Journal of Modern Applied Statistical Methods

Volume 17 | Issue 2 Article 17

4-16-2019

Calibration of Measurements

Edward Kroc
University of British Columbia, ekroc@stat.ubc.ca

Bruno D. Zumbo
University of British Columbia, bruno.zumbo@ubc.ca

Recommended Citation
Kroc, E., & Zumbo, B. D. (2018). Calibration of measurements. *Journal of Modern Applied Statistical Methods, 17*(2), eP2780. doi: 10.22237/jmasm/1555355848

Cover Page Footnote
We would like to thank Paul Gustafson for helpful comments and advice about the initial formulations of many of the ideas that appear in this paper. Both authors would also like to register their gratitude to the UBC-Paragon Research Initiative that helped to partially fund this research, as well as to the anonymous reviewers whoh suggested many useful edits.

Journal of Modern Applied Statistical Methods
November 2018, Vol. 17, No. 2, eP2780
doi: 10.22237/jmasm/1555355848

בס"ד

ISSN 1538 − 9472

Calibration of Measurements

Edward Kroc
University of British Columbia
Vancouver, BC

Bruno D. Zumbo
University of British Columbia
Vancouver, BC

Traditional notions of measurement error typically rely on a strong mean-zero assumption on the expectation of the errors conditional on an unobservable "true score" (classical measurement error) or on the data themselves (Berkson measurement error). Weakly calibrated measurements for an unobservable true quantity are defined based on a weaker mean-zero assumption, giving rise to a measurement model of differential error. Applications show it retains many attractive features of estimation and inference when performing a naive data analysis (i.e. when performing an analysis on the error-prone measurements themselves), and other interesting properties not present in the classical or Berkson cases. Applied researchers concerned with measurement error should consider weakly calibrated errors and rely on the stronger formulations only when both a stronger model's assumptions are justifiable and would result in appreciable inferential gains.

Keywords: Measurement error, Berkson error, differential error, misclassification

Introduction

The classical framework for modeling the (additive) measurement error associated with some random variable of interest X is as follows:

$$X = W + \varepsilon, \quad \mathrm{E}(\varepsilon \mid X) = 0, \tag{1}$$

where W is the quantity that is actually observed, and ε has a mean-zero distribution (often normal) conditional on the true value X. In a psychometrics context, one considers X to be a *platonic true score*, which means it is the actual quantity of interest, latent or otherwise, while W is simply a fallible measurement of this quantity. This defines a true score in terms of validity (Klein & Cleary, 1967; Lord & Novick, 1968).

doi: 10.22237/jmasm/1555355848 | Accepted: June 3, 2018; Published: April 16, 2019.
Correspondence: Bruno D. Zumbo, bruno.zumbo@ubc.ca

Note that (1) imposes no restrictions on the structure or values of the random variables under consideration beyond the conditional expectation condition. In general, no distinction is required between continuous and categorical random variables; nor should equation (1) be interpreted as defining a true score X as a function of a measurement W. Something similar is often implied in the psychometrics literature, where (1) is usually rewritten as $W = X + \varepsilon$ and the measurement W is interpreted to be a function of the underlying (latent) variable X of interest. This is traditional, but not at all implied by the literal mathematics of these equations. No distinctions are made between the equations $X = W + \varepsilon$ and $W = X + \varepsilon$; all that matters are the conditions imposed on the error structure via some conditional expectation restriction, as in (1).

Model (1) is the measurement error model associated with the classic statistical theory (Berkson, 1950; Cochran, 1968). It is related to the measurement error models of the classical test theory (Novick, 1966) and the classical econometrics literature (Hausman, 2001), where the conditional mean-zero assumption is replaced by a formally stronger requirement. In the former setting, the error is defined in terms of reliability of the measurement, while in the latter setting, the true score is independent of an associated mean-zero error ε. See Kroc & Zumbo (2019) for a detailed and more formal discussion the relationships between these (and other) measurement error models.

Model (1) is often used within a regression framework, where the intent is to infer a relationship between the explanatory variables $\mathbf{X} = (1, X_1, \ldots, X_n)^{\mathrm{T}}$ and a response variable Y, for example:

$$Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} + \delta \tag{2}$$

where $\boldsymbol{\beta}$ is a vector of coefficients and $\delta \sim \mathrm{N}(0, \sigma^2)$.

When a set of random variables $\mathbf{X}$ is subject to measurement error as in (1), the standard coefficient estimates $\hat{\boldsymbol{\beta}}$ are biased, often with inflated standard errors, leading to considerable inflation of type I error rates (Shear & Zumbo, 2013). In the univariate setting, classical measurement error always biases the estimate β_1 towards the null (Berkson, 1950).

In contrast to model (1), consider measurement error of the Berkson type, defined by

$$X = W + \varepsilon, \quad \mathrm{E}\left(\varepsilon \,|\, W\right) = 0 \tag{3}$$

where the mean-zero assumption on the errors is now conditional on the observed value, W. Alternatively, this model is sometimes specified under the formally stronger condition that the observed value (observed score) W is independent of an associated mean-zero error ε.

The concept of measurement error is intrinsic to many applied disciplines, as it is often impractical or impossible to directly measure the true process or quantity of interest. Although the measurement error model in (1) has enjoyed a wide range of applications (for some archetypal examples see Hausman, 2001; Lord & Novick, 1968; Cochran, 1972; Heid, Kuchenhoff, Miles, Kreienbrock, & Wichmann, 2004), it is often inappropriate, especially within an observational research setting. Relying on models with such incorrect assumptions can lead to erroneous or completely meaningless inferences. The aim of this study is to explore these issues, indicate some instances where they apply, and suggest some ways the theory may be extended to treat measurement error in these more general cases.

A weakening of the classical mean-zero assumption in (1) defines what is described herein as a weakly calibrated measurement, which is a differential measurement in the sense its resultant errors may correlate with the true (or observed) quantity of interest. By definition, these measurements are more common than the traditional, or strongly calibrated, ones. This will be reflected in practice as well, as our weakened definition is easier to justify, especially within a data collection framework that is observational and uncontrolled. Moreover, this type of measurement error opens up the possibility to use theoretical or previous information about how such measurements are likely to interact with their errors to produce more reasonable, and perhaps better, inferences, in terms of mitigating bias and/or reducing variability.

Our generalized definitions introduce a level of dependence on the sample with which one aims to make inferences about parameters of the underlying target population. It will be explained why such a framework is necessary for coherent discussion of measurement in many practical settings where the structure of the measurement is affected by the sample one chooses – or is compelled – to study.

Inevitably, many measurements taken within an observational framework will satisfy neither the strong nor weak notions of calibration proposed herein. There is the necessity to use such data to generate estimates, inferences, and models, and therefore it is appropriate to discuss some of the issues practitioners should keep in mind. Crucially, as with Rubin and others (e.g. Rosenbaum & Rubin, 1984), it is recommended care be given to decide which measurements are likely to be calibrated or not for their respective target quantities of interest, and

then incorporate a routine leg of sensitivity analysis into their broader analyses to test the susceptibility of their research conclusions to varying degrees of (unknown) measurement error.

Motivating Examples

Before making formal definitions and examining the mathematical consequences, let us consider several motivating examples to justify why such definitions are required. These examples will show why the traditional models of measurement error displayed in (1) and (3) are insufficient for many applied problems. Such examples will also suggest the proposed definition for a *weakly calibrated measurement*, the main focus of this paper.

Surveys and Opinion Polling

Suppose one wishes to conduct a survey of the public approval of the federal government of Canada. This latent variable X may be studied by way of an observable quantity W which is the answer to a survey question: "*How much do you approve of the actions of the current Parliament since its formation in 2015?*" (For the purposes of this example, assume X can be described by a unidimensional real-valued random variable. A more complex but realistic formulation may consider X to have many dimensions, corresponding to the different dimensions of public approval that could be subsequently captured by a well-designed, multi-item survey.) Notice the aim is to measure the approval of a government's actions, not approval of the political party that happens to be in power. Unfortunately, these two factors are invariably confounded.

Suppose the survey question is answered by a reasonably representative random sample of voting-age Canadians, with ignorable nonresponse, and sampled individuals must record responses on a typical 5-point Likert scale (1-5), with 1 denoting strong disapproval and 5 denoting strong approval. It may be supposed that the equation $X = W + \varepsilon$ captures the relationship between the latent random variable X of interest, and the measurement W. However, neither the classical nor the Berkson conditional mean-zero error assumptions are justifiable in this setting.

First, consider the classical case: what is known about the conditional distribution of ε given X? Specifically, could the assumption $\mathrm{E}(\varepsilon \mid X) = 0$ be plausible? Such a condition ignores known psychology about how people respond to opinion polls. For example, it is well documented that supporters of a political

party in power are more likely to ignore negative actions or broken promises of that government, while staunch political opponents are more likely to fixate on and amplify every negative action to buttress their own confirmation biases (e.g. Niemi, Weisberg, & Kimball, 2001; Green, Palmquist, & Schickler, 2004). Within the context of the current example, this means, for example, if $X^{-1}(\{4, 5\})$ identifies true supporters of the current government's actions, it would be expected to find $E(\varepsilon \mid X \in \{4, 5\}) < 0$, as individuals with antagonistic political allegiances to the governing party are less likely to rate their actions highly even if they would approve of such actions in a politically neutral context. Similarly, it would be expected to find $E(\varepsilon \mid X \in \{1, 2\}) > 0$, as individuals who consider themselves politically aligned with the federal party are more likely to ignore actions they dislike due to political loyalty. Clearly, such a measurement fails to meet the criteria of the classical error model (1). For similar reasons, neither is such a measurement of Berskon-type.

However, if X is studied by drawing a reasonably representative sample of voters that is balanced across party affiliations, then it may be reasonable to suppose that these two miscalibrations essentially cancel out; i.e., $E(\varepsilon \mid X \in \{1, 2\}) + E(\varepsilon \mid X \in \{4, 5\}) = 0$. If it is assumed that $E(\varepsilon \mid X \in \{3\}) = 0$, one can still get an unskewed picture of the average true response $E(X)$ by studying the unadjusted measurements W. This is what is referred to herein as a measurement weakly calibrated (on the balanced sample of voters), defined formally in the next section.

Misclassification

It will be formally shown below that no binary measurement W of a binary true score X can ever satisfy the requirements of either (1) or (3). This immediately implies that any theory regarding these two measurement error models is insufficient for discussing problems of binary misclassification. Intuitively however, it can still be reasonable to expect the condition that the measurement, or misclassification, error when measuring binary X by binary W should be balanced, on average. More simply, if the chance that one misclassifies $X = 0$ by $W = 1$ is the same as the chance of misclassifying $X = 1$ by $W = 0$, one can still accurately study the average true response by working only with the naive measurements W.

Ecological Research

A large branch of ecological research is concerned with tracking species abundance and migratory patterns across sometimes vast geographical ranges. The eBird program managed by the Cornell Lab of Ornithology and the National Audubon Society uses a combination of professional and citizen crowd-sourced data to monitor these processes across North America at all times of year. Two such species are the Glaucous-winged Gull (*Larus glaucescens*) and the Western Gull (*L. occidentalis*), large, highly-visible coastal omnivores whose joint range spans the entirety of the North American Pacific seaboard. In our notation, one may consider an individual bird's true species identity (classification) as the random variable of interest X, and each individual classifier's species diagnostic as a separate measurement W_i. Note, for example, three professional ornithologists and two citizen-scientists would each generate a measurement W_i, five in total. Some W_i and W_j, $i \neq j$, may be recorded for the same bird, while others do not overlap at all. Nevertheless, all certainly measure the categorical quantity of interest, X.

Apart from the obvious misclassification concerns that arise with multiple observers of varying identification skills contributing to a single dataset, the particular species in question provide a whole host of other complicating measurement factors. The two species look similar, and the main diagnostic tool used by observers in the field (especially among relatively inexperienced observers) is a difference in primary wingtip color, with those of the Glaucous-winged Gull usually being varying shades of grey and those of the Western Gull usually a solid black. However, Glaucous-winged and Western Gulls frequently hybridize in Oregon and Washington where their ranges overlap, with the resulting offspring dispersing mostly northward into Washington and southern British Columbia (Bell, 1996, 1997). Such hybrids can develop any shade of wingtip color in adulthood, from light grey to solid black. The Glaucous-winged Gull also exhibits natural clinal variation in the darkness of these wingtips, from light grey at the northern edge of the species range in Alaska, to near-black at the southern extent (Bell, 1997). These complications can make reliable measurements incredibly challenging.

Moreover, there are large subpopulations of both species that are migratory. In the winter months, the Salish Sea region of British Columbia and Washington is home to light wingtipped Alaskan Glaucous-winged Gulls, dark wingtipped Californian Western Gulls, as well as the resident (nonmigratory) hybrids and clinal variants present in the region year-round. As a result, species counts in the Salish Sea during the winter months are particularly subject to measurement error.

Such errors are clearly incredibly complex and could not reasonably hope to satisfy the stringent conditions imposed by the classical or Berkson models. However, due to the seasonally changing composition of the study population, it may be argued that the misclassification errors essentially cancel out over time; i.e., measurements are weakly calibrated (over time). In fact, a sampling scheme can be specifically designed to take advantage of this phenomenon, thereby inducing weak calibration into our pool of measurements. Identification and tracking of many other North American species pose similar measurement challenges.

Calibration of Measurements

Recall the archetypal (additive) measurement error model:

$$X = W + \varepsilon\,, \tag{4}$$

where X is the quantity intended to observe (the so-called *true value* or *true score*), W is the surrogate for X that is actually observed, and ε is the additive discrepancy between the two. The observable W is a *measurement* for X, and X is measured with error ε by W.

It should be noted that this definition of *true score* does not align with classical test theory (Lord & Novick, 1968), where a true score is defined as the expectation of the observed scores over (infinitely many) independent and identical repetitions of the measurement process. Instead, as noted above, it is a *platonic true score* or *construct* (Klein & Cleary, 1967; Lord & Novick, 1968; Borsboom & Mellenbergh, 2002). Such a definition does not force any *a priori* structure on a true score, nor does it rely on a frequentist interpretation of equation (4).

Calibrated Measurements

Implicit in the specification of (4) is the existence of an underlying measurable space $(\Omega, \mathcal{F})$ on which the true score X and the observable proxy W (and thus the error ε) are marginally defined. Naturally, one interprets Ω as the *population* on which such random variables are definable, which either coincides with or contains the population on which one ultimately aims to study via estimation of population parameters (e.g. the population mean) and corresponding inferences.

For a subset $\mathcal{S} \in \mathcal{F}$, one can state the following definitions:

- Measurement W in (4) is strongly calibrated to X on $\mathcal{S}$ if $\mathrm{E}(\varepsilon \mid \mathcal{S}, X) = 0$.
- Measurement W in (4) is Berkson calibrated to X on $\mathcal{S}$ if $\mathrm{E}(\varepsilon \mid \mathcal{S}, W) = 0$.
- Measurement W in (4) is weakly calibrated to X on $\mathcal{S}$ if $\mathrm{E}(\varepsilon \mid \mathcal{S}) = 0$.

Usually, $\mathcal{S}$ is the sample used to study the population properties of the true score X. When the discussion does not depend on the choice of conditioning set $\mathcal{S}$, or when the choice is obvious, the notation will be suppressed, writing simply $\mathrm{E}(\varepsilon)$ rather than $\mathrm{E}(\varepsilon \mid \mathcal{S})$.

The notation introduced in the above conditional expectation conditions requires some clarification. It is meant to be a generalized shorthand of the $\mathrm{E}(U \mid V)$ notation for random variables U, V, where the conditional expectation is understood to be taken over the σ-algebra generated by the random variables V; i.e.

$$\mathrm{E}(U \mid V) = \mathrm{E}(U \mid \sigma(V)),$$

where $\sigma(V) = \{V^{-1}(G): G \in \mathcal{G}\}$ is a sub-σ-algebra of $\mathcal{F}$, and $\mathcal{G}$ denotes the image σ-algebra of $V : (\Omega, \mathcal{F}) \to (\Gamma, \mathcal{G})$. Usually, Γ denotes $\mathbb{R}^d$ or $\mathbb{Z}^d$ and $\mathcal{G}$ denotes the Borel sets on $\mathbb{R}^d$ or the power set on $\mathbb{Z}^d$, respectively.

With this in mind, one defines

$$\mathrm{E}(U \mid \mathcal{S}, V) = \mathrm{E}\left(U \mid \left\{\mathcal{S} \cap V^{-1}(G) : G \in \mathcal{G}\right\}\right).$$

Notice $\left\{\mathcal{S} \cap V^{-1}(G) : G \in \mathcal{G}\right\}$ is still a sub-σ-algebra of $\mathcal{F}$. When $\mathcal{S} = \Omega$, the definition of a strongly calibrated measurement recovers the classical measurement error model in (1), and the definition of a Berkson calibrated measurement recovers the Berkson measurement error model in (3). These are models of nondifferential measurement error, whereas the notion of weak calibration allows for the possibility of differential error. Clearly, any strongly or Berkson calibrated measurement is also weakly calibrated on the same sample.

Calibrated measurements are common in the physical sciences, where the notion of calibration has a very physical antecedent. Indeed, the notion of calibrating a measurement as one would calibrate an experimental design has appeared in the literature since at least the work of Wald (1940). The assumptions of strong or Berkson calibration are often quite reasonable given some measurement apparatus of a well-defined, observable physical phenomenon,

especially when a treatment is manipulated experimentally. Both ideas have been used to great effect in many fields. One notable example from environmental epidemiology comes from the work of Heid et al. (2004) where, among other measurements, the concentration of a particular gas within a single residence was measured (strong calibration), and then a single concentration level was assigned as a common value of exposure to all occupants of the single residence (Berkson calibration).

The choice of the term "calibration" is deliberate to also distinguish the idea from that of unbiasedness. Superficially, a calibrated measurement could be called an unbiased measurement, because its defining property is that the (conditional) expected value of the measurement equals the expectation of the random variable of interest. However, the notion of bias is only sensible within the context of estimators. Estimators are statistics formed from a given sample (actual or theoretical), and unbiased estimators equal their targets of estimation in expectation. But measurements in our framework do not attempt to estimate anything; indeed, it is the measurements that precisely comprise the sample of study itself. The notion of calibration does not rely on the combination of measurements into a sample statistic, and so does not concern estimation. Consequently, the idea of bias is simply not applicable.

This distinction becomes clearer in comparing this view of measurement with the one from Lord and Novick (1968). Specifically, in Chapter 8.4 of their text, they define unbiasedness of their measurements in terms of the notion of parallel tests. This gels with the traditional usage of the unbiasedness terminology, because the existence of even hypothetical parallel tests generates a population space from which all observed measurements can be considered sampled. In contrast, the current notion of measurement relies on no such theoretical space of parallel measurements on which true scores for individuals are fixed. Our defining of expectations for our various notions of calibration are taken over a single sample, generated via $\mathcal{S} \subseteq \mathcal{G}$, and for a single measurement. Consequently, the usual frequentist interpretation of expectation should not be applied in this setting; the expectations in our definitions of calibration are meant only to indicate a typical value (an integral) of a random variable.

Because of this distinction, a set of measurements in this framework is not useful for studying properties of individual sample points (e.g. individual respondents). If the target of study is a property of an individual, then multiple measurements are required to apply our notions of calibration. In this case, the sample $\mathcal{S}$ would comprise a singleton; thus, calibration (of any type) for a particular measurement can only hold if the measurement is error-free on the

sample. However, it is easy to see how our definitions can be extended to encompass a notion of calibration over multiple measurements for an individual. In fact, this is a special case of a generalized definition of calibration to a composite set of measurements (see below). Then, one simply requires the sample $\mathcal{S}$ of interest to contain only a single individual.

Extending the notion of calibration from a single measurement to a composite set of measurements is straightforward. As an example, if X represents a student's mathematics ability, a typical construct from the social science literature, and the aim is to measure this quantity by administering a 10-question examination, then a standard measurement for X could be the unweighted sum of the scores of the 10 exam questions. More generally, one may consider the model

$$X = a_1 W^{(1)} + a_1 W^{(2)} + \ldots + a_m W^{(m)} + \varepsilon ,$$

where X is some latent variable to be measured by the (weighted) composite score given by the linear combination of the observable $W^{(k)}$s. The associated error, ε, may then be treated as a single error of the composite score as a whole (as written), or as a composite of errors itself, $\varepsilon = \varepsilon^{(1)} + \ldots + \varepsilon^{(m)}$, whichever is more appropriate for the problem at hand. In any case, one may then speak about whethe or not the composite score is calibrated to the latent variable X; i.e. if $\mathrm{E}\left(\varepsilon \mid \mathcal{S}, W^{(1)}, \ldots, W^{(m)}\right) = 0$, (strong calibration), or if only $\mathrm{E}\left(\varepsilon \mid \mathcal{S}\right) = 0$ (weak calibration).

Miscalibrated Measurements

When (4) holds for some quantity of interest X and an observable proxy W, but $\mathrm{E}\left(\varepsilon \mid \mathcal{S}\right) \neq 0$, the measurement is *miscalibrated* for X on $\mathcal{S}$. Miscalibrated measurements are common in observational studies, especially in the social sciences when trying to measure properties based on human perception, opinion, or interpretation. For example, people are likely to under or over-report feelings of self-worth or depression based on various social stigmas (real or perceived), regardless of how controlled a research setup may be.

Indeed, it is not sufficient to simply define the variables X and W to characterize the type of resulting measurement error. For example, people often over-report their annual income when applying for credit, while others (possibly the same people) are likely to underreport this same figure when filing a tax return. In both cases, the true value of interest is the respondent's actual annual income,

and the measurement observed is the self-reported response to a superficially identical enquiry: *What is your annual income?* In both scenarios, measurements are likely to be miscalibrated, but the direction of this miscalibration, as given by the sign of $E(\varepsilon)$, changes depending on the context in which the measurement is taken.

Social pressures may induce people to underreport their age, over-report their activity level, exaggerate their like or dislike for a political party or position given their preconceived alliances, etc. These measurements are miscalibrated. Although they are often treated analytically as true scores observed without error, such analyses are inherently distorted. More sophisticated analyses that incorporate an idea of measurement error tend to always include the assumption of strong calibration, something that is simply untenable in most social science contexts, particularly in surveys and opinion polling. The ubiquitous presence of miscalibrated measurements in the social sciences makes drawing accurate inferences inherently more difficult, and sometimes impossible.

In some lucky cases, it may be possible to combine miscalibrated measurements in such a way as to produce a composite measurement that is plausibly close to calibrated. Consider an ecological example investigating the fecundity of a population of Mallards (*Anas platyrhynchos*). Fecundity is a latent variable that captures the reproductive success of a breeding population. One may define the measure of fecundity of a Mallard pair to be the simple sum of the hen's maximal clutch size (*CS*) and the number of young successfully fledged (*FL*). These counts both typically range between 0 and 12 (Drilling, Rodger, & McKinney, 2002), with $FL \leq CS$.

It is often difficult to exactly observe *CS* and *FL*. Clutches can usually only be observed by disturbing the incubating hen, and eggs may fall out of the nest or be stolen by predators while observation takes place (Götmark, 1992). Moreover, as eggs are not laid simultaneously (Hill, 1984), there is no correct time to count the clutch that would guarantee the observation of maximum clutch size. Consequently, it may be necessary to use a proxy for *CS*: the number of eggs hatched; i.e., the initial brood size, denoted by CS_{obs}. Necessarily, $CS_{\text{obs}} \leq CS$. Brood size is easy to observe and non-disruptive, making it an attractive measurement for *CS*. Similarly, it is often difficult to observe young birds all the way until they have properly fledged. Commonly, young are observed until a certain age, after which all surviving young are declared fledged. Mallards typically fledge between 50-60 days of age (Drilling et al., 2002), but most mortality events occur within the first two weeks of life. Therefore, it may be possible to only observe them until day 14, FL_{obs}. Necessarily, $FL_{\text{obs}} \geq FL$.

The two measurements are miscalibrated by definition, that is:

$$CS = CS_{\text{obs}} + \varepsilon_{CS}, \quad FL = FL_{\text{obs}} + \varepsilon_{FL},$$

where $\mathrm{E}(\varepsilon_{CS}) \geq 0$ and $\mathrm{E}(\varepsilon_{FL}) \leq 0$. However, it may in fact be feasible to suppose the composite measure for fecundity is approximately weakly calibrated:

$$Fec = CS + FL = CS_{\text{obs}} + FL_{\text{obs}} + \delta,$$

where $\mathrm{E}(\delta) \approx 0$. This calibration relies on the assumption that the magnitude of *CS* and *FL* errors are likely to somewhat balance each other out, something the observation process suggests may be reasonable.

Notice the error need not balance on common values of *CS* and *FL*, as would be required for strong calibration. Although it is theoretically possible, it is more difficult to justify in practice. Moreover, such a strong assumption does not necessarily add strength to any resultant inferences one may wish to make on the sample population. As will be discussed, weakly calibrated measurements can be just as precise as strongly calibrated measurements. In fact, they can be more precise than perfect measurements; i.e. those measured without error.

The biggest challenge of dealing with miscalibrated measurements is there is often no way of directly quantifying the degree of miscalibration. The structure of the measurements themselves must be relied on to make reasonable assumptions about the degree of miscalibration, as in the fecundity example, or sometimes information from past studies may be used to decide how the errors are likely to be probabilistically distributed. Providing a clear definition of the true value of interest and the measurement observed can provide clues about the distributional structure of the error: e.g. is it always of the same sign, or is there some subpopulation likely to produce less calibrated measurements than the rest of the population? Even with this type of information though, it is unlikely to have a clear picture of the distribution or of the measurement error to completely adjust for it in subsequent inferences. In practice therefore, it will often be vital to perform some type of sensitivity analysis on the measurement assumptions, for example by varying the degree of miscalibration and examining how this affects inferences based on these quantities.

Uncalibrated Measurements

There certainly exist situations where (4) is a reasonable model, yet the error term may not even be integrable. In certain situations where the intent is to model extremely volatile processes, the value of a certain stock on a financial market perhaps, the most reasonable error distribution may be extremely heavy-tailed and so nonintegrable. The corresponding measurements would be classified as *uncalibrated* within our current framework, although that is not to say the analytical issues are intractable in this alternative setting. A proper treatment would require finer assumptions on the structure of the probability measures themselves, rather than merely some kind of moment condition on the errors. Although these issues are not taken up in the current paper, it is important to note that there exists a need to clarify the meaning of measurement in this more delicate setting.

Estimation and Inference with Calibrated Measurements

Whenever a sample $\mathcal{S}$ is used to perform estimation of or inference on some population parameter, the assumption is usually made the sample is representative of the actual population phenomena one aims to study. More formally, if θ_X represents the (population) target of estimation and φ denotes its estimator on a sample $X_1,\ldots, X_N$, where $\#(\mathcal{S}) = N$, one typically assumes that

$$\mathrm{E}_{\mathcal{G}}\left[\varphi\left(X_1,\ldots,X_N\right)\mid\mathcal{S}\right] \coloneqq \mathrm{E}_{\mathcal{G}}\left[\varphi\left(X_1\left(\omega_1\right),\ldots,X_N\left(\omega_N\right)\right)\right] = \theta_X \tag{5}$$

where $\mathcal{S} = \{\omega_1,\ldots,\omega_N\}$, and the expectation is taken over some σ-algebra $\mathcal{G}$ such that $\sigma(\mathcal{S}) \subseteq \mathcal{G} \subseteq \mathcal{F}$. In simpler language, this is simply the statement that φ is an unbiased estimator of θ_X on all samples $\mathcal{S}$ such that $\sigma(\mathcal{S}) \subseteq \mathcal{G}$. Good sampling methodology, such as simple random sampling, can often ensure such a condition holds, in which case $\mathcal{S}$ is *φ-representative* for θ_X on $\mathcal{G}$. From a frequentist perspective, one often requires the stronger condition that $\mathcal{G} = \mathcal{F}$, although this is not strictly necessary to make sensible inferences, and indeed, sometimes it can be more plausible to assume the existence of a strictly smaller $\mathcal{G} \subseteq \mathcal{F}$ in practice.

Unbiasedness of the Naive Sample Mean Estimator

By far, the most common estimand of interest in scientific research is the *typical* response of an observable. Although many formulations of what is *typical* exist,

some more appropriate than others depending on both the scientific question under investigation and the actual distribution of the quantity being measured, probably the most common instantiation for continuous quantities is given by the population mean.

If W is weakly calibrated to X on some $\mathcal{G} \subseteq \mathcal{F}$, and if $\mathrm{E}(X) < \infty$, then $\bar{W}$ is an unbiased estimator of the population mean μ_X on $\mathcal{G}$:

$$\begin{aligned} \mathrm{Bias}_{\mathcal{G}}\left(\bar{W}\right) &:= \mathrm{E}_{\mathcal{G}}\left[\bar{W} - \mu_X\right] \\ &= \mathrm{E}_{\mathcal{G}}\left[\frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}}\mathrm{E}\left(W_i \mid \mathcal{S}\right)\right] - \mu_X \\ &= \mathrm{E}_{\mathcal{G}}\left[\frac{1}{N_{\mathcal{S}}}\sum_{i=1}^{N_{\mathcal{S}}}\mathrm{E}\left(W_i \mid \mathcal{S}\right) - \mathrm{E}\left(\varepsilon_i \mid \mathcal{S}\right)\right] - \mu_X \\ &= \mathrm{E}_{\mathcal{G}}\left(\bar{X}\right) - \mu_X = 0 \end{aligned} \tag{6}$$

Note (6) holds whether W is strongly, Berkson, or weakly calibrated to X on $\mathcal{G}$.

Variability of the Naive Sample Mean Estimator

Any strongly calibrated measurement necessarily has an associated error that is uncorrelated to X, the quantity of interest studied via the measurement W. This is not true of weakly calibrated measurements. Indeed, if one calculates

$$\mathrm{Cov}_{\mathcal{G}}\left(\varepsilon, X\right) = \mathrm{E}_{\mathcal{G}}\left[W \cdot \mathrm{E}\left(\varepsilon \mid \mathcal{S}, X\right)\right] \tag{7}$$

which is necessarily zero only if $\mathrm{E}\left(\varepsilon \mid \mathcal{S}\right) = 0$ is assumed.

The consequences of (7) are manifest for quantifying the uncertainty involved in any estimation procedure that uses the measurement W. Regarding the naive (observable) sample mean estimator $\bar{W}$, one sees that if W is a measurement for X, then

$$\mathrm{Var}_{\mathcal{G}}\left(\bar{W}\right) = \frac{1}{N}\mathrm{Var}_{\mathcal{G}}\left(X\right) + \frac{1}{N^2}\sum_{i=1}^{N}\mathrm{Var}_{\mathcal{G}}\left(\varepsilon_i\right) - \frac{2}{N^2}\sum_{i=1}^{N}\mathrm{Cov}_{\mathcal{G}}\left(X_i, \varepsilon_i\right). \tag{8}$$

It is assumed X_i does not correlate with X_j and that ε_i does not correlate with ε_j, for all $i \neq j$. This will be the case if the measurements W_i are considered to be drawn

as an independent sample. Moreover, if identical distributions is assumed, then (8) simplifies to

$$\mathrm{Var}_{\mathcal{G}}\left(\bar{W}\right)=\frac{1}{N}\left[\mathrm{Var}_{\mathcal{G}}\left(X\right)+\mathrm{Var}_{\mathcal{G}}\left(\varepsilon\right)-2\cdot\mathrm{Cov}_{\mathcal{G}}\left(X,\varepsilon\right)\right]. \tag{9}$$

The naive estimator $\bar{W}$ is consistent for μ_X if, as before, W is weakly calibrated to X on $\mathcal{G}$, and if $\mathrm{E}(X) < \infty$. This follows from an application of Markov's inequality. Asymptotic normality also follows if it is assumed either of the regularizing conditions of Lyapunov's or Lindeberg's central limit theorems, and then apply the chosen theorem (Durrett, 2010). In most practical settings, these are mild conditions which amount to assuming little more than finitude of the second moments of X and ε.

Although it is tempting to intuit that equations (8) or (9) imply that $\bar{W}$ is a noisier estimator of the population mean than the true sample mean $\bar{X}$, due to the presence of measurement error, this conclusion is not necessarily true. In fact, an estimator derived from observables measured with error may be less noisy than an estimator derived from true scores. For this to occur, such an estimator would have to take advantage of the extra correlation in the errors with the true (unobserved) values of X_i to more precisely estimate the population parameter of interest, in this case μ_X. More precisely, this will occur if and only if the last two terms in (8) or (9) sum to a negative value.

Such a situation is impossible for strongly calibrated measurements by definition. Indeed, since $\mathrm{E}\left(\varepsilon \mid \mathcal{S}\right)=0$ is assumed, it follows that $\mathrm{Cov}_{\mathcal{G}}(X, \varepsilon) = 0$. Consequently, the last term in (8) drops out and the classical result (Berkson, 1950) is recovered such that

$$\mathrm{Var}_{\mathcal{G}}\left(\bar{W}\right)\geq\mathrm{Var}_{\mathcal{G}}\left(\bar{X}\right). \tag{10}$$

Similarly, one may derive the inverse inequality for Berkson calibrated measurements, because $\mathrm{E}\left(\varepsilon \mid \mathcal{S}\right)=0$ implies $\mathrm{Cov}_{\mathcal{G}}(\mathrm{X}, \varepsilon) = \mathrm{Var}_{\mathcal{G}}(\varepsilon)$.

For weakly calibrated measurements, an inequality like (10) cannot be concluded so directly. If the covariance of the error with the unobserved true value is small relative to the variance of the error itself for each item $i = 1,\ldots, N$, then (10) holds. This will be the case when one does not expect the measurement errors to correlate in any meaningful way with the true quantity X being measured by proxies W_i, but this property is not an inherent feature of the definition.

Traditionally, this extra property is often implicitly assumed when working with imperfect measurements, and indeed it is sometimes a reasonable assumption. However, there are scenarios when this assumption fails. In such a setting, measurement error afflicted estimator is actually *more precise* than the corresponding unobserved estimator built from the actual true scores themselves. This can lead to either unexpected inferential gains or deceiving levels of precision when miscalibration is present.

Consider an example wherein of a survey of N people about their sex lives. One of the questions asked is, "*During a typical week, how many times do you have sex?*" Such a question comes equipped with a multitude of social and cultural baggage, which will influence the candor (and perhaps the genuine recall) of some respondents. For this particular type of question, one might expect to see a regressing effect, where respondents will be more likely to report answers close to what the perceived average, or norm, is already. This means that respondents with true scores at both extremes of the response spectrum are likely to mitigate their responses to appear more normal (this effect can be present even in anonymous surveys; see Fisher, 1993). In the case of the question about frequency of sex, respondents who typically do not have sex at all may be more likely to respond that they do, while respondents who typically have a lot of sex may be more likely to underreport their frequency.

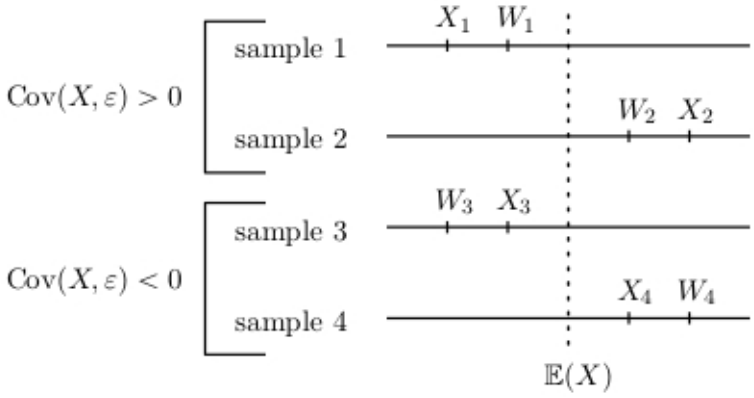

Figure 1. When Cov(X, ε) > 0, the measurements W_i are forced to lie more tightly around the population mean, $\mathbb{E}(X)$, on average

These reporting biases reflect distinct cultural pressures that point to a common expected norm of social behavior (in this case, some sex, but not too much, is what the respondent may perceive to be socially expected). In order for this survey question to actually be a (weakly) calibrated measurement for the true response among the target population, the total magnitude of underreported scores must approximately equal the total magnitude of overreported scores. One way to ensure this would be to demand the sampled population contain an equal number of people who underreport as who over-report, and the magnitude of these under and over-reports cancel each other out. In such a case, the measurement error afflicted estimator $\bar{W}$ of μ_X will have smaller variance than the true score estimator $\bar{X}$, as $\text{Cov}(X, \varepsilon) > 0$. In this case, the measurement errors actually pull the reported responses toward the population average (see Figure 1).

What is more likely to occur in the above scenario is for the sample to contain an equally mixed collection of individuals who feel compelled to under or over-report the frequency of their sexual activity, but there is no good reason to believe that the magnitudes of these measurement errors will approximately balance out; i.e., it cannot be assumed W is calibrated to X. Consequently, the estimator $\bar{W}$ for μ_X will contain some amount of bias and simultaneously appear *overly precise*, as the structure of the measurement errors are likely to display the discussed composite mitigating effect. Interpreting such estimates becomes doubly challenging in such a context, as the measurement errors work against us twice as hard. The necessity of simulating the inferential effects of different degrees of miscalibration is the clearest way out of this interpretive dead-end.

Returning to the fecundity example, suppose the aim is to estimate the mean fecundity of the population. The intent is to calculate the mean value of $CS + FL$, the error-free score. Because this is not possible, however, one may end up calculating the mean value of $CS_{\text{obs}} + FL_{\text{obs}}$, an (approximately) unbiased estimator of the population mean by the assumption of (approximate) weak calibration. Now, if the further assumption holds that the two measurement errors are sufficiently positively correlated with our response of interest, i.e., if

$$\text{Cov}\left(\varepsilon_{CS}, Fec\right) + \text{Cov}\left(\varepsilon_{FL}, Fec\right) \gg 0, \tag{11}$$

then, by equation (8), the unbiased estimate of average fecundity derived from $CS_{\text{obs}} + FL_{\text{obs}}$ is actually *more precise* than if the alternative unbiased estimate of average fecundity derived from the unobserved was used, but true, $CS + FL$.

Unfortunately, assumption (11) is untestable, although the individual research problem may offer support for such an assumption. For the current

example, assumption (11) would imply the measurement errors are typically bigger (not in magnitude, but in the well-ordered sense) for Mallard pairs with higher fecundity. This assumption happens to be true for at least the second term of our covariance condition, that is $\mathrm{Cov}(\varepsilon_{FL}, Fec) \gg 0$. Because $\varepsilon_{FL} \leq 0$, the greater the number of observed fledglings, the closer the error is to 0.

Unfortunately, this argument works equally well in the opposite direction for the other term in (11). There is a guaranteed negative correlation between CS_{obs} and ε_{CS}. Thus, depending on the magnitude of this covariance, (11) may fail. Furthermore, the errors ε_{CS} and ε_{FL} happen to be somewhat correlated: if the first is very large, then the second is necessarily very small, further complicating the residual variance analysis.

Bias reduction of miscalibrated measurements can be accomplished by creating composite measurements. As in the above example, both *CS* and *FL* are reasonable measurements of fecundity on their own, although they are both fundamentally miscalibrated. Nevertheless, combining these two measurements reduces the bias due to their individual miscalibrations, resulting in a more accurate measurement, without necessarily producing any loss of precision (by weak calibration).

Regression with Calibrated and Miscalibrated Measurements

Consider the setting of an idealized simple linear regression, where i.i.d. measurements are drawn from a target population and model the effect of some observable predictor X of interest on a response Y, *without measurement error*, using the following model:

$$Y = \beta_0 + \beta X + u \tag{12}$$

If it is not possible to perfectly measure X, then one is forced to consider the following model, where W is the measurement for X related via the typical additive setup:

$$Y = \beta_0' + \beta' W + u' = \beta_0' + \beta' \left(X - \varepsilon \right) + u' \ . \tag{13}$$

The intent is to perform the naive regression using the measurements W, and then relate the output back to the model of actual interest in (12). It is the effect of X on Y that is of interest, not the effect of the proxy W on Y. Under the standard linear

regression assumptions, the output of the naive analysis can be summarized in Table 1.

When the measurement W is only weakly calibrated to X, then β' is either attenuated towards the null, or exaggerated away from it, depending on the average structure of the measurement errors as illustrated in Figure 1. When W is miscalibrated for X, then the effect of measurement error on the simple regression coefficient also depends on the magnitude of the first moments of X and ε (this is because $\mathrm{E}(\varepsilon) \neq 0$ under miscalibration, and so $\mathrm{E}(X) \neq \mathrm{E}(W)$). In general, both weakly calibrated and miscalibrated measurements are not guaranteed to produce attenuating effect estimates, as the correlations between the measurements and their errors can be negative, a common situation in practice as has already been seen. This once again reflects the ability of weakly calibrated measurements to leverage the information contained in the interaction between those measurements and their errors to yield possible inferential gains.

When W is strongly calibrated to X, there are two sources of inflation as β' is attenuated towards the null. When W is Berkson calibrated to X, then the only source of inflation comes from the variability in the measurement errors, ε, themselves. When only weak calibration is assumed, then the residual variance is not necessarily inflated, the interaction between the measurements and their errors becomes important. Once again, estimators derived from weakly calibrated measurements can be more precise than those derived from true scores. However, in the context of simple linear regression, the corresponding naive estimator is likely to be affected by non-negligible bias.

Table 1. Summary of how the naive regression in (13) relates to the measurement error free model (12); see Buzas, Tosteson, and Stefanski (2003) for derivations

Calibration type	Slope β'	Residual variance $\sigma^2_{u'}$
Strong	$\beta \cdot \frac{\sigma_X^2}{\sigma_X^2 + \sigma_\varepsilon^2}$	$\sigma_u^2 + \beta^2 \sigma_X^2 \cdot \frac{\sigma_\varepsilon^2}{\sigma_X^2 + \sigma_\varepsilon^2}$
Berkson	β	$\sigma_u^2 + \beta^2 \sigma_\varepsilon^2$
Weak	$\beta \cdot \frac{\sigma_{XW}}{\sigma_W^2} + \frac{\sigma_{uW}^2}{\sigma_W^2}$	$\sigma_u^2 + \beta^2 \sigma_X^2 - \frac{(\beta \sigma_{XW} + \sigma_{uW})^2}{\sigma_W^2}$

The Misclassification Problem

Nothing about our notions of calibration require continuity of either the true or observed scores. Nevertheless, it is often convenient to treat the discrete setting somewhat distinctly, the so-called misclassification problem. This is perhaps especially useful when the quantity of interest is nominal in nature, as in medicine when measuring a person's gender or whether or not one smokes.

Consider the basic problem of the misclassification of a binary random variable. Under model (4), if $X, W \in \{0, 1\}$, then the error ε follows a multinomial distribution on $\{-1, 0, 1\}$. It is easy to see that such a (nontrivial) measurement W for X cannot be strongly calibrated. Indeed, if one lets p_{-1}, p_0, p_1 define the probability mass function of ε, where $p_{-1} + p_0 + p_1 = 1$, then one may calculate

$$\mathrm{E}\left(\varepsilon \mid \mathcal{S}, W = 0\right) = 0 \cdot p_0 + 1 \cdot p_1 = p_1,$$

and similarly,

$$\mathrm{E}\left(\varepsilon \mid \mathcal{S}, W = 1\right) = p_{-1},$$

for any sample $\mathcal{S} \subseteq \Omega$. Therefore, the only way W can be strongly calibrated to X is if $p_1 = p_{-1} = 0$; i.e. if W is in fact a perfect measurement, free of error. The same argument also shows that any (nontrivial) measurement W for X cannot be Berkson calibrated.

Thus, the notion of strong calibration (the classical measurement error model) is not useful for the basic misclassification problem. Weak calibration is quite appropriate however, because any measurement error given by a balanced multinomial distribution (i.e. one where $p_{-i} = p_i$ for all i) automatically corresponds to a weakly calibrated measurement.

Within the context of the classical test theory, this failure was already known to Lord and Novick (1968); see Chapter 2.9 of their text. To our knowledge, it has not yet been noted that the more general classical measurement error model (strong calibration) is also useless here. However, as our notion of true score coincides with the platonic one introduced by Sutcliffe (1965) in his analysis of the binary classification problem, and as our the definition of weak calibration represents a considerable weakening of the classical assumptions, it should not be too surprising the same difficulties are not encountered.

When X and W are binary variables, the naive sample mean $\bar{W}$ is the sample proportion of successes, $W = 1$, to failures, $W = 0$. If the error ε_i is given by the

same balanced multinomial distribution for all i (necessarily a weakly calibrated measurement), then equation (9) implies (10) because

$$2 \cdot \mathrm{Cov}\left(X, \varepsilon\right) = 2 \cdot \Pr\left(X = 1, \varepsilon = 1\right) < 2 \cdot \Pr\left(\varepsilon = 1\right) = p_{-1} + p_1 = \mathrm{Var}\left(\varepsilon\right).$$

Unlike the continuous case, the only time a binary measurement W of a binary true score X can produce an estimator of the sample proportion with smaller variance than that generated by the true score is if $p_{-1} = 0$ and $p_1 < 2 \cdot \Pr(X = 1, \varepsilon = 1)$. This implies two things: first, the measurement W must be miscalibrated for X; and second, the measurement W enjoys 100% specificity, in the sense that if $X = 0$ then $W = 0$.

A simple calculation shows that $\mathrm{Cov}(X, \varepsilon) = \Pr(X = 1)p_1$ if W is a calibrated measurement for X, and that $\mathrm{Cov}(X, \varepsilon) = \Pr(X = 1)p_{-1}$ otherwise. In both cases, $\mathrm{Cov}(X, \varepsilon) \geq 0$; this result paired with the content of Table 1 proves that a binary observable subject to measurement error, either calibrated or miscalibrated, will always attenuate the linear main effect β' of the classical one-way ANOVA model (13) towards the null, in stark contrast to the continuous setting.

Any good measurement W for X should be both accurate and precise on average. The mean square of the measurement error ε is the most classical way to quantify this balance. This quantity captures a calibration-variance tradeoff in direct analogy with the bias-variance tradeoff of the mean-squared error of a typical point estimator. It is clear that

$$\begin{aligned} \mathrm{E}\left(\varepsilon^2\right) &= \mathrm{E}\left(\left(X - W\right)^2\right) \\ &= \mathrm{Var}\left(\varepsilon\right) + \mathrm{Calib}\left(\varepsilon\right)^2 \end{aligned}$$

where $\mathrm{Calib}(\varepsilon) = \mathrm{E}(X - W)$ is the natural measure of how calibrated the measurement W is for X. Using this metric, the simple trinomial error generated by measuring a binary response X by a binary measurement W yields a mean-squared error of $\mathrm{E}(\varepsilon^2) = p_{-1} + p_1$. From this, one sees that there are many miscalibrated measurements W that are "good" proxies for X, while there are also many calibrated measurements W that are "poor" proxies for X. For example, under a uniform trinomial error distribution, $\mathrm{E}(\varepsilon^2)$ is maximized, even though the corresponding measurement is calibrated for X. Such a measurement, although perfectly calibrated, is hardly useful from an analytical perspective.

The mean-squared error is by no means the only quantifier one may want to consider when assessing the "goodness" of a particular measurement. An alternative metric is discussed in the Appendix that has some interesting theoretical structure.

Final Remarks

The quality of models depends on how well the data, and the observational or experimental process or object of study, respect the mathematical assumptions required of those models. Although classical measurement error theory relies heavily on assumptions of strong or Berkson calibration, the restrictions imposed by these assumptions are often not respected in practice, especially in the environmental, ecological, and social sciences. By relying on the use of weakly calibrated measurements only, one may continue to use the inferential tools of classical statistics while simultaneously better approximating the real-world processes one attempts to study.

Outside of a tightly controlled, laboratory setting, weak calibration is, by and large, a more scientifically accurate representation of an errors-in-measurement phenomenon. Weak calibration does not restrict any fibers of the error measures to be constant, unlike the strong or Berkson calibrated models. Consequently, the errors themselves can be imbued with some of the structure of the process attempted to be measured. This information can sometimes be exploited analytically for inferential gains, either by relying on theoretical knowledge of the observational process or on a validation sample of measurements. Moreover, weak calibration allows us to simultaneously treat the theories of continuous errors in measurement and discrete misclassification models.

The mathematical assumptions made as analysts are often restrictions imposed on the phenomenon of study, not it's *a priori* properties. The covariance conditions imposed by an assumption of weak calibration are much weaker than what is commonly assumed, and can often produce inferential results that are just as good as, or perhaps better than, those yielded by strong calibration. Unless the restrictions placed on a phenomenon by an assumption of strong or Berkson calibration are likely to hold (see Heid et al., 2004 for a good practical discussion), it is recommended treating errors in measurement as only weakly calibrated, if indeed they can be reasonably assumed calibrated at all.

As Tukey once advised, "Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which

can always be made precise" (Tukey, 1962, pp. 13-14). There is little point in analyzing a phenomenon if one is unwilling to use assumptions and apply models that are likely to reasonably reflect the structure of the object of study. With regards to the errors in measurement problem, weaker assumptions are far less likely to produce erroneous answers.

Acknowledgements

The first author would like to thank Paul Gustafson for helpful comments and advice about the initial formulations of many of the ideas that appear in this paper. Both authors would also like to register their gratitude to the UBC-Paragon Research Initiative that helped to partially fund this research, as well as to the anonymous reviewers who suggested many useful edits.

References

Bell, D. A. (1996). Genetic differentiation, geographic variation and hybridization in gulls of the *Larus glaucescens-occidentalis* complex. *The Condor, 98*(3), 527-546. doi: 10.2307/1369566

Bell, D. A. (1997). Hybridization and reproductive performance in gulls of the *Larus glaucescens-occidentalis* complex. *The Condor, 99*(3), 585-594. doi: 10.2307/1370471

Berkson, J. (1950). Are there two regressions? *Journal of the American Statistical Association, 45*(250), 164-180. doi: 10.2307/2280676

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs: A comment on Schmidt and Hunter. *Intelligence, 30*(6), 505-514. doi: 10.1016/s0160-2896(02)00082-x

Buzas, J. S., Tosteson, T. D., & Stefanski, L. A. (2003, April). *Measurement error* (Institute of Statistics Mimeo Series No. 2544). Raleigh, NC: North Carolina State University.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10*(4), 637-666. doi: 10.2307/1267450

Cochran, W. G. (1972). Some effects of errors of measurement on linear regression. In L. M. Le Cam, J. Neyman, & E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1)

(pp. 527-539). Berkeley, CA: University of California Press. Available from https://projecteuclid.org/euclid.bsmsp/1200514110

Drilling, N. R., Rodger, T., & McKinney, F. (2002). Mallard (*Anas platyrhynchos*). In A. Poole & F. Gill (Eds.), *The birds of North America online*. Retrieved from http://bna.birds.cornell.edu/bna/species/658

Durrett, R. (2010) *Probability: Theory and examples* (4th ed.). Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511779398

Fisher, R. J. (1993). Social desirability bias and the validity of indirect questioning. *Journal of Consumer Research, 20*(2), 303-315. doi: 10.1086/209351

Götmark, F. (1992). The effects of investigator disturbance on nesting birds. In D. M. Power (Ed.), *Current ornithology* (Vol. 9) (pp. 63-104). Boston, MA: Springer. doi: 10.1007/978-1-4757-9921-7_3

Green, D. P., Palmquist, B., & Schickler, E. (2004). *Partisan hearts and minds: Political parties and the social identities of voters*. New Haven, CT: Yale University Press.

Hausman, J. (2001). Mismeasured variables in econometric analysis: Problems from the right and problems from the left. *Journal of Economic Perspectives, 15*(4), 57-67. doi: 10.1257/jep.15.4.57

Heid, I. M., Kuchenhoff, H., Miles, J., Kreienbrock, L., & Wichmann, H. E. (2004). Two dimensions of measurement error: Classical and Berkson error in residential radon exposure assessment. *Journal of Exposure Analysis and Environmental Epidemiology, 14*, 365-377. doi: 10.1038/sj.jea.7500332

Hill, D. A. (1984). Laying date, clutch size and egg size of the Mallard *Anas platyrhynchos* and Tufted Duck *Aythya fuligula*. *Ibis, 126*(4), 484-495. doi: 10.1111/j.1474-919x.1984.tb02075.x

Klein, D. F., & Cleary, A. C. (1967). Platonic true scores and error in psychiatric rating scales. *Psychological Bulletin, 68*(2), 77-80. doi: 10.1037/h0024761

Kroc, E., & Zumbo, B. D. (2019). *A transdisciplinary view of measurement error models and the variations of $X = T + E$*. Manuscript under review.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley Publishing Company.

Niemi, R. G., Weisberg, H. F., & Kimball, D. C. (Eds.). (2001). *Controversies in voting behavior*. Washington, D.C.: CQ Press.

Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1-18. doi: 10.1016/0022-2496(66)90002-2

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association, 79*(387), 516-524. doi: 10.2307/2288398

Shear, B. R., & Zumbo, B. D. (2013). False positives in multiple regression: Unanticipated consequences of measurement error in the predictor variables. *Educational and Psychological Measurement, 73*(5), 733-756. doi: 10.1177/0013164413487738

Sutcliffe, J. P. (1965). A probability model for errors of classification. I. General considerations. *Psychometrika, 30*(1), 73-96. doi: 10.1007/bf02289748

Tukey, J. W. (1962). The future of data analysis. *Annals of Mathematical Statistics, 33*(1), 1-67. doi: 10.1214/aoms/1177704711

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics, 11*(3), 284-300. doi: 10.1214/aoms/1177731868

Appendix

It is sometimes possible to reparameterize a multinomial random variable as a sum of independent Bernoulli random variables. In the classic situation considered in section *Uncalibrated Measurements*, the measurement error ε is a trinomial categorical random variable (the simplest type of nontrivial multinomial), with measure specified by $\Pr(\varepsilon = -1) = a$ and $\Pr(\varepsilon = 1) = b$, for some $0 \leq a, b \leq 1$. One may reparameterize this random variable as the difference of two independent, though not necessarily identically distributed random variables, $Y_i \sim \text{Ber}(p_i)$, $i = 1, 2$ if and only if $0 \leq a \leq 1,\ 0 \leq b \leq \left(1 - \sqrt{a}\right)^2$. This region of reparameterization is illustrated in Figure 2 below.

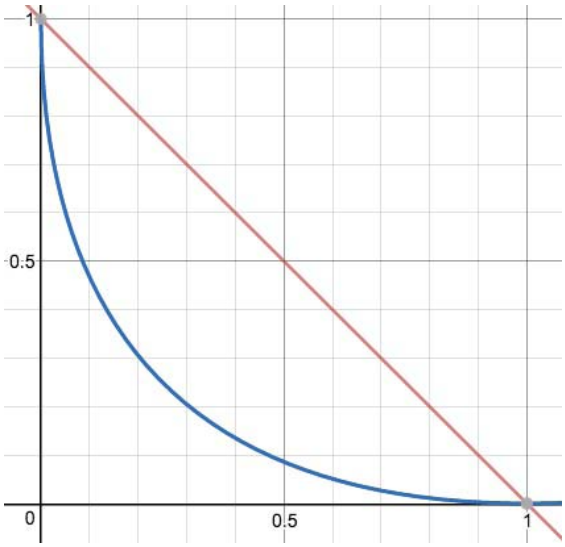

Figure 2. The trinomial categorical random error ε is parameterized by the values a, $b \in [0, 1]$, where $a + b \leq 1$, the region outlined by the right triangle defined by the coordinate axes and the red line in the figure. The region between the coordinate axes and the blue curve, $b = \left(1 - \sqrt{a}\right)^2$, illustrates where a trinomial categorical random variable can be reparameterized as a difference of independent Bernoulli random variables, $\varepsilon = Y_1 - Y_2$. Given a pair (a, b) in this closed region, one may define $p_1 = \frac{1}{2}\left(1 + b - a + \sqrt{(a - b - 1)^2 - 4b}\right)$, $p_2 = b / p_1$, with $Y_i \sim \text{Ber}(p_i)$

In some sense different from what was seen in the *The Misclassification Problem* section, this reparameterization region captures those error distributions that correspond to measurements W that are "good enough" for X. Calibrated measurements produce error distributions that fall on the diagonal $a = b$, yet a useful measurement should produce an error that is not too high in variance as well, forcing all "good" measurements to lie reasonably close to the coordinate axes in Figure 2. In this imprecise sense, the region of parameterization in Figure 2 seems to capture those measures that attain a reasonable balance between minimal miscalibration and variance.

Recall that the mean-squared error of our categorical trinomial error was given by $\mathrm{E}(\varepsilon^2) = a + b$. The level curves of this surface are lines parallel to $1 = a + b$, the red line of domain in Figure 2. Clearly then, this measure of calibration-variance tradeoff does not match up with that captured by the region of reparameterization in Figure 2.

It is currently unclear to us how this reparameterization of errors provides precise insight into the quality of a proposed measurement, even though the reparameterization seems to be capturing something in this vein. It should be noted that this region of reparameterization can be extended to a general multinomial error, though predicated upon considerably more algebraic effort. Furthermore, it should be noted that this reparameterization phenomenon appears to partition the family of multinomial distributions into two distinct classes: one where reparameterization by sums of independent Bernoulli random variables is possible, and one where this is not possible. The implications of this observation are still unknown to us. These facts are recorded here in the hope that other practitioners may be interested in elucidating them further.