reasonable plausibility.

We demonstrate these techniques for the inspections for scratches. We assume that it is possible to create scratches with specified width and depth (with sufficient precision). First, Z_5 (light conditions) could be made irrelevant by screening off the inspections from ambient light. Second, we deal with the part's color by limiting the evaluation to its most challenging values (worst-case evaluation). In this case, the values $\breve{z}_3^{A0}$ and $\breve{z}_3^{R0}$ that maximize the *FAP* and *FRP* are the same (namely, black). We will average out the effect of appraiser fatigue Z_4 by experimental randomization. Thus, we need an experimental design in the two remaining covariates scratch width Z_1 and depth Z_2, suitable for fitting the link function q. We could take the 7×7 grid $\{0, 15, \ldots, 120\} \times \{0, 20, \ldots, 90\}$. The experiment now involves creating 49 parts with scratches with width and depth defined by the design points and color $\breve{z}_3^{A0}$ ($=\breve{z}_3^{R0}$, black). These parts are judged repeatedly and in randomized order by appraisers under situations that are representative for the variation in fatigue under normal conditions. This should be carefully planned, one option being that appraisers take part in the experiment on, say, four moments distributed evenly over a shift and do their normal work in between. The characteristic curve $q_{\breve{z}_3^{A0}, \overline{z}_4}(z_1, z_2)$ is fitted on the results.

As an example, the contours in Figure 5 represent the fitted characteristic curve (for fictitious data). Note two properties crucial in understanding the performance of the measurement system: random and systematic errors. The random error can be represented by the distance between the contours showing 0.05 and 0.95 rejection probabilities. Ideally, this distance is small enough to fit in the marginal region $X = 0.5$. The systematic error is the *location* of the region between the 0.05 and 0.95 contours; ideally, this region falls as much as possible in the marginal region.

Figure 5 suggests that the scratch inspections perform well with regards to scratch depth, because the location and width of the horizontal area in between the 0.05 and 0.95 contours match well with the specifications of the inspection's true values. As for the scratch width aspect, we note that the inspections seem to be systematically off, although the random error seems relatively modest. It should be kept in mind that these contours represent the worst-case situation that the part color is black and that the characteristic curves for other colors are expected to be closer to the ideal behavior implied by the measurand definition.

Estimation of the *FAP* and *FRP* from the estimated characteristic curves requires the determination of P(scratch) and F_{Z_1} and F_{Z_2}, which requires a representative sample of scratched parts. This returns us to the earlier described challenge that defects are a remote-tail phenomenon. Sampling from the stream of rejected parts is likely to give a substantial overrepresentation of wide and deep scratches. Sampling from the total parts population and then using gold-standard measurements to identify scratched parts implies an enormous effort in the plausible scenario that P(scratch) is very small. From the cases that we have been involved in, our impression is that it is quite common that reliable estimation of *FAP* and *FRP* is problematic. Typically, there are one or two covariates for which it is very difficult to obtain representative samples, and this precludes the determination of *FAP* and *FRP* either directly (traditional estimation from sample proportions) or indirectly (by determining the characteristic curve and population distribution separately and obtaining the *FAP* and *FRP* from equations such as [6] and [7]).

To the contrary, fitting of the characteristic curves in the most important covariates seems generally feasible, and we have come to prefer the following approach. Instead of reporting *FAP* and *FRP*, we report the points $z^{0.05}$, δ, and $z^{0.95}$. For a characteristic curve in one argument, these are the z values for which $q(z) = 0.05$, 0.50, and 0.95, which represent the decision threshold and limiting values demarcating the grey area where inspection results are highly random. For the scratch inspections, we have a characteristic curve with two arguments. One could now choose a point $\mathbf{z}^{0.95} = \left(z_1^{0.95}, z_2^{0.95}\right)$ such that parts with scratches with $\boldsymbol{Z}_1 \geq z_1^{0.95}$ and $\boldsymbol{Z}_2 \geq z_2^{0.95}$ have a rejection probability of at least 0.95. Note that this is generally not a unique point. In Figure 5 we chose $\mathbf{z}^{0.95} = (53, 42)$, and we could now characterize the measurements as follows: "Parts with scratches with width >53 μm and depth >42 μm are rejected with a least 0.95 probability."

Similarly, one can choose a point $z^{0.05}$ such that scratches with $Z_1 \leq z_1^{0.05}$ or $Z_2 \leq z_2^{0.05}$ are rejected with 0.05 probability or less. In Figure 5, this is the point $z^{0.05} = (12, 29)$, with its coordinates determined by the horizontal and vertical asymptotes of the 0.05 contour. Parts with scratches with width $<12\,\mu m$ or depth $<29\,\mu m$ are accepted with a probability of at least 0.05.

In addition to the practical problems in obtaining reliable estimates, there are other practical drawbacks of reporting the validity of measurement systems in terms of an *FAP* and *FRP* and thus motivate an evaluation in terms of $z^{0.05}$, δ, and $z^{0.95}$ instead. Changes in the production process (and therefore in the distributions of the covariates) immediately render the estimated *FAP* and *FRP* invalid. Further, it is often desirable to quantify the validity of a measurement system in values that are not tied to a specific population of parts and manufacturing context; for example, for a manufacturer of measuring equipment wishing to specify the performance of its equipment. For such reasons, we have come to favor approaches where we fit the characteristic curve in the most important covariates (and report $z^{0.05}$ and $z^{0.95}$ points) but are not overly concerned with estimating *FAP* and *FRP*.

STEPWISE APPROACH

The experiences and (sometimes complex) arguments introduced and explained in the previous sections motivate a certain approach to the statistical evaluation of categorical measurement systems. We suggest that the modeling of the behavior of such systems naturally follows seven steps, which we discuss below.

Step 1: Measurement Model and Identification of Covariates

Measurements are on a categorical scale $\mathscr{C} = \{0, 1, \ldots, a-1\}$ (binary, nominal, ordinal, or other). The measurement system is evaluated for application to a population of items $\mathscr{I}$. Each item in $\mathscr{I}$ has properties $Z_1, \ldots, Z_k$ (usually not measured directly and sometimes even not more than a construct). In addition, the appraisals are done under circumstances that have properties $Z_{k+1}, \ldots, Z_l$.

In this first step, the inquirer critically reviews whether the measurement protocol is clearly specified and identifies the covariates $Z_1, \ldots, Z_k$ and $Z_{k+1}, \ldots, Z_l$. In addition, the inquirer carefully considers the definition of the measurand in terms of some of the $Z_1, \ldots, Z_k$, which defines the true values X as a mapping X: $\mathbb{R}^k \to \mathscr{C}$. The true value is usually unobservable in principle, and the inquirer should consider whether reference or gold-standard measurements R: $\mathscr{I} \to \mathscr{C}$ are available for some items. For identifying the covariates, the inquirer may consider these categories:

1. Covariates related to the measurand.
2. Covariates related to test conditions and appraisers performing the measurements.
3. Covariates related to properties of the items but not related to the measurand.

Step 2: Statistical Model

The actual measurements are a stochastic map Y: $\mathscr{I} \times \Omega \to \mathscr{C}$. The second step concerns the statistical model for this stochastic behavior in terms of which the measurement system will be evaluated. In general, the Y cannot be assumed i.i.d. conditional on the events $\{X=0\}, \ldots, \{X=a-1\}$. This is accounted for by the two main elements of the statistical model: q and F_Z. The characteristic curves $q(z) = P(Y=0|Z=z)$ attribute variability in $P(Y=0)$ to the covariates Z. The distributions F_Z of the covariates model variability of properties of the items and appraisal conditions in $\mathscr{I}$.

Step 3: Reducing Dimensionality of the Problem

In case there are more than, say, two covariates, we suggest reducing the dimensionality of the problem by applying the simplification strategies mentioned before:

- Sidelining of covariates
- Averaging out by experimental randomization
- Worst-case evaluation

Step 4: Functional Form

In the fourth step, the inquirer chooses a functional form for the characteristic curve. It is

convenient to find one that can be parameterized in such a way that random and systematic measurement error can be discerned (this is allowed by the parameters α and δ in the car parts example). A straightforward option is logistic curves:

$$q(z) = P[Y = 0|Z = z] = (1 + \exp\{-\alpha(z - \delta)\})^{-1}. \quad [8]$$

Erdmann and De Mast (2012) explored options for nonsymmetrical characteristic curves. An option for a two-dimensional characteristic curve, such as in the scratches example, is the bivariate logistic function q_b, proposed by Ali et al. (1978):

$$\begin{aligned} q_b(z_1, z_2) &= P[Y = 0|Z_1 = z_1, Z_2 = z_2] \\ &= (1 + \exp\{-\alpha_1(z_1 - \delta_1)\} \\ &\quad + \exp\{-\alpha_2(z_2 - \delta_2)\} \\ &\quad + (1 - v)\exp\{-\alpha_1(z_1 - \delta_1) \\ &\quad - \alpha_2(z_2 - \delta_2)\})^{-1}. \end{aligned}$$

The δs represent the vertical and horizontal position of the asymptotes of the 0.50 contours in Figure 5, and the αs determine the distance between the 0.05 and 0.95 contour lines. The parameter v can be varied for $-1 \le v \le 1$ and is a parameter of association. When $v = 0$, the characteristic curve equals $q_b(z_1, z_2) = q_1(z_1) \times q_2(z_2)$, where q_1 and q_2 are univariate logistic functions of the form [8]. This restricted form is suited when the rejection of items can be seen as the result of stochastically independent evaluations on the underlying characteristics Z_1 and Z_2.

The restriction may be tested by means of a likelihood ratio test or may be argued on a priori grounds. Alternatively, the bivariate normal distribution function may be used:

$$\begin{aligned} q_b(z_1, z_2) &= \Phi_{\mu,\Sigma}(z_1, z_2), \text{with } \boldsymbol{\mu} = (\delta_1, \delta_2) \text{and} \\ &= \begin{pmatrix} \alpha_1^{-2} & v \\ v & \alpha_2^{-2} \end{pmatrix}. \end{aligned}$$

Again, the value $\nu = 0$ corresponds to the restricted form suited when the rejection of items can be seen as the result of stochastically independent evaluations on the underlying characteristics Z_1 and Z_2.

Step 5: Evaluation Metrics

In step 5, the inquirer determines what metrics he or she wishes to determine from the fitted model in order to evaluate the measurement system. We strongly advocate that such metrics be based on the notion of measurement error; that is, somehow quantifying the statistical distribution of the discrepancy between the true value and the measurement values. For nominal and binary measurements, measurement error should be defined in terms of misclassification, and this motivates an evaluation in terms of the *FAP* and *FRP* or the random error components *IAP* and *IRP*. For ordinal scales, one could, in addition, estimate a probability of incorrect order (De Mast and Van Wieringen 2010).

As discussed, for pass–fail inspection, one often comes across the combination of challenges that obtaining representative samples from the populations of good and defective items is not possible, and working with a representative sample from the total population of items is hampered by the enormous sample sizes needed to obtain a reasonable number of parts with z-values in the steep and right part of the characteristic curve. In such cases, reporting $z^{0.05}$, δ, and $z^{0.95}$ points, as described above, is a good alternative to an evaluation in terms of the *FAP* and *FRP*.

Step 6: Experimental Design and Experiment

Lastly, the inquirer needs to set up the measurement system analysis study. For the covariates that are retained as arguments of the characteristic curve, one needs an experimental design suited for fitting the chosen link function q. For the covariates that are dealt with by an averaging-out strategy, one needs to ascertain that their values during the experiment are representative for regular conditions. This was illustrated for the inspection of scratches, where spreading the trials over a day improves the representativeness of appraiser fatigue. In many situations, it may not be possible to stage representative test conditions, and an averaging-out strategy may not be feasible. In addition to the experiment, one needs representative samples of parts for fitting the F_Z distributions.

Step 7: Execution, Analysis, and Conclusion

The execution and analysis of the experiment follow from steps 1 through 6, but in several case

studies we have found that residual analysis and goodness-of-fit studies are very important in this type of study. For example, in each of the studies by Van Wieringen and De Mast (2008), De Mast and Van Wieringen (2010), and Erdmann et al. (2013), items were identified with anomalous results, drawing attention to properties of the items that are poorly captured in measurement instructions and algorithms. It is important to identify such items with salient results, not only because of their possibly large impact on the quality of estimates but also because a close investigation of such parts can greatly help in improving the measurement system.

CONCLUSIONS

Categorical scales have a simple mathematical structure. But this does not mean that the underlying empirical reality, with which they are homomorphic, is simple as well. Simple methods, such as κ statistics and *FAP* and *FRP* estimated from sample proportions, are treacherous. They evaluate measurement systems in one or two numbers between 0 and 1. At best, the extreme values have a clear interpretation, but it is difficult to give a tangible meaning to intermediate values and substantiate what the difference is between, say, $\kappa = 0.6$ and $\kappa = 0.8$. As far as interpretations are offered, these typically depend critically on rather strict assumptions about conditional independence and the representativeness of samples, assumptions that are almost always violated in such applications.

The approaches that we advocate are more involved in terms of statistical modeling. Research is still ongoing to design effective approaches and new modeling techniques for common situations, and once these have found their way into statistical software packages, we do not believe that they will pose insurmountable challenges for practitioners. The approaches based on characteristic curves have a clear relation to the important notion of measurement error, which provides a solid basis for their interpretation. In addition, the results offer much more detailed insight into the functioning of a measurement system than a summary in one or two numbers. In particular, the distinction between systematic and random error is very valuable for improving a measurement system that performs poorly.

ACKNOWLEDGMENT

The authors thank Stefan Steiner for his inspiring comments and helpful ideas.

ABOUT THE AUTHORS

Jeroen de Mast is professor of methods and statistics for operations management at the University of Amsterdam. He also holds a position as principal consultant at the Institute for Business and Industrial Statistics (IBIS UvA). He is recipient of the ASQ's Brumbaugh, Nelson, and Feigenbaum awards.

Thomas Akkerhuis is a Ph.D. student and consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam.

Tashi Erdmann is a postdoctoral researcher and senior consultant at the Institute for Business and Industrial Statistics of the University of Amsterdam.

REFERENCES

Automotive Industry Action Group. (2003). *Measurement System Analysis: Reference Manual*, 3rd ed. Detroit, MI: Automotive Industry Action Group.

Ali, M. M., Mikhail, N. N., Haq, M. S. (1978). A class of bivariate distributions including the bivariate logistic. *Journal of Multivariate Analysis*, 8:405–412.

Allen, M. J., Yen, W. M. (1979). *Introduction to Measurement Theory*. Monterey, CA: Brooks/Cole.

Boyles, R. A. (2001). Gauge capability for pass–fail inspection. *Technometrics*, 43:223–229.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Conger, A. J. (1980). Integration and generalization of kappa for multiple raters. *Psychological Bulletin*, 88:322–328.

Danila, O., Steiner, S. H., MacKay, R. J. (2008). Assessing a binary measurement system. *Journal of Quality Technology*, 40:310–318.

Danila, O., Steiner, S. H., MacKay, R. J. (2010). Assessment of a binary measurement system in current use. *Journal of Quality Technology*, 42:152–164.

Danila, O., Steiner, S. H., MacKay, R. J. (2012). Assessment of a binary measurement system with varying misclassification rates. *Journal of Quality Technology*, 44:179–191.

De Mast, J. (2007). Agreement and kappa type indices. *The American Statistician*, 61:148–153.

De Mast, J., Erdmann, T. P., Van Wieringen, W. N. (2011). Measurement system analysis for binary inspection: Continuous versus dichotomous measurands. *Journal of Quality Technology*, 43(2):99–112.

De Mast, J., Van Wieringen, W. N. (2007). Measurement system analysis for categorical data: Agreement and kappa type indices. *Journal of Quality Technology*, 39(3):191–202.

De Mast, J., Van Wieringen, W. N. (2010). Modeling and evaluation repeatability and reproducibility of ordinal classifications. *Technometrics*, 52(1):94–106.

Erdmann, T. P., Akkerhuis, T. S., De Mast, J., Steiner, S. H. (2013). The statistical evaluation of a binary test based on combined samples: A case study. Submitted for publication.

Erdmann, T. P., De Mast, J. (2012). Assessment of binary inspection with a hybrid measurand. *Quality and Reliability Engineering International*, 28(1):47–57.

Erdmann, T. P., De Mast, J., Warrens, M. J. (in press). Some common errors of experimental design, interpretation and inference in agreement studies. *Statistical Methods in Medical Research*.

Feinstein, A. R., Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43:543–549.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:378–382.

Fleiss, J. L., Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, 33:613–619.

Hand, D. J. (1996). Statistics and the theory of measurement, with discussion. *Journal of the Royal Statistical Society, Series A*, 159(3): 445–492.

Hunter, J. S. (1980). The national system of scientific measurement. *Science*, 210:869–874.

Irwig, L. M., Bossuyt, P. M. M., Glasziou, P. P., Gatsonis, C., Lijmer, J. G. (2002). Designing studies to ensure that estimates of test accuracy are transferable. *British Medical Journal*, 324:669–671.

International Organization for Standardization. (1995). *Guide to the Expression of Uncertainty in Measurement*, 1st ed. Geneva: International Organization for Standardization.

Joint Committee for Guides in Metrology. (2008). *International Vocabulary of Metrology—Basic and General Concepts and Associated Terms (VIM)*, 3rd ed. ISO/IEC Guide 99–12:2007, Geneva: ISO.

Kimothi, S. K. (2002). *The Uncertainty of Measurements: Physical and Chemical Metrology—Impact and Analysis*. Milwaukee, WI: ASQ Quality Press.

Kraemer, H. C., Periyakoil, V. S., Noda, A. (2002). Kappa coefficients in medical research. *Statistics in Medicine*, 21:2109–2129.

Lord, F. M., Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*, Oxford, UK: Addison-Wesley.

Montgomery, D. C., Runger, G. C. (1993). Gauge capability and designed experiments. Part I: Basic methods. *Quality Engineering*, 6(1):115–135.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103:677–680.

Van Wieringen, W. N., De Mast, J. (2008). Measurement system analysis for binary data. *Technometrics*, 50:468–478.

Vardeman, S. B., Van Valkenburg, E. S. (1999). Two-way random effects analyses and gauge R&R studies. *Technometrics*, 41(3): 202–211.

Wallsten, T. S. (1988). Measurement theory. In: Kotz, S., Johnson, N., Eds. *Encyclopedia of Statistical Sciences.* 8th ed. New York: Wiley.