Rethinking Calibration as a Statistical Estimation Problem to Improve Measurement Accuracy

Song S. Qian[a,e], Sabrina Jaffe[a], Emanuela Gionfriddo[b,d], Hongjun Wang[c], Curtis J. Richardson[c], Nipunika H. Godage[b]

[a]*Department of Environmental Sciences The University of Toledo Toledo OH*

[b]*Department of Chemistry and Biochemistry The University of Toledo Toledo Ohio 43606*

[c]*Nicholas School of the Environment Duke University Durham North Carolina 27710*

[d]*Current address: Department of Chemistry University at Buffalo The State University of New York Buffalo New York 14260-3000*

[e]*Corresponding author*

Abstract

Calibration in analytical chemistry is crucial for ensuring the accuracy and reliability of measurements. Proper calibration strategies minimize errors, enhance reproducibility, and maintain compliance with regulatory requirements. Without it, data integrity could be compromised, leading to incorrect conclusions and potentially flawed decisions in both research and industrial applications. Calibration strategies can be affected by the type of analytical instrumentation utilized as well as the time and resources available to the analyst. In this work, we reevaluated the commonly used calibration method as a statistical estimation problem to highlight the substantial uncertainty inherent in the resulting measurements and proposed a Bayesian hierarchical modeling (BHM) approach, which offers enhanced accuracy and consistency for calibration-based methods without changing the current experimental settings. Using data from three types of calibration problems, we showed that (1) the notable variability of a typical calibration-based method is due largely to the relatively limited sample size used for fitting the calibration curve, (2) the BHM approach effectively mitigated this uncertainty by pooling relevant information from multiple data points within a test and combining information from calibration curve coefficients across similar calibration curves, and (3) replications are necessary for effective estimation of measurement uncertainty. Our findings demonstrate that the accuracy and consistency of all

Email address: song.qian@utoledo.edu (Song S. Qian)

calibration-based measurement methods can be significantly enhanced by replacing the conventional regression method with the more robust BHM modeling approach.

Keywords: Bayesian statistics, calibration, ELISA, hierarchical modeling, missing data problem, shrinkage estimator

---

1. Introduction

Calibration-curve based methods are extensively used in analytical chemistry to determine analyte concentrations, constituting over 90% of chemical analytical work (Miller and Miller, 2010). This process typically involves two distinct steps: (1) developing a calibration curve through empirical modeling (regression), using known concentrations of standard solutions as the predictor and corresponding instrument-generated responses as the response variable, and (2) estimating the unknown analyte concentration of a sample using its instrumental response and the inverse function of the regression model from step (1). Although instrumental responses of calibration standard solutions and samples with unknown concentrations are often measured simultaneously, these two steps are independent. The accuracy of estimated unknown analyte concentrations depends on the statistical characteristics of the regression model.

Typically, calibration processes rely on goodness-of-fit statistics (e.g., coefficient of variation or R^2-value for linear regression) as a summary of the regression model's "quality," while predictive accuracy statistics are rarely reported. However, because these statistics measure the fit and prediction of the instrumental responses (response variable), they are not directly relevant to the classical calibration problem, which aims to estimate unknown analyte concentrations (the missing predictor variable values in the combined data from both standard solutions and calibration samples). To evaluate the uncertainty of the estimate, statistical characteristics of the classical calibration estimator (inverse function of the regression model) must be derived. Such characteristics are represented by the sampling distribution of the estimator (DeGroot, 1986). An estimate is a statistic of the data. Because data are random samples from a probability distribution (e.g., y_i in equation (1) are random samples from a normal

distribution), the statistic (estimate) is also a random variable and its probability distribution is called the sampling distribution. Intuitively, a sampling distribution describes the distribution of repeatedly estimated concentration values obtained following the same sample collection and data analysis procedure. It describes the statistical characteristics (uncertainty) of an estimator, often represented using the confidence interval.

The statistical concept of uncertainty of an estimator includes bias and precision. Bias is the difference between the mean of the sampling distribution and the (unknown) true value, and precision is measured by the variance of the sampling distribution. When the sampling distribution mean coincides with the true value, the estimator is "unbiased." The variance of an unbiased estimator is often called random error in analytical chemistry and the bias is known as the systematic error (Miller and Miller, 2010). Common statistical methods used in analytical chemistry produce unbiased estimators. Unbiased estimators often have high variances and the estimated standard curve parameters and analyte concentrations can be highly sensitive to small variability in the data, especially when the sample size used for developing the calibration standard curve is small.

The theoretical sampling distribution for the calibration estimator (the inverse function of the standard curve) is only available when the calibration curve is linear, a Cauchy distribution for which the mean and variance do not exist (Osborne, 1991). As such, quantifying the uncertainty of the estimated concentration is contentious to statisticians and almost never properly documented in practice (Mandel, 1958; Osborne, 1991). Nevertheless, the calibration estimator is often highly variable, especially when a limited number of standard calibration levels are used (Qian et al., 2015; Qian, 2016).

Because we do not use the mean of repeatedly measured concentrations in practice, the error of a specific estimate must be judged by its accuracy (the difference between the specific estimate and the true value). An unbiased estimator with a large variance can result in individual estimates with large deviations from the true value (low accuracy overall), whereas a (slightly) biased but precise estimator can result in highly accurate estimates (Figure 1), a consequence of the basic statistical trade-off between bias and variance.

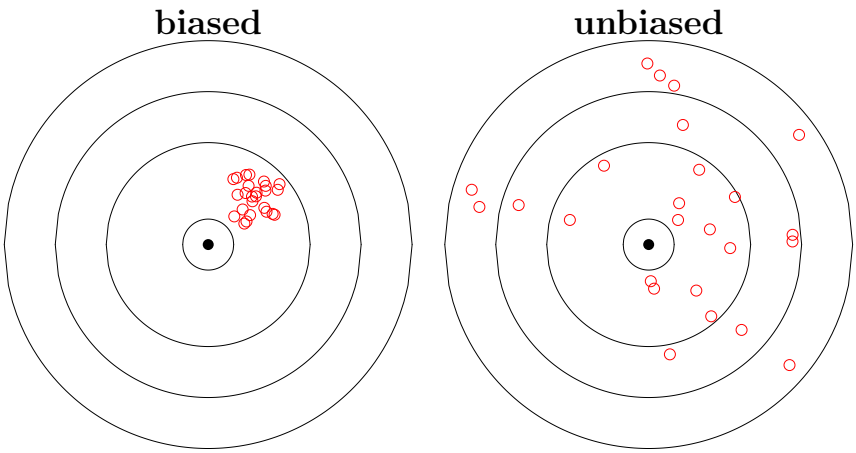

Figure 1: The bias-variance trade-off – a biased estimator with a low variance (left panel) can often out-perform an unbiased estimator because of the foundamental bias-variance trade-off in statistics.

Statisticians have long learned that "certain deliberately induced biases can drastically improve estimation properties when there are several parameters to be estimated" simultaneously (Efron, 1975). Such bias is usually induced by incorporating relevant information outside of the data (i.e., the data from the standard solutions) used for estimating the parameters of interest. For this reason, we find the Bayesian method appealing, because we can incorporate relevant information beyond the data to formulate a proper informative prior distribution to improve estimation accuracy (Efron and Morris, 1975). Furthermore, the difficulty of quantifying uncertainty in the classical calibration problem can be circumvented by using a Bayesian method for estimating the posterior distribution of unknown analyte concentrations (DeGroot, 1986). An early Bayesian calibration model for a linear standard curve (Hunter and Lamboy, 1981) illustrated computational difficulties of using the Bayesian statistics. This application used non-informative priors because of the computational difficulties, hence did not leverage additional information beyond the data. With the advent of modern simulation-based computational methods, specifically, Markov chain Monte Carlo simulation (MCMC) (Gilks et al., 1996), computational difficulties are no longer a hindrance. We now need to identify sources of relevant information to help reducing estimation uncertainty.

In the context of a calibration problem, two pertinent sources of information can be

utilized to improve estimation accuracy. Firstly, the samples needed to build the calibration curve, quality assurance (QA) samples with a known amount of analyte, and test samples with unknown amounts of analyte are usually analyzed in the same batch. In other words, there are multiple unknown analyte concentrations to be estimated in a single batch. The situation aligns with Stein's paradox (Efron and Morris, 1977), which suggests that biased estimators, such as the classical James-Stein estimator (Stein, 1956; James and Stein, 1961), empirical Bayes methods (Efron, 1996), and the Bayesian hierarchical modeling (BHM) approach (Gelman et al., 2014), can derive relevant priors for the multiple unknown analyte concentrations thereby improving the overall estimation accuracy when estimating more than three variables simultaneously. Secondly, laboratories conducting routine analyses accumulate multiple calibration curves over time. Through the use of BHM, we can achieve a similar improvement in estimation accuracy for the calibration curve coefficients (Gelman et al., 2004). In both the classical and Bayesian methods, the approach involves shrinking individually estimated (and unbiased) quantities (multiple unknown concentrations within a calibration test and calibration curve coefficients among multiple tests) towards their corresponding overall averages.

The reason of this improvement in estimation accuracy can be understood intuitively as follows: when we empirically estimate a quantity (with inherent error), the estimate may be either overestimated or underestimated compared to the underlying true quantity. In isolation, we are unsure of the direction of the estimation error, making an unbiased estimator desirable. However, when there are multiple estimates of the same quantity from, for example, parallel or inter-laboratory studies, the overall average of these estimates serves as a reference to determine whether an estimate is likely too high or too low. Consequently, shrinking the estimates towards their overall mean is advantageous in reducing overall estimation uncertainty, as demonstrated by a mathematical proof (Efron and Morris, 1975). The James-Stein estimator and BHM calculate the levels of shrinkage based on the strength of the information derived from the relative magnitudes of the estimation uncertainty of individual estimates compared to the spread among individual estimates. Efron and Morris (1977) presented a popular science introduction of the Stein's paradox.

In this study, we demonstrate the benefit of using the BHM approach for improving the measurement accuracy of calibration methods. The BHM approach allows effective sharing of information within a single calibration test and across multiple tests and has been shown to reduce estimation and prediction uncertainty in many statistical estimation problems (Gelman et al., 2014). Through three examples representing three types of calibration problems we illustrate (1) the methodological details of applying BHM to calibration problems, (2) how sample size used to develop the calibration curve affects estimation accuracy, and (3) why replication samples are necessary for quantifying estimation uncertainty in a traditional calibration method.

2. Methods

We first discuss the calibration problem as a statistical problem of estimating missing data and the difficulty in quantifying the estimation uncertainty. We propose to use a Bayesian approach to the same problem such that the estimation uncertainty can be quantified. As a statistical problem, the estimation uncertainty is directly linked to the sample size. In the calibration problem, the sample size is the number of standard solutions used for fitting the calibration curve. We show that the limited sample size (determined by the current analytical infrastructure) is a significant concern for the measurement accuracy and how a modern statistical method originated in the 1960s can be used to leverage available information from the existing analytical method to improve measurement accuracy without changing current lab procedures. Statistical details relevant to the calibration problem are included in the online Supporting Information.

2.1. Calibration as a Missing Data Estimation Problem

The classical calibration problem involves a statistical estimation process that aims to estimate both the calibration curve coefficients and the unknown analyte concentrations. In this process, we measure the instrumental responses of I standard solutions $(y_i, i = 1, \cdots, I)$ and J calibration test samples with unknown analyte concentration $(y_0^j, j = 1, \cdots, J)$ together. The goal is to simultaneously estimate the calibration curve coefficients (θ) and the missing concentration values $(x_0^j, j = 1, \cdots, J)$

together. This can be formulated as a regression problem with missing (predictor variable) data, expressed by equation (1):

$$\begin{aligned} y_i &= f(x_i, \theta) + \varepsilon_i \\ y_0^j &= f(x_0^j, \theta) + \varepsilon_j \\ \varepsilon_{ij} &= N(0, \sigma^2) \end{aligned} \tag{1}$$

where, f represents the calibration model (e.g., the four-parameter logistic function of equation 7) and $x_i, i = 1, \cdots, I$ are the known standard solution concentrations. The model error term ε is assumed to follow a normal distribution with mean 0, and variance σ^2, which is a measure of the measurement error of the instrumental response.

The joint likelihood function of the unknown parameters θ, σ^2 and x_0^j is:

$$L = \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y_i - f(x_i,\theta))^2}{2\sigma^2}} \times \prod_{j=1}^{J} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(y_0^j - f(x_0^j,\theta))^2}{2\sigma^2}}, \tag{2}$$

The maximum likelihood estimator of x_0^j coincides with the two-step process of fitting the regression model to data from standard solutions and estimating calibration sample concentrations using the inverse function of the fitted calibration curve.
Mathematically, the inverse function method is expressed as:

$$\hat{x}_0^j = f^{-1}(y_0^j, \hat{\theta}). \tag{3}$$

Here, $\hat{x}_0^j$ represents the estimated concentration of the jth calibration sample. To quantify the uncertainty of the estimated concentrations (x_0^j), we must derive the sampling distribution of $\hat{x}_0^j$, from which we derive the confidence interval as the measure of estimation uncertainty. However, the theoretical sampling distribution exists only for a linear calibration problem, which is a Cauchy distribution for which the mean and variance do not exist. This problem led us to the Bayesian approach using MCMC. For comparison to the Bayesian approach, we use the Monte Carlo simulation method described in Qian (2016) to approximate the sampling distribution.

2.2. Bayesian Calibration

Under the Bayesian framework, we estimate the posterior distribution of all unknown quantities, which is

$$\pi(\theta, \sigma^2, x_0^j \mid \text{data}) \propto \pi(\theta, \sigma^2, x_0^j) \times L \tag{4}$$

where $\pi(\theta, \sigma^2, x_0^j \mid \text{data})$ represents the posterior distribution of all unknown parameters, L is the likelihood function (equation (2)), and $\pi(\theta, \sigma^2, x_0^j)$ is the prior distribution.

When using non-informative (or flat) priors (i.e., $\pi(\theta, \sigma^2, x_0^j) \propto 1$), the Bayesian posterior distribution is essentially the normalized likelihood function. (For computational stability, we used the default weakly informative prior (Gelman et al., 2008) implemented in the software `Stan`.) Although an analytic solution of equation (4) can be challenging even for the simplest linear calibration problem (Hunter and Lamboy, 1981), the Markov chain Monte Carlo (MCMC) simulation-based computation method can be readily used for estimating the marginal posterior distributions of x_0^j, θ, and σ^2 (see Qian et al. (2022) for details).

2.3. Bayesian Hierarchical Modeling

Because of the presence of multiple test samples with unknown analyte concentrations and multiple calibration tests over time, we use a BHM approach to reduce estimation uncertainty at two levels. The statistical basis of this improvement is the Stein's paradox (Stein, 1956) from the classical statistics and the empirical Bayes approach (Robbins, 1956) from Bayesian statistics. Stein's paradox showed that when there are multiple parameters (e.g., multiple unknown analyte concentrations) to be estimated simultaneously, adjusting the individually estimated (unbiased) estimates towards the overall mean can improve the overall estimation accuracy. This adjustment is known in statistics literature as shrinking. That is, shrinking the range of the multiple (unbiased) estimates. Mathematical theory (Efron, 1975) suggests that unbiased estimates are more likely to be far away from the overall mean. As a result, shrinking unbiased estimates towards the overall mean will always improve the overall estimation accuracy. In other words, estimation methods that result in a shrinkage effect (shrinkage estimators) are always better overall than their unbiased counterparts. The empirical Bayes approach achieves similar shrinkage effect by imposing a common prior on the multiple parameters and estimating the (common) prior distribution parameter(s) from the data.

Within a single calibration test, we normally have multiple calibration samples with unknown concentrations. Accordingly, we impose a common prior distribution for all

unknown concentrations to reflect our knowledge that these concentrations are likely different from each other, while acknowledging our ignorance of their relative magnitude:

$$\log(x_0^j) \sim N\left(\mu_{x_0}, \sigma_{x_0}^2\right) \tag{5}$$

The log transformation of x_0 is used because concentration variables can usually be approximated by the log-normal distribution (Ott, 1995). The prior distribution $N\left(\mu_{x_0}, \sigma_{x_0}^2\right)$ is known as the hyper-distribution and its parameters (μ_{x_0} and $\sigma_{x_0}^2$) are hyper-parameters, with μ_{x_0} as the estimate of the overall average concentration of all samples. With the common prior and data from multiple samples, we can estimate the hyper-parameters without additional information. In other words, the common prior distribution is estimated from the data (hence the early name "empirical Bayes"). The estimated overall average is a result of having multiple samples in the same calibration test, which represents information gained from combining these samples together. Equation (5) facilitates information sharing among multiple test samples within a test.

Likewise, in cases where labs routinely repeat the same test, resulting in multiple sets of estimated calibration curve coefficients, we can use another Bayesian hierarchical framework to unite these multiple tests by analyzing data from the multiple tests together. As calibration curve coefficients are regression coefficients, we can impose the same normal prior distribution to reflect our knowledge that model coefficients vary from test to test, but we cannot foresee the relative magnitude of them among multiple tests:

$$\beta_k \sim N(\mu_\beta, \sigma_\beta^2). \tag{6}$$

Similarly, weakly informative priors can be used for μ_β and σ_β. Again, μ_β is the estimated mean of multiple calibration curve coefficients and the hierarchically estimated β_k is closer to the overall mean (μ_β) than its unbiased counterpart. Equation (6) facilitates information sharing across multiple tests.

Equations (5) and (6) can be combined in a single hierarchical model for estimating all unknown concentrations and calibration curve coefficients from multiple tests. Sharing information within a test and across multiple tests achieve "deliberately induced biases" in the estimated unknown concentrations and calibration curve coefficients.

2.4. Examples

We use three examples to illustrate the BHM approach: a nonlinear enzyme-linked immunosorbent assay (ELISA) calibration curve example to illustrate the effects of pooling information within a test and across multiple tests and the effects of sample size in calibration accuracy, a linear calibration curve example to discuss the importance of replications in evaluating the accuracy of a calibration problem, and a matrix effect example to show other relevant sources of information.

In these three examples, we have one or more quality assurance (QA) samples with known analyte concentration. We evaluate the accuracy of different methods by comparing the estimated concentrations of QA samples to their known concentrations and measure the accuracy using the absolute differences between the estimated and the known concentration values. To evaluate the uncertainty of the Bayesian methods, we use the estimated posterior distributions of the QA concentrations (represented by random samples from their posterior distributions) to derive the posterior distribution of the accuracy. For the classical (MLE) calibration method, we use the Monte Carlo simulation method for assessing regression model uncertainty discussed in Gelman and Hill (2007) and Qian (2016). These simulation methods use random samples of the estimated regression model coefficients to represent regression model uncertainty: each set of random coefficients represents a possible calibration curve, from which we estimate the QA sample concentration. As a result, uncertainties of both the Bayesian methods and the classical calibration method are represented by random numbers from their respective distributions.

2.4.1. The Nonlinear Calibration Example

The nonlinear calibration example uses data collected during the "Toledo Water Crisis" of August 1-3, 2014, when a "Do-Not-Drink" order was issued due to one microcystin (MC) measurement (2.7 μg/L) exceeded the Ohio drinking water quality standard (1 μg/L) on August 1, 2014. The measurement was made in City of Toledo's drinking water plant using a nonlinear ELISA with kits from Eurofins-Abraxis. The kit recommends the use of a sigmoid function (the four parameter logistic function) to describe the relationship between the instrumental response (y) and MC concentration

(x, equation (7)).

$$y = \theta_4 + \frac{\theta_1 - \theta_4}{1 + \left(\frac{x}{\theta_3}\right)^{\theta_2}} + \varepsilon \tag{7}$$

This example has data from six ELISA tests carried out during the crisis so that we can illustrate the computational details and make direct comparisons between the BHM estimates and the classical calibration estimates. Data are from Qian (2016).

Following the US EPA's standard protocol, the standard commercial kit from Abraxis uses six standard solutions with known MC values ranging from 0 to 5.55 μg/L with two replicates each. Each ELISA kit from Abraxis come with one QA sample with a known MC concentration of 0.75μg/L, near the low end of the standard solutions range. The resulting 12 data points are either (1) directly used for fitting the calibration curve ($n = 12$) or (2) transformed by averaging the response replicate observations for each non-zero standard solution and dividing them by the average of the zero solution replicates to generate 5 "relative" responses for fitting the calibration curve ($n = 5$).

2.4.2. The Linear Calibration Example

A calibration problem with a linear standard curve is the simplest calibration problem. We used water quality (orthophosphate, PO_4) monitoring data from the Stream and Wetland Assessment Management Park (SWAMP) at Duke University in Durham, NC. The measurement method is based on the reaction of ammonium molybdate and potassium antimonyl tartrate in acid medium with orthophosphate (PO_4). The resulting phosphomolybdic acid is then reduced to intensely colored molybdenum blue by ascorbic acid (Ascorbic Acid Method, SM 4500-PE). The intensity of the color is proportional to the PO_4 concentration in the sample. Six standard solutions with known PO_4 concentrations (5, 10, 20, 50, 100, and 200 μg/L) and two blank samples (0) are used to develop the linear calibration curve (i.e., $y = \beta_0 + \beta_1 x + \epsilon$). We used a subset of calibration tests from 2017 to 2021 in this paper.

The EPA's PO_4 measurement protocol does not require replicate samples. Without a replicate, the inverse-function estimated concentration is the value resulted in the

observed instrumental response according to the fitted calibration standard curve. If there were replicates, we would obtain multiple estimated concentration values for the same test sample to provide additional information on the estimation uncertainty with respect to analyte concentration. Without a replicate, we rely on the overly optimistic measurement uncertainty information (e.g., the estimated residual standard deviation) from the fitted standard curve (Hunter and Lamboy, 1981).

The QA sample from the Duke lab has a known concentration of 50 μg/L, which has the same concentration as one of the standard solutions used for fit the calibration curve and is very close to the mean of the standard solution concentrations (48.125 μg/L, including the two blanks), where the fitted regression model has the lowest uncertainty (Weisberg, 2005). Because a linear regression line is anchored at the point of $(\bar{x}, \bar{y})$, that is, the fitted linear model goes through the point defined by the standard solution concentration mean and the response mean, including all the lines represented by our Monte Carlo simulation. Given that no replication is used in this data set, the Monte Carlo simulation method estimated accuracy is overly optimistic. As a result, we can only compare the accuracy of the Bayesian methods.

2.4.3. Matrix Effects

Concentrations of xenobiotics in human blood are often used to measure human exposure to organic pollutants, such as pesticides and pharmaceuticals. Measuring xenobiotic concentrations can be expensive and complicated, and ultimately, the concentration in a blood sample is estimated using a calibration method. A recent study explored the use of biocompatible solid phase microextraction (SPME) coupled with liquid chromatography mass spectrometry (LCMS) for xenobiotic analysis in plasma (Godage et al., 2024).

This study emphasized the feasibility of using non-human plasma-based standard solutions to reduce costs. Due to differential binding between target analyte and different plasma, calibration curve coefficients are known to vary from plasma to plasma (the matrix effect). When the matrix effect between human plasma and non-human plasma (e.g., bovine plasma) is stable, bovine plasma-based standard solutions can be used as substitutes for human plasma standard solutions, which can greatly reduce

measurement costs. Using calibration data for acetochlor (a widely used herbicide and known human carcinogen and thyroid disruptor) as an example, we illustrate the use of BHM for improving estimation accuracy (evaluated using QA samples) and establishing matrix effects.

We use data from acetochlor, a chloroacetanilide type herbicide, to illustrate how the use of BHM improves estimation accuracy. The raw data from the experiments included results from multiple days, and only the data from the first day were used in our analysis (including the calibration standard solutions and quality control samples). There are three groups of quality assessment samples spanning across the concentration spectrum of the target analyte concentration to represent low, median, and high levels. As in the other two examples, we analyzed the data using both the conventional inverse-estimation method and the BHM method. In both methods, we used predesignated standard solutions and treated quality control sample concentration values as unknown to evaluate estimation accuracy. There are six mediums (human, bovine, rabbit, and rate plasma, and PBS solution) for each xenobiotic.

For the inverse-estimation method, we fit each of the six calibration curves independently and calculate the "unknown" concentrations using the respective resulting linear functions. The BHM method combines data from all six test mediums to estimate the quality control sample concentrations and the six calibration curve coefficients together.

The example includes (1) the determination of the calibration curve format (a log-log linear model), (2) a comparison of the estimation accuracy using three sets of quality control samples, and (3) the estimated matrix effects (differences in fitted calibration curve coefficients, intercept and slope, among five mediums: human, bovine, rabbit, and rat plasma, as well as phosphate-buffered saline, or PBS, buffer).

2.4.4. Methods Comparison

We use "uncertainty" as a general concept of the state of being unsure how close the estimated concentration is to the unknown true value. Accordingly, we measure the uncertainty using "accuracy" defined as the absolute difference between the estimated

and the true value, which can only be directly estimated when quality assurance (QA) samples with known concentrations are available. Accuracy can be attributed to both the variance and bias of an estimator (Figure 1). Both bias and variance summarize the collective behavior of the estimation method. In our case, we use random samples, either from the posterior distribution or the Monte Carlo method, to capture the behavior of the accuracy: using the median of the random samples to approximate the expected accuracy and the ranges of the middle 50% and 95% of the random samples (the 50% and 95% credible intervals) as a measure of consistency of the method.

2.4.5. The Role of Sample Size

In a regression problem, we assume that the response variable is a normal random variable with the mean defined as a function of the predictor variable (the mean function) and a constant residual variance. The model-fitting process quantifies both the mean function coefficients and the residual variance based on available data. The estimation uncertainty is largely determined by the amount of information in the data, measured by the degrees of freedom (typically, the sample size minus the number of model coefficients to be estimated). The regression model coefficient uncertainty is a function of the model's residual variance. When a regression model's degrees of freedom is four or fewer, the uncertainty of the estimated residual variance is undefined, meaning that we cannot simultaneously quantify the model coefficients and the residual variance with confidence. In other words, a regression model with a degree of freedom of 4 or fewer cannot be reliably used for prediction, although a regression model fit with a small sample size is more likely to have an impressive goodness-of-fit statistics. In the extreme case of fitting a linear model with two data points, we always obtain a perfect R^2 value of 1 and a $\hat{\sigma}^2 = 0$. However, with a degrees of freedom of 0, we have no information about the model's predictive accuracy (i.e., the resulting model is unreliable for prediction).

2.4.6. Computational Details

We implemented our methods in R (R Core Team, 2022), using the package `rv` (Kerman and Gelman, 2007) for Monte Carlo simulation, and `rstan` (Stan Development Team,

2022a) to access the Bayesian computation software `Stan` (Stan Development Team, 2022b) for our Bayesian computation. Data and commented code used for this work are available at author's GitHub repository (`GitHub.com/songsqian/calibration`).

3. Results and Discussion

3.1. The Role of Sample Size

We illustrate the role of sample size by fitting the nonlinear four-parameter logistic function (equation (7)) using data from the first ELISA test from Toledo Water Crisis. We fit the regression model using a Bayesian regression method (with non-informative priors) and illustrate the strong correlation between model coefficients and the residual variance. We first use $n = 5$ (relative response data) and then use $n = 12$.

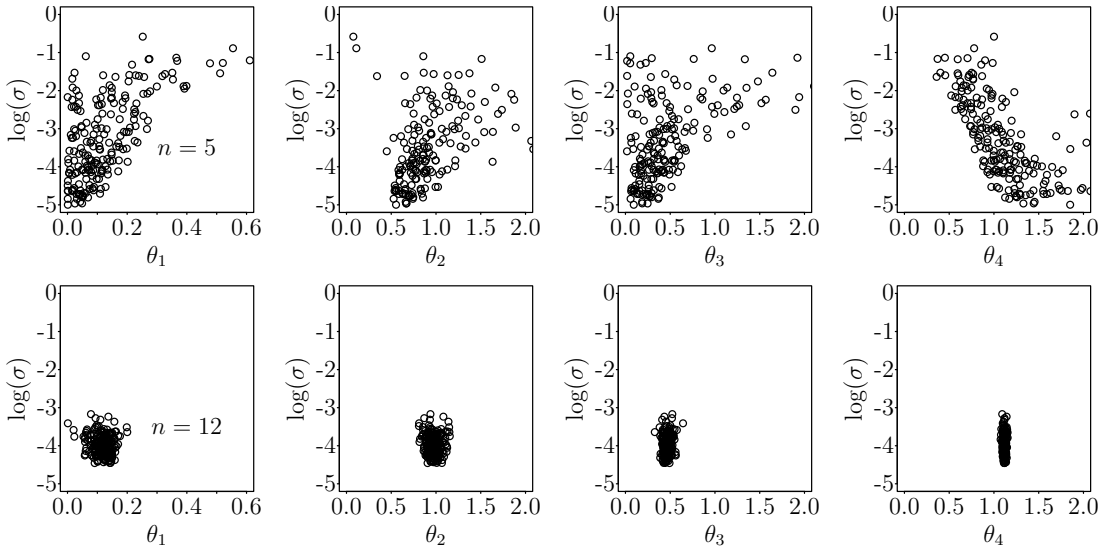

Figure 2: Uncertainty in the fitted standard curve (four-parameter logistic function) coefficients (θ_1–θ_4) and residual variance are represented by their joint posterior distribution, expressed using bi-variate scatter plots of the log standard deviation (σ in equation (1)) against coefficients using random samples from the joint posterior distribution. The standard curve fit with mean relative response ($n = 5$) has a much higher uncertainty (top row) than the curve fit with un-transformed raw data ($n = 12$, bottom row).

When fitting the model using relative responses ($n = 5$), the marginal posterior distributions of θ_i and σ^2 are funnel-shaped (Figure 2, top row): the range of θ_i is highly dependent on the value of σ^2. With $n = 5$ (or 1 degree of freedom), we have no

information to adequately quantify σ^2. As a result, we are highly uncertain about the model coefficients. When fitting the same model with $n = 12$ (8 degrees of freedom), we are able to better quantify σ^2 (Figure 2, bottom row). As a result, the estimated θ_i are more stable. In all six tests, the apparent R^2-values for the models fit with $n = 5$ are larger and the residual standard errors are smaller than the same for models fit with $n = 12$. Using the Bayesian estimation method, we can evaluate the estimation uncertainty using the posterior distribution of the unknown analyte concentration. The variances of the estimated MC concentrations (expressed as the standard deviations of posterior distributions) using the inverse-function method is orders of magnitude higher when using the standard curve fitted with $n = 5$ than the same using $n = 12$ (Figure 3). In other words, the R^2-value and residual standard deviation should not be used as the sole criterion of evaluating a calibration curve. Fitting a calibration curve using average responses of standard solution replicates is inadvisable. In the rest of the paper, we present the Toledo Water Crisis example results of the model with $n = 12$ only.

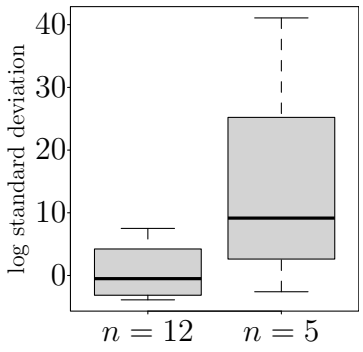

Figure 3: Estimation uncertainty (expressed as standard deviation of the estimated MC concentrations) using the inverse-function method is a function of the sample size used for fitting the standard curve. The boxplots represent the 80 estimated log standard deviations from the 40 water samples (each with a replicate) in one of the 6 ELISA tests conducted during the Toledo Water Crisis.

3.2. Hierarchical Modeling Within and Across Multiple Calibration Curves

We apply the BHM approach on two levels. One is within each test, by considering all unknown analyte concentrations as exchangeable (i.e., they are different but otherwise uncertain) and imposing a common prior distribution. That is, we fit the calibration

curve and estimate unknown concentrations one test at a time. The other is across all tests by considering the standard curve coefficients as exchangeable and imposing a common prior across the 6 tests. As a result, the six calibration curves are fitted and all unknown concentrations are estimated together.

In the six Toledo Water Crisis ELISA tests, we have, for each test, 11 to 40 test samples with unknown MC concentrations and one QA sample with known MC concentration of 0.75 μg/L. We compare the estimated MC concentrations of the QA sample from using the inverse-function method, and the two BHM approaches (Figure 4). The estimation uncertainty of the inverse-function based method is estimated based on a Monte Carlo simulation where random samples of the standard curve coefficients were used to represent the model uncertainty. Some random samples of the coefficients result in non-real value solutions (log of negative values) and must be excluded. As a result, uncertainty of the inverse-function method in Figures 4 and 5 is underestimated.

We note that the inverse-function method estimated QA sample concentrations vary more from test to test compared to the estimates from the BHM estimates, also reflected in the comparison of the estimated accuracy among different methods (Figure 5).

For the two linear calibration examples from SWAMP monitoring data, we only compare the three Bayesian methods: Bayesian without using the hierarchical modeling approach (Bayes), BHM within test only (BHM_1), and BHM within and across tests (BHM_2) (Figures 6). As with the ELISA data, BHM within a test resulted in a substantial reduction in bias. The addition of cross-test BHM produced marginal improvement.

For the mass-spectromertry data, we only show the comparison of BHM within and across the calibration curves and the classical model averaged over the estimates from 6 different mediums. Because the calibration curve is fit on log-concentration scale, the absolute difference between the estimated and true log QA sample concentration is the log ratio of the estimated over the true. As such the absolute value of 1 minus the ratio is the accuracy in percentage (Figure 7). The accuracy in this case can be measured as within a certain percentage of the true concentration.

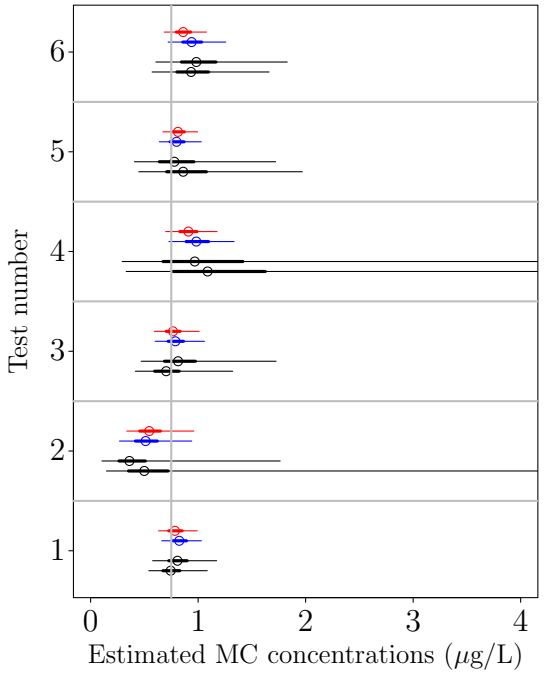

Figure 4: The estimated QA sample concentrations (two replicates) using inverse-function method (black), BHM within test (blue), and BHM within and across test (red) are compared. The open circles are the estimated means, the thick and thin lines are the 50% and 95% credible intervals, respectively. The true QA sample concentration (0.75 μg/L) is shown by the gray vertical line. The inverse-function method produced two estimates, one for each replicate. Each Bayesian method produced one estimate for each QA sample.

Although pooling information within a test is unquestionably justifiable, pooling information across tests requires fitting multiple calibration curves together, which is impractical. In a separate study, Jaffe et al. (2024) discussed the use of a sequential updating algorithm to implement the cross-test BHM. The algorithm is based on the definition of the hyper-distribution of calibration coefficients (equation (6)), which serves as the common prior for coefficients of individual calibration curves. As such, when the hyper-distribution is estimated from existing data, it can be considered as the prior for the coefficients of the next calibration test (Efron, 1996). Consequently, we can update the hyper-distribution one test at a time sequentially. The sequential updating algorithm can be automated by incorporating it into a computer application, thereby replacing the currently used spreadsheet-based applications.

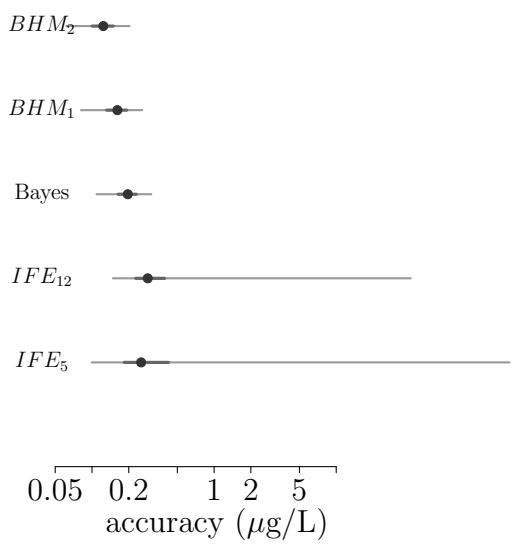

Figure 5: ELISA estimation accuracy (mean absolute errors of the 6 ELISA tests) of the inverse-function estimators (IFE_5 and IFE_{12}), Bayesian method with flat prior (Bayes), BHM within test (BHM_1), and BHM within and across tests (BHM_2) are compared to the known control sample concentration of 0.75 μg/L. The black dots are the estimated means, the thick and thin lines are the 50% and 95% credible intervals, respectively. IFE_5 is for reference only. Its 95% credible interval range was greatly underestimated.

3.3. Matrix Effect

In the xenobiotics measurement example, Godage et al. (2024) first explored whether the linear calibration curve should be fit in the original scale or the (natural) log-log scale. Graphical display (scatter plots) of the measured response and known concentration values indicate that the calibration curve is better defined in the log-log scale (Figure 8). A log-log linear relationship of $\log(y) = \beta_0 + \beta_1 \log(x) + \epsilon$ indicates a proportional relationship between x and y, that is, for every 1% increase in x, y would increase by β_1% (Qian, 2016). In the original scale, the log-log model is a power model $y = \alpha_0 x^{\beta_1} \times e^{\epsilon}$ (where $\alpha_0 = e^{\beta_0}$). We can easily judge whether the response is proportional to the concentration by comparing the estimated β_1 to 1. Based on Figure 8, we used the log-log model.

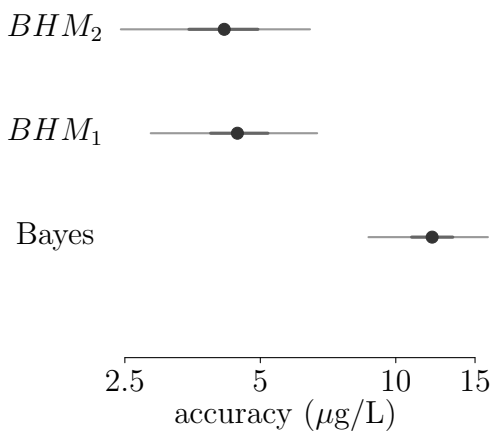

Figure 6: Estimation accuracy of the linear calibration example are compared among the Bayesian estimators: Bayesian method with flat prior (Bayes), BHM within test (BHM_1), and BHM within and across tests (BHM_2). The known QA sample concentration is 50 μg/L [total phosphorus]

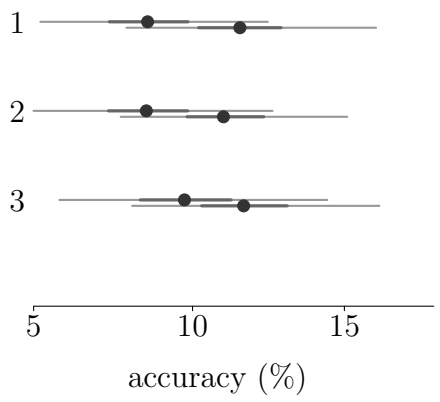

Figure 7: Estimation accuracy, measured by the relative biases are compared between the inverse-estimation method (bottom line of each pair) and the BHM method (top line of each pair). The three pairs of lines represent (from top to bottom) quality control sample concentrations at the low (2.5-5 μg/L), median (10-25 μg/L), and high (50-75 μg/L) groups. A bias of 5%, for example, represents that, in the concentration scale, the estimated concentration is within 5% of the true concentration.

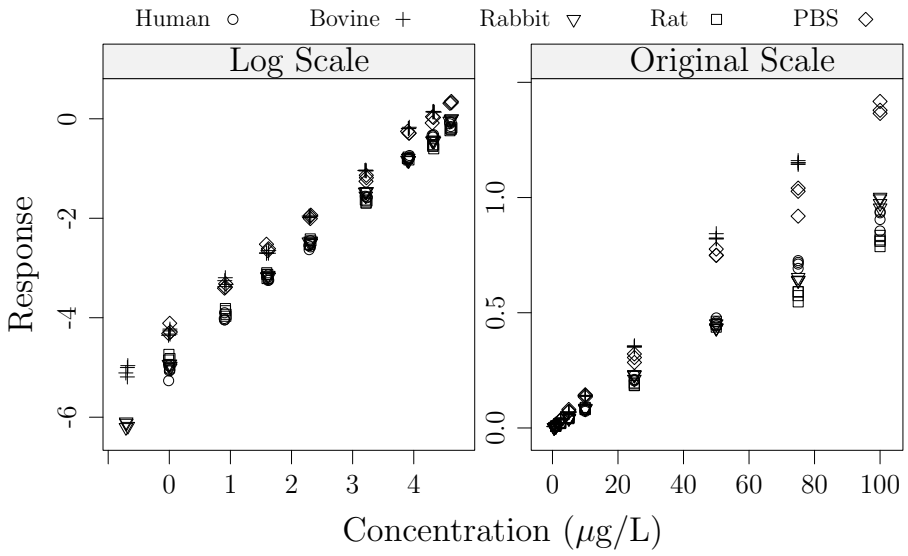

Figure 8: Scatter plot of the SPME-LCMS data for acetochlor (the response, y-axis, is the area of the analyte). In the original scale (right panel), the difference among different mediums lies in the slope and the intercepts nearly the same (~ 0), whereas in the log-log scale (left panel) the difference lies in the intercept and the slopes are nearly the same (~ 1). The response is analyte area and the concentration is in μg/L.

The advantages of BHM (or multilevel model) for estimating multiple (related) regression model coefficients is well-documented in the literature (e.g., Gelman and Hill (2007); Qian et al. (2010)). In our analysis we find that the estimated regression model coefficients for the six mediums are independent of each other with low estimation uncertainty (Figure 9). As a result, using animal plasma (or PBS) as standard solution medium is feasible.

4. Conclusions

We identified an inherent statistical weakness of the calibration-based measurement methods, that leads to high levels of uncertainty in resulting measurements. As these methods are used for quantifying concentrations of nearly all substances and affecting all aspects of our daily life, the importance of our study cannot be overstated. Our study discussed that the underlying statistical reason for the method's high level of estimation uncertainty is the small sample size (usually limited to around 6 standard solutions) used in fitting the calibration curve.

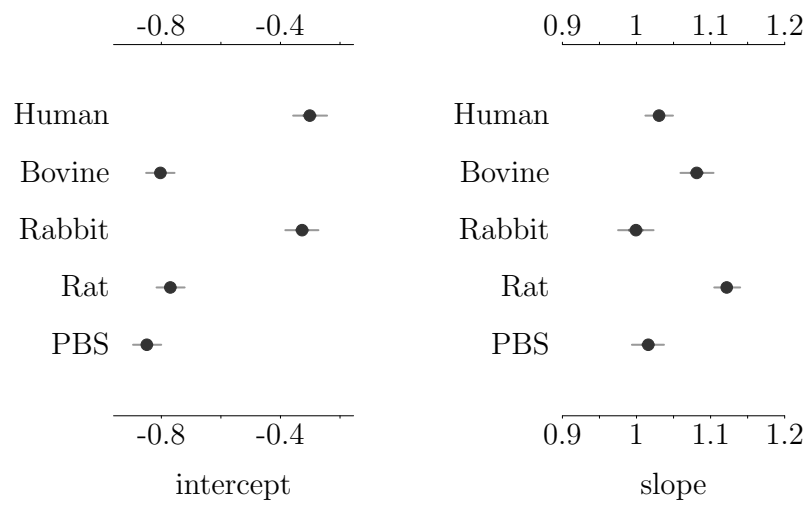

Figure 9: BHM estimated calibration curve coefficients are of low estimation uncertainty, indicating that using non-human plasma as substitution for human plasma standard solution is feasible.

Our study recommends the use of Bayesian hierarchical modeling (BHM) approach to improve measurement accuracy by leveraging relevant information already available in a typical chemical laboratory. In all three examples, the BHM method consistently outperform the conventional inverse-function method in terms of measurement accuracy, measured by the absolute difference between the estimated concentration values and the known (or true) concentration values, the pertinent measure for calibration problems.

Although the BHM approach presented in this paper is not new (both the frequentist and Bayesian statistical theories showed the benefit of sharing information among relevant sources in improving estimation accuracy (Stein, 1956; James and Stein, 1961; Robbins, 1956)), our contributions lie in three areas. First, we illustrated the accuracy problem of calibration methods, a problem which is often oversimplified and approximated in practice based on overly optimistic uncertainty information from the fitted standard curve. Second, we demonstrated the effectiveness of the BHM approach in improving the accuracy. In a separate paper Jaffe et al (Jaffe et al., 2024) introduced our third contribution, which made this approach easily applicable as discussed in the previous section.

We envision that a web-browser-based application (e.g., one based on the R Shiny package (Chang et al., 2023)) would be developed to incorporate our sequential

updating algorithm for each laboratory to replace the currently used custom spreadsheet-based software. Such a browser-based app would necessarily be lab- and analyte-specific, as is the currently used spreadsheet-based software, to accumulate the updated hyper-distributions of calibration curve coefficients.

Acknowledgement

This work was partially supported by the Ohio Sea Grant small Grant Program (SSQ and SJ), the University of Toledo's summer research initiative program (SSQ and EG), the University of Toledo start-up funds (EG), and Perkin Elmer for enabling the use of the QSight 220 for this work (EG).

Declaration

- Authors declare no conflict of interest/competing interests.
- Authors' contribution
  - SSQ – Conceptualization, study and data analysis design, drafting manuscript, securing funding
  - SJ – Data analysis, editing and reviewing manuscript
  - EG – Study design, data curation, editing/reviewing manuscript, securing funding
  - HW & CJR – data curation, reviewing manuscript
  - NHG – data curation

5. Supporting Information

In a GitHub repository `github.com/songsqian/Calibration`, we included:

- a document of statistical basis of our manuscript, which includes a discussion on the difference between the fitted and predictive uncertainty of a regression model, statistical details of treating a calibration problem as a statistical estimation of

missing data problem, the use of Monte Carlo simulation to evaluate a linear or nonlinear regression model, and the trade-off between an unbiased and a biased estimator;

- Source data used in the paper; and
- Computer code with documentation.

All analyses were conducted in R. The raw data in this repository were kept unaltered as we received them. Any data manipulation we performed are recorded in the R code. The matrix effect example data spreadsheets included fitted standard curves. They were not used in our analysis. The Toledo water crisis data were originally published by the City of Toledo in the form of a PDF file of photo-copied lab spreadsheet printouts. The data we used were manually digitized by Qian (2016). The PO4 monitoring data spreadsheet was provided by the Duke University Wetland Center.

References

W. Chang, J. Cheng, J.J. Allaire, C. Sievert, B. Schloerke, Y. Xie, J. Allen, J. McPherson, A. Dipert, and B. Borges. *shiny: Web Application Framework for R*, 2023. https://shiny.posit.co/, https://github.com/rstudio/shiny.

M.H. DeGroot. *Probability and Statistics*. Addison-Wesley Publishing Company, Inc., second edition, 1986.

B. Efron. Biased versus unbiased estimation. *Advances in Mathematics*, 16:259–277, 1975.

B. Efron. Empirical Bayes methods for combining likelihoods. *Journal of the American Statistical Association*, 91(434):538–550, 1996.

B Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association*, 70(350):311–319, 1975.

B. Efron and C. Morris. Stein's paradox in statistics. *Scientific American*, 236:119–127, 1977.

A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, New York, 2007.

A Gelman, G.L. Chew, and M. Shnaidman. Bayesian analysis of serial dilution assays. *Biometrics*, 60:407–417, 2004.

A. Gelman, A. Jakulin, M.G. Pittau, and Y.S. Su. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, 2(4):1360–1383, 2008.

A. Gelman, J.B. Carlin, H.S. Stern, David B. Dunson, Aki Vehtari, and D.B. Rubin. *Bayesian Data Analysis*. CRC Press, Boca Raton, Florida, 3rd edition, 2014.

W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London. UK, 1996.

Nipunika H. Godage, Song S. Qian, Erasmus Cudjoe, and Emanuela Gionfriddo. Enhancing quantitative analysis of xenobiotics in blood plasma through cross-matrix calibration and bayesian hierarchical modeling. *ACS Measurement Science Au*, 4(1): 127–135, 2024.

W.G. Hunter and W.F. Lamboy. A Bayesian analysis of the linear calibration problem. *Technometrics*, 23(4):323–328, 1981.

S. Jaffe, D. Gossiaux, R.M. Errera, E. Gionfriddo, and S.S. Qian. A Bayesian hierarchical modeling approach for improving measurement accuracy of microcystin concentrations. *Submitted*, in review:Available at `github.com/songsqian/calibration`, 2024.

W. James and C. Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium Mathematics, Statistics and Probability*, volume 1, pages 361–379. University of California Press, Berkeley, California, 1961.

J. Kerman and A. Gelman. Manipulating and summarizing posterior simulations using random variable objects. *Statistics and Computing*, 17(3):235–244, 2007.

J. Mandel. A note on confidence intervals in regression problems. *The Annals of Mathematical Statistics*, 29(3):903–907, 1958.

J.N. Miller and J.C. Miller. *Statistics and Chemometrics for Analytical Chemistry*. Pearson/Prentice Hall, Essex, UK, 6th edition, 2010.

C. Osborne. Statistical calibration: A review. *International Statistical Review*, 59(3): 309–336, 1991.

W.R. Ott. *Environmental Statistics and Data Analysis*. Lewis Publishers, Boca Raton, Florida, 1995.

S.S. Qian. *Environmental and Ecological Statistics with R*. Chapman and Hall/CRC Press, Boca Raton, Florida, 2nd edition, 2016.

S.S. Qian, T.F. Cuffney, I. Alameddine, G. McMahon, and K.H. Reckhow. On the application of multilevel modeling in environmental and ecological studies. *Ecology*, 91:355–361, 2010.

S.S. Qian, J.D. Chaffin, M.R. DuFour, J.J. Sherman, P.C. Golnick, C.D. Collier, S.A. Nummer, and M.G. Margida. Quantifying and reducing uncertainty in estimated microcystin concentrations from the ELISA method. *Environmental Science and Technology*, 49(24):14221–14229, 2015.

S.S. Qian, M.R. DuFour, and I. Alameddine. *Bayesian Applications in Environmental and Ecological Studies with R and Stan*. Chapman & Hall/CRC Applied Environmental Statistics. CRC Press, Boca Raton, FL, 2022. ISBN 9781351018760.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022. URL `https://www.R-project.org/`.

H. Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163. The University of California Press, Berkeley, California, 1956.

Stan Development Team. RStan: the R interface to Stan, version 2.21.5, 2022a. URL `http://mc-stan.org/rstan.html`.

Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 2.30*, 2022b. URL `http://mc-stan.org/manual.html`.

C. Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 197–206, Berkeley, California, 1956. University of California Press.

S. Weisberg. *Applied Linear Regression*. Wiley, New York, 3rd edition, 2005.