NUCLEAR SCIENCE AND ENGINEERING: **173**, 15–27 (2013)

Meta-Analysis Options for Inconsistent Nuclear Measurements

Tom Burr,* Brian Williams, Stephen Croft,
Morgan White, and Ken Hanson

Los Alamos National Laboratory
Los Alamos, New Mexico 87545

Received December 9, 2011
Accepted May 7, 2012

Abstract—*Meta-analysis aims to combine results from multiple experiments. For example, a neutron reaction rate or cross section is typically measured in multiple experiments, and a single estimate and its uncertainty are provided for users of the estimated reaction rate. It is often difficult to combine estimates from multiple laboratories because there can be important differences in experimental protocols among laboratories and because laboratories do not always provide all the information needed to assess the estimate's uncertainty, particularly if total uncertainty (random and systematic) is required. The paper illustrates that explicit measurement error models are essential for understanding measurement processes and for guiding how to combine multiple measurements, whether the measurements are consistent or not. We emphasize that both the consensus estimate and its estimated uncertainty depend on the assumed measurement error model, and we investigate measurement error model selection options for two examples.*

I. INTRODUCTION

This paper considers options to combine estimates of the same measurand from multiple laboratories to reach a consensus value. Individual laboratory estimates are sometimes inconsistent in the sense that their differences are larger than their individual uncertainty estimates suggest.[1] Such inconsistencies can arise from important differences in experimental protocols among laboratories, making it difficult to combine multiple estimates across experiments.

Laboratories do not always provide all the information needed to assess an estimate's uncertainty, particularly if total uncertainty (random and systematic) is required. However, one purpose of a meta-analysis is to discover inconsistencies that could result in more tightly controlled assay protocols and therefore more consistent assay results. We emphasize that both the consensus estimate and its estimated uncertainty depend strongly on the assumed measurement error model, and measurement error model selection options are investigated.

The following sections include a background discussion for multilaboratory experiments, statistical modeling, model selection, two examples, and suggestions for future directions. The examples arise from experiments to measure nuclear reaction rates and illustrate to what extent inconsistencies across laboratories can be detected as a function of the number of laboratories and the size of the inconsistencies relative to inevitable random assay errors.

II. BACKGROUND

Measurement error models are essential for understanding measurement processes and for guiding how to combine multiple measurements, whether the measurements are consistent or not. This background section introduces three relatively simple measurement error models used throughout. We use the term "measurement error" to mean a random variable whose probability distribution characterizes assay errors. Following convention, we sometimes qualitatively refer to the standard deviation of the measurement error distribution as its uncertainty. Alternatively, the qualitative term "uncertainty" is sometimes used to describe the confidence interval (CI) associated with an estimate.

*E-mail: tburr@lanl.gov

II.A. Model 1: Simplest Possible Measurement Error Model

One of the simplest possible measurement error models (model 1) from laboratory i for estimating a measurand in a multilaboratory experiment from n laboratories is

$$M_{ij} = T + R_{ij} \ , \tag{1}$$

where

$M_{ij} = j$'th measured value from laboratory i

T = true value

R_{ij} = random error of the j'th replicate from laboratory i, which can often be well modeled as having a normal (Gaussian) distribution, denoted as $R_i \sim N(0, \sigma_{R_i}^2)$, where $N(\mu, \sigma^2)$ is the Gaussian distribution with mean μ and variance σ^2 (Refs. 1 through 5).

Although Eq. (1) is sometimes adequate and sometimes each laboratory provides a high-quality and accurate estimate of its random error variance $\sigma_{R_i}^2$, this paper's focus is the situation in which the laboratory values are not consistent ("inconsistent" is defined in Sec. V.B), and so alternatives to Eq. (1) are necessary. As an example alternative to Eq. (1), each laboratory could have a laboratory-specific systematic error, as described next.

II.B. Model 2: Laboratory-Specific Systematic Errors

One alternative to Eq. (1), which we refer to as model 2, is to allow for laboratory-specific systematic errors,[2,5] arising, for example, from laboratory-specific estimation of detector efficiency. If laboratory-specific systematic errors can occur, than a more reasonable measurement error model from laboratory i for estimating a measurand in a multilaboratory experiment can be expressed as

$$M_{ij} = T + S_i + R_{ij} \ , \tag{2}$$

with all terms the same as in Eq. (1), except S_i is added, where S_i is the unknown laboratory-specific systematic error (bias) of laboratory i. One typically assumes and can defend that errors (even laboratory-specific systematic errors) are random at some stage (such as being random across hypothetical or real recalculations of detector efficiency), which we denote as $S_i \sim N(0, \sigma_{S_{lab}}^2)$. Whether the assumption of a Gaussian distribution is important depends on the goals and context. The most common model assumes measurements have nonzero covariance if and only if they are made by the same laboratory during the same experiment, so S_i is the same for two measurements of the same measurand by the same laboratory.

II.C. Model 3: Some Subset of Laboratories Has Larger-Than-Estimated Random Error Variance

As another alternative to Eq. (1), which we refer to as model 3, Hanson[6] proposed that some subset of the laboratories is optimistic and that in fact for some of the laboratories, the true random error standard deviation is $\gamma\sigma_{R_i}$, where $\gamma > 1$ and γ is the same value for all the optimistic laboratories. This results in model 3 given by a mixture of two Gaussians for the laboratory result m for a given true value T as

$$f_2(m|T) = (1-\beta) f_1(m|T, \sigma_R, \gamma = 1) + \beta f_1(m|T, \sigma_R, \gamma) \ , \tag{3}$$

where the mixing fraction β satisfies $0 \le \beta \le 1$, the scaling factor $\gamma > 1$, and f_1 is the Gaussian distribution with standard deviation $\gamma\sigma_R$. Implied by Eq. (3) is the fact that prior to observing data from all n laboratories, there is no ability to infer which laboratories are optimistic, only that a fraction β of the laboratories is optimistic.

Cacuci and Ionescu-Bujor[1] considered a variation of Eq. (3) in which some or all of the n laboratories have "unrecognized errors," leading to larger-than-reported within-laboratory random error variance. Using our notation for consistency with this paper, their model is $M_{ij} = T + R_{A_{ij}} + R_{B_{ij}}$, where R_A is recognized by the laboratory and R_B is not. Cacuci and Ionescu-Bujor[1] assumed it would be known which laboratories have and which laboratories do not have unrecognized errors and also assumed that the ratio of the variances $\sigma_{R_{B_{i1}}}^2 / \sigma_{R_{B_{i2}}}^2$ of the unrecognized errors was known for every nonredundant pair (i_1, i_2) of laboratories, introducing a single unknown scaling factor s that determines the sum of the variances $\sum_{i=1}^{n} \sigma_{R_{B_i}}^2$ of the unrecognized errors R_{B_i}. In examples 1 and 2 in Sec. V, there is only one measurement per laboratory, although each measurement could be an average of several replicates. Therefore, the error model in Ref. 1 is similar to our Eq. (2) except that Ref. 1 assumes only random errors, not laboratory-specific systematic errors S_i. And, our Eq. (2) assumes that the variance of S_i is the same for all laboratories, denoting that common variance as $\text{var}(S) = \sigma_{S_{lab}}^2$. Analytical expressions for least informative priors and corresponding Bayesian posterior distributions are then provided by Ref. 1.

As shown in Sec. V, our emphasis is different from that of Ref. 1 in that we assess to what extent the $M_1, M_2, \ldots, M_n$ values can indicate whether model 1, model 2, or model 3 is more appropriate, and we use a numerical implementation of a Bayesian[7] approach described below (rather than analytical calculation) that is available in open source software. For comparison, we also provide a maximum likelihood (ML) approach.

II.D. Variance Propagation Implications for Models 1, 2, and 3

Suppose we want to combine the results of n_1 measurements of the same measurand from laboratory 1 and n_2 measurements from laboratory 2 to estimate T. Using Eq. (2), we obtain the true variance of an unweighted average (see Sec. V for weighted averages) of n_1 measurements of the same measurand from laboratory 1 and n_2 measurements from laboratory 2 as

$$\sigma^2_{(\overline{M}_1+\overline{M}_2)/2} = (1/4)\,\mathrm{var}(\{n_1 T + n_1 S_{lab1} + R_{1,1} + R_{1,2} + \cdots + R_{1,n_1}\}/n_1 + \{n_2 T + n_2 S_{lab2} + R_{2,1} + \cdots + R_{2,n_2}\}/n_2) \ , \tag{4}$$

where var( ) is the variance of the quantity in ( ), which reduces to $\sigma^2_{(\overline{M}_1+\overline{M}_2)/2} = (\sigma^2_{S_{lab}}/2) + (\sigma^2_R/2n_1)$ if $\sigma^2_{R_{lab1}} = \sigma^2_{R_{lab2}} = \sigma^2_R$ and $n_1 = n_2$. Notice that the random error variance σ^2_R gets reduced by a factor of $1/(2n_1)$ while the systematic error variance $\sigma^2_{S_{lab}}$ gets reduced only by a factor of 2 because only two laboratories generated realizations of S_i. Extensions to Eq. (2) for nonconstant within-laboratory random error variance and/or unequal number of replicates per laboratory are straightforward.[2,8] Equation (4) is one basis for assigning a standard deviation to the consensus estimate of T based on Eq. (2) by using a method of moments approach that sets observed variances equal to expected variances.[2,3,5] For example, the sample variance of $\{\overline{M}_1, \overline{M}_2, \ldots, \overline{M}_n\}$ across laboratories $s^2 = \sum_{i=1}^n (\overline{M}_i - \overline{\overline{M}})^2/(n-1)$, where $\overline{\overline{M}}$ is the overall mean (the mean of the laboratory measurements, some of which could be average measurements over multiple experiments) has the expected value $\sigma^2_{S_{lab}} + \sum_{i=1}^n \sigma^2_{R_i}/n$ if we assume one measurement per laboratory as in examples 1 and 2 of Sec. V. Therefore, the observed sample variance of $\{\overline{M}_1, \overline{M}_2, \ldots, \overline{M}_n\}$ provides a simple option to estimate $\sigma^2_{S_{lab}}$. Specifically, assuming each $\sigma^2_{R_i}$ is known, $\hat{\sigma}^2_{S_{lab}} = \max(0, s^2 - \sum_{i=1}^n \sigma^2_{R_i}/n)$, and the "hat" notation denotes a parameter estimate. Alternatively, standard approximations based on the estimated curvature of the likelihood in the case of ML estimation or on a numerical option using Markov Chain Monte Carlo[7] (MCMC) in the case of Bayesian estimation are available to assign a standard deviation to the consensus estimate of T. See examples 1 and 2 in Sec. V. As explained in Sec. V, we prefer the MCMC-based option but for completeness also investigate the ML option and a moments-based option when available.

III. META-ANALYSIS

There is no single definition of meta-analysis, but generally, a meta-analysis involves combining information from multiple experiments using statistical methods with the goals to aggregate and/or contrast findings from several related studies or experiments. Viechtbauer[9] briefly reviews some freely available software for meta-analysis and then describes the `metaphor` package for meta-analysis in `R` (Ref. 10). Bax et al.[11] describe meta-analysis with interactive explanations that include assessment of publication bias, tests for unequal laboratory variances, and both fixed and random effects models. Fixed effects models are concerned only with the specific errors S_i in the study. Random effects models regard S_i as random realizations from a process to be characterized, for example, by estimating $\sigma^2_{S_{lab}}$. In the examples in Sec. V, we regard the errors S_i as random effects. And, unlike Refs. 9 and 10, this paper emphasizes the role of the assumed underlying measurement error model in determining the consensus value and the estimate of its standard deviation. Although not considered in this paper, a slight extension of our Bayesian approach for model 2 would allow us to estimate each of the errors S_i. The resulting estimate $\hat{S}_i = a_i(\overline{M}_i - \overline{\overline{M}})$, where the shrinkage factor $a_i < 1$ depends on the relative sizes of $\hat{\sigma}^2_{S_{lab}}$ and $\sigma^2_{R_i}$.

One benefit of multilaboratory experiments is the potential to assess whether there is evidence for laboratory-specific systematic errors S_i such as in Eq. (2), and if so, how the S_i are distributed. For example, Burr and Doss[12] and Burr[13] develop a nonparametric modeling option for S_i (in meta-analysis for clinical trials in medical experiments) using a mixture of conditional Dirichlet processes. A somewhat similar approach is taken by Jara et al.,[14] but for more general situations that are beyond our scope. The approach by Burr and Doss[12] allows the data (see Sec. V.G) to suggest to what extent the distribution of the S_i agrees with, for example, the Gaussian distribution.

IV. STATISTICAL MODELING

Equations (1), (2), and (3) will be our models 1, 2, and 3, respectively. The S_i terms are a candidate explanation for differences among laboratory results being larger than their individual uncertainty estimates ($\sigma^2_{R_i}$) suggest. Model 3 ignores the possibility of laboratory-specific biases but allows for underestimation of $\sigma^2_{R_i}$ by some of the laboratories.

Equation (2) is among the simplest defensible models that allows for nonoverlapping error bars such as implied by the nonoverlapping Gaussian distributions centered on each measured value in Fig. 1 and extending to

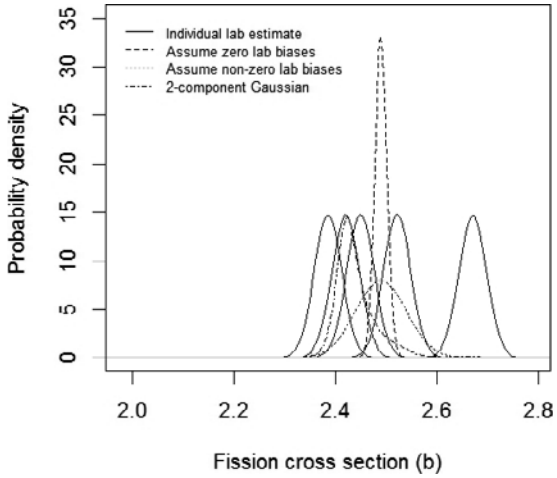

Fig. 1. Example of discrepant data from Hanson.[6] Five individual laboratory means (2.385, 2.521, 2.449, 2.420, and 2.670) and standard deviations (0.027, 0.027, 0.027, 0.027, 0.027), respectively.

TABLE I

Root-Mean-Squared Error for Using Model 2; Model 3 with $\beta = 0.1$, $\gamma = 10$; and Model 3 with Fitting β and γ, where the True Model is Model 2*

n	Model 2	Model 3, $\beta = 0.1$, $\gamma = 10$	Model 3, Fit β and γ	$BIC_2 - BIC_3$ to Choose Model	
2	0.074	0.089	0.078	0.074	1000
3	0.059	0.076	0.072	0.060	903
5	0.048	0.071	0.063	0.049	929
10	0.033	0.058	0.045	0.045	951
20	0.023	0.043	0.032	0.024	958
40	0.016	0.035	0.027	0.017	969
80	0.012	0.028	0.025	0.012	987
100	0.010	0.024	0.026	0.011	982

*The last column uses the model having the smallest BIC value, and the second entry in the last column is the number of simulations out of 1000 for which the smallest BIC value corresponded to the correct model. Entries are repeatable to approximately ± 0.001.

$\pm k\sigma_{R_i}$, where k is usually approximately 2 to 3. Because the effects of the S_i are not included in σ_{R_i}, if Eq. (2) is an appropriate measurement error model, then nonoverlapping error bars are not surprising and indicate the presence of laboratory-specific biases.

Another defensible model based on a mixture of Gaussian distributions is described in Eq. (3). Statistical modeling is the art and science of using domain knowledge and/or experimental data to select an appropriate data model or likelihood. Modern methods to compare and select models use the likelihood or integrated likelihood.[7] The basic concept is that better models explain the data better, in a way that can be numerically quantified via the likelihood.

In addition to model selection based on the likelihood, parameter estimation and residual or model diagnostics are necessary for a complete analysis. Examples are given in Sec. V. We note here that it is traditional to state the uncertainty of a consensus result using a single standard deviation, which is appropriate if the composite result can be adequately modeled as having a Gaussian distribution centered on the true value of the measurand T. However, as will be made clear in examples 1 and 2, it is sometimes necessary to use more caution in the expression of the uncertainty in the consensus value. For example, the posterior probability distribution for a quantity might not have approximately a Gaussian distribution (see Fig. 4).

Note that model 2 assumes a mature assay protocol in that all laboratories understand their sources of variation and have high-quality estimates of σ_{R_i}. Further, the modeling assumption $S_i \sim N(0, \sigma^2_{S_{lab}})$ implies that all laboratories have equal understanding of their systematic sources of variation because there is only one variance $\sigma^2_{S_{lab}}$. Such a model could arise due to each laboratory having a very similar procedure to estimate detector efficiency, for example. And, if that procedure to estimate detector efficiency is not repeated across replicate experiments by a laboratory, then estimation error in detector efficiency would contribute to laboratory-specific systematic error. Model 3 is quite different, assuming that some laboratories understate their true σ_{R_i}, which would imply that there is not yet an established and reproducible protocol for doing the experiment. In the example problems, some limited ability to distinguish between models 2 and 3 will be evident, as seen in Tables I and II through the Bayesian Information Criterion (BIC).

V. EXAMPLES

Hanson[6] describes meta-analyses for measuring reaction rates but after rejecting model 1 assumed a special case of Eq. (3) without considering alternative models. One of the reaction rates is the neutron-induced fission cross section for ^{239}Pu at an incident neutron energy of 14.7 MeV. The number of experiments ranges from 5 to 16 depending on the particular measurand considered. Example 1 has five laboratory results. Example 2 has 16 laboratory results.

V.A. Example 1

As a preliminary to evaluating the real laboratory results, Hanson[6] uses a synthetic data set from five

TABLE II

Root-Mean-Squared Error for Using Model 2; Model 3 with $\beta = 0.1$, $\gamma = 10$; and Model 3 with Fitting β and γ, where the True Model Is Model 3.*

n	Model 2	Model 3, $\beta = 0.1$, $\gamma = 10$	Model 3, Fit β and γ	$BIC_2 - BIC_3$ to Choose Model	
2	0.137	0.147	0.138	0.137	0
3	0.092	0.061	0.072	0.086	130
5	0.052	0.029	0.016	0.046	237
10	0.040	0.010	0.010	0.030	417
20	0.024	0.007	0.007	0.017	493
40	0.016	0.005	0.005	0.011	479
80	0.010	0.003	0.003	0.008	482
100	0.009	0.003	0.003	0.007	493

*The second entry in the last column is the number of simulations out of 1000 for which the smallest BIC value corresponded to the correct model. Entries are repeatable to approximately ±0.001.

laboratories with the estimated cross section (measured in barns) provided as 2.385, 2.521, 2.449, 2.420, and 2.670 from the five laboratories, with each laboratory claiming $\sigma_{R_i} = 0.027$.

First, note that if the simple model [Eq. (1)] $M_i = T + R_i$ is adequate and if one measurement per laboratory is assumed so the index j is dropped, then it is well-known[15,16] that the likelihood function

$$f(m_1, m_2, \ldots, m_5) = \frac{1}{\sqrt{2\pi\sigma_{R_1}^2}} \frac{1}{\sqrt{2\pi\sigma_{R_2}^2}} \cdots \frac{1}{\sqrt{2\pi\sigma_{R_5}^2}} \times e^{-\left[\frac{(m_1-T)^2}{2\sigma_{R_1}^2} + \frac{(m_2-T)^2}{2\sigma_{R_2}^2} + \cdots + \frac{(m_5-T)^2}{2\sigma_{R_5}^2}\right]}$$

can be maximized as a function of T, resulting in a ML estimate for T (which is the same as the weighted least-squares estimate in this case) of

$$\hat{T} = \frac{1}{\sum_{i=1}^{5} \frac{1}{\sigma_{R_i}^2}} \sum_{i=1}^{5} \frac{m_i}{\sigma_{R_i}^2} \ ,$$

which reduces to the ordinary unweighted average $\hat{T} = \sum_{i=1}^{5} m_i/5$ if all the laboratory variances $\sigma_{R_i}^2$ are equal. Therefore, if one assumes the model $M_i = T + R_i$ with $\sigma_{R_i} = 0.027$, and $R_i \sim N(0, \sigma_{R_i}^2)$, then $\hat{T} = 2.489$ is the consensus estimate, with standard deviation $0.027/\sqrt{5} = 0.012$, a 95% confidence interval (CI) is $2.489 \pm 1.96 \times 0.012$, which is (2.47, 2.51).

We will show that examples 1 and 2 exhibit evidence of having inconsistent measurements, so other options to calculate a consensus value will be presented. For example, as another option to consider, because Eq. (1) is not supported by the data (see Sec. V.B), one might informally opt to use the sample standard deviation

$$s = \sqrt{\frac{1}{5-1} \sum_{i=1}^{5} (m_i - \bar{m})^2} = 0.113$$

and the 0.975 quantile of the t distribution with 4 degrees of freedom (2.78) to obtain an approximate 95% CI of $2.489 \pm 2.78 \times 0.113/\sqrt{5}$, or (2.35, 2.63), which is again centered on the mean $\bar{M}$ but is much wider than the previous CI that arose from believing the laboratory claims that each $\sigma_{R_i} = 0.027$ and that Eq. (1) is adequate. We say "informally opt" because there is no explicit attempt to choose an appropriate measurement error model for the $m_1, m_2, \ldots, m_5$ values. However, Ref. 1 shows that this option arises as a special case of the approach in Ref. 1. These two options and other options described below for examples 1 and 2 to calculate a consensus value are summarized in Table III for examples 1 and 2.

TABLE III

Consensus Values and Estimated Standard Deviation of Consensus Values for T in First Cell Entry, and 95% CIs for T in Second Cell Entry for Examples 1 and 2 for Models 1, 2, and 3*

Example	Model 1 Assumed Known σ_{R_i}	Model 1 Estimated σ_R	Model 2 Using MCMC Known σ_{R_i}	Model 3 Fix $\beta = 0.1$, $\gamma = 10$ Using MCMC Known σ_{R_i}	Model 3 Estimate β and γ Using MCMC Known σ_{R_i}
1	2.49, 0.012 (2.47, 2.51)	2.49, 0.050 (2.35, 2.63)	2.49, 0.040 (2.41, 2.57)	2.43, 0.015 (2.40, 2.46)	2.44, 0.033 (2.40, 2.53)
2	2.44, 0.013 (2.42, 2.47)	Not available	2.50, 0.038 (2.43, 2.58)	2.45, 0.023 (2.41, 2.50)	2.45, 0.027 (2.40, 2.51)

*Model 2 is mildly or strongly preferred over all other models, with model 3 (with estimation of β and γ) as the second-best model.

V.B. Evidence for Laboratory Results Being Inconsistent with Eq. (1)

In Fig. 1, using the assumed value $\sigma_{R_i} = 0.027$ for each laboratory to depict a fixed-width Gaussian distribution centered on each m_i, it appears that 2.670 is an outlier and that 2.521 is also possibly an outlier if we assume all $S_i = 0$. More formally, Hanson[6] computed

$$\chi^2 = \frac{\sum_{i=1}^{5} \{\overline{M}_i - \overline{\overline{M}}\}^2}{\sigma_R^2} = 69.9 \ ,$$

which corresponds to a very small p value of approximately 10^{-14} for 4 degrees of freedom in χ^2 for testing the null hypothesis that Eq. (1) is adequate . Following convention, the p value is defined as $p = Prob(\chi^2_{4df} > 69.9) \approx 10^{-14}$, which is the probability that χ^2 with 4 degrees of freedom exceeds 69.9. The term $\overline{M}_i$ is the laboratory i mean (in this case, an average of one measurement, but in general laboratories might report the average of n_i measurements), and the term $\overline{\overline{M}}$ is the average (unweighted in this case) of the individual laboratory means.

Alternatively, to assess whether the model $M_{ij} = T + R_{ij}$ [Eq. (1)] is adequate, the probability that the sample range of five measurements from a Gaussian distribution having the same mean T and standard deviation σ_R is greater than approximately 10.6 σ_R (the observed range from the five laboratories) is much less than 0.0001 (estimated on the basis of 10^6 simulations). So, also by using this alternate statistical test, we find the p value to be very small, and we conclude that the model $M_{ij} = T + R_{ij}$ is almost surely not adequate, so meta-analysis of inconsistent measurements is required.

By "inconsistent measurements," we mean measurements that are inconsistent with Eq. (1), $M_{ij} = T + R_{ij}$. Similarly, Cacuci and Ionescu-Bujor[1] considered the laboratory data to be inconsistent if any of the distances $|m_i - m_j| > \sigma_i + \sigma_j$. Specifically, if any of $|m_i - m_j| > \sigma_i + \sigma_j$, then Ref. 1 assumed that a known subset of laboratories underestimated σ_{R_i} because of the presence of unrecognized random errors. This maximum distance option to decide whether there is evidence for inconsistency among laboratory results does not account for the number of laboratories n and has a false alarm rate (the rate of declaring laboratory results to be inconsistent when Eq. (1) is adequate) that depends on the magnitudes of σ_i and σ_j (because the standard deviation of $m_i - m_j$ is $\sqrt{\sigma_i^2 + \sigma_j^2}$, not $\sigma_i + \sigma_j$). If there is sufficient information to suggest that the effect of unrecognized errors leads to the true random error standard deviation in each laboratory being larger than the reported σ_{R_i}, then results in Ref. 1 (that use a maximum entropy prior in a Bayesian approach) provide a new meta-analysis option for a special case of Eq. (3) in which all laboratories are assumed to have unrecognized errors.

As a complement and extension to Ref. 1, our inference task requires us to assess candidate measurement error models such as Eqs. (2) and (3) and/or to consider alternatives to the Gaussian distribution. That is, because the Eq. (1) model $M_{ij} = T + R_{ij}$ is not adequate for this small example with five laboratory results, it is appropriate to develop alternative data models. For example, suppose instead of $R_{ij} \sim N(0, \sigma^2_{R_i})$, we let R_{ij} have a scaled t distribution, $R_{ij} \sim (\sigma_{R_i}/\sqrt{3})t_{\nu=3}$, where $t_{\nu=3}$ is the t distribution with 3 degrees of freedom [the smallest integral degrees of freedom for which the t distribution has a variance, and that variance is equal to 3 (Ref. 17)]. That is, we still assume the laboratory i random standard deviation is σ_{R_i} but change the distributional assumption from Gaussian to the broader-tailed t distribution with 3 degrees of freedom. Then, the p value based on the sample range of five observations from a scaled $t_{\nu=3}$ changes from <0.0001 to approximately 0.002 (on the basis of 10^6 simulations). Therefore, even with the $t_{\nu=3}$ distributional assumption, the model $M_{ij} = T + R_{ij}$ with R_{ij} having a t distribution is probably not adequate because although arbitrary, it is convention to regard a p value of 0.05 or smaller as being strong evidence against a candidate model.

These model checks and associated p values suggest that the model $M_{ij} = T + R_{ij}$ cannot be made adequate for the data from the five laboratories by changing the distributional assumption for R_{ij} from Gaussian to something more diffuse such as a t distribution.

V.C. Model Parameter Estimation

Although Eq. (1) is not adequate for these data, either Eq. (2) with nonzero S_i or Eq. (3) with underestimated laboratory σ_{R_i} values has the potential to adequately describe this data set. This section describes parameter estimation for Eqs. (2) and (3).

Informally, without using statistical inference, Hanson[6] claims that using Eq. (3), $\beta = 0.1$ and $\gamma = 10$ seemed to fit the data from the five laboratories. We will consider the "fix $\beta = 0.1$ and $\gamma = 10$" option but also include the option to estimate β and γ. Using the optimizer `nlm` in `R` to implement the ML option to estimate β and γ in model 3, we obtain $\hat{T} = 2.425$, $\hat{\beta} = 0.56$, and $\hat{\gamma} = 5.85$. With β set to 0.1 and γ set to 10 as in Hanson,[6] we obtain $\hat{T} = 2.486$. For model 2, we obtain $\hat{T} = 2.489$ and $\hat{\sigma}_{S_{lab}} = 0.087$. For comparison, the ordinary unweighted average of the five laboratory means is 2.489. Because in this example each laboratory's estimate of σ_{R_i} is 0.027, the model 2 consensus estimate of T is the same as the unweighted average, 2.489.

To estimate model parameters and the standard deviations of those estimates, for completeness we experimented with both ML estimation and numerical Bayesian estimation using MCMC as implemented in the `metrop` function in the `mcmc` package for `R` (Ref. 10). We note here that all the MCMC results in this paper were obtained

using `metrop` and the prior probability distribution for each unknown variance was uniform from 0 to a large upper limit as determined by the sample variance of the relevant data. Not surprisingly, ML results can be similar to results using a numerical Bayesian approach based on the estimated maximum posterior when the posterior parameter distribution is symmetric and the prior is diffuse (has a large uncertainty as defined by its standard deviation). See example 2 in Sec. V.F where the usual convention of expressing the uncertainty in a quantity (even a consensus quantity) using a single standard deviation can be appropriate. However, because ML estimation typically is only asymptotically unbiased and because the quality of ML-based CIs is sometimes difficult to assess, we generally recommend the MCMC approach.

In Fig. 1, the estimated posterior distribution for the consensus of each of three meta-analysis options is indicated as a dashed curve. Model 1 assumes zero laboratory biases [$M_{ij} = T + R_{ij}$, Eq. (1)], model 2 assumes nonzero laboratory biases [Eq. (2)], and model 3 assumes a two-component Gaussian mixture [Eq. (3)]. Notice that model 1 ignores the apparent contradictions among some of the laboratory results, resulting in an overly optimistic estimate (the estimated posterior is too narrow) of the uncertainty in the consensus estimate. Model 2 results in the same consensus value, 2.489 as model 1, but with a much wider posterior distribution. Model 3 assigns positive probability that the two largest observations should be downweighted so that the consensus value is smaller, 2.425. Notice the skewed shape for the posterior probability in model 3, which arises because there is recognized uncertainty regarding which subset of laboratories has underestimated σ_{R_i}. Model 2 via MCMC results in an estimated standard deviation of $\hat{T}$ of 0.040 while model 3 results in a more optimistic value of 0.033 if β and γ are estimated and of 0.015 if β is fixed at 0.1 and γ is fixed at 10. For comparison, the method of moments approach, which is available for model 2 (but not for model 3), leads to $\hat{T} = 2.489$ with an estimated standard deviation of $\hat{T}$ of 0.050. For all reported MCMC results, the posterior estimates of T and of the standard deviation of the estimate of T are based on 10^6 MCMC samples, which we repeated twice to ensure negligible uncertainty due to using a finite (but large) number of MCMC observations.

V.D. Model Selection

Model 2 assumes that each laboratory knows its random error standard deviation σ_{R_i} (or can estimate σ_{R_i} adequately) but that systematic errors are nonnegligible and are distributed as $S_i \sim N(0, \sigma^2_{S_{lab}})$. Both models 2 and 3 assume the random errors from each laboratory have a Gaussian distribution, but model 3 assumes that one or more laboratories understate σ_{R_i} and ignores the possibility of laboratory-specific systematic errors.

Model 2 is most appropriate when the assay protocol is relatively well established and consistent across laboratories. Even with an established assay protocol, systematic uncertainties can arise, for example, in estimation of the ^{239}Pu induced fission neutron output and energy spectrum, from uncertainties in determining the efficiency of the neutron detectors, the flight path distance, and the possible time offset in the time-of-flight measurement (Haight et al.[18]).

Alternatively, such uncertainties can be included in the random error variance σ_{R_i}, but model 3 allows for the possibility that some such random error sources are neglected by some laboratories, suggesting that assay protocol is less established than in model 2. Model 3 is referred to as the scale-contaminated normal distribution and has been studied, for example, by Gleason[19] and by Cacuci and Ionescu-Bujor,[1] who study a version of model 3 in which some or all laboratories have unrecognized errors that we denoted previously as R_B to distinguish them from the recognized errors denoted as R_A.

To investigate whether observed data can indicate which model is preferred, we use a simulation study. In the simulation study, we first let the true model be model 2 and then let the true model be model 3. We also compare inference options appropriate for models 2 and 3 for both situations. The parameters for the simulated data were the same or nearly the same as the corresponding estimated parameters ($T = 2.489$; $\sigma_{S_{lab}} = 0.1$; $\sigma_{R_i} = 0.027$; and set $\beta = 0.1$ and $\gamma = 10$).

The ML for each model can be used to compare models, using the well-known BIC. The BIC is not the only option for comparing models but is among those that have stood the test of time.[4] The BIC is defined as $\text{BIC} = -2\log(\text{ML}) + k\ln(n)$, where ML is the maximum value of the likelihood (the likelihood evaluated at the parameter values that maximize the likelihood); k is the number of model parameters; and, in our examples, n is the number of laboratories. Models with a smaller-valued BIC are the most plausible models. As a rule of thumb, it is suggested that if the BIC value for model 2, BIC_2, is 10 or more less than BIC_3, then model 2 is quite strongly preferred over model 3. A few quantitative studies have been published to support this type of BIC calibration. See Aitken[7] and the references therein.

Figure 2 plots $\text{BIC}_2 - \text{BIC}_3$ for $n = 3$, 5, 10, 20, 40, 80, and 100 laboratories (one observation per laboratory) in the case that model 2 is correct, with $\sigma_S = 0.1$ and $\sigma_R = 0.027$ (0.1 and 0.027 are values suggested from analysis of variance applied to model 1) (see Refs. 2, 4, and 8).

Realistically, n will be from approximately 3 to 20, but the larger values of n are included here for completeness. The horizontal line at $\text{BIC}_2 - \text{BIC}_3 = -10$ is given for reference. The upper and lower tick marks are the 0.025 and 0.975 quantiles estimated from 1000 simulations of each case. The "M" plotting symbols mark the

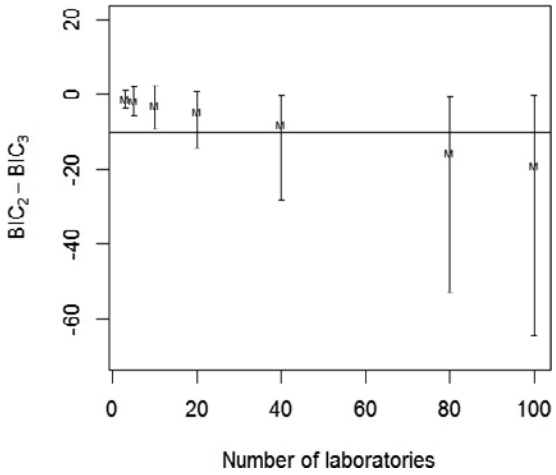

Fig. 2. BIC_2 - BIC_3 for $n = 3, 5, 10, 20, 40, 80$, and 100 laboratories in the case that model 2 is correct, with $\sigma_S = 0.1$ and $\sigma_R = 0.027$. The plotting symbol "M" denotes the mean of the BIC_2 - BIC_3 distribution at each value of n.

mean difference. Notice that the distribution of $BIC_2 - BIC_3$ values is not generally symmetric around the mean.

Figure 3 is the same as Fig. 2, but model 3 is the correct model with $\beta = 0.1$ and $\gamma = 10$ as used by Hanson.[6] In this case, the horizontal line at $BIC_2 - BIC_3 = 10$ is given for reference.

One key finding is that Figs. 2 and 3 suggest that at least for small n (perhaps $n \leq 10$ laboratories), it is difficult to distinguish between models 2 and 3 no matter which model is the true model. And, the estimated consensus value based on ML does change as the assumed model changes from model 2 to model 3. Therefore, in another simulation experiment, we will use the BIC to choose a model and evaluate to what extent this type of model selection and inference can outperform simply using either model 2 or model 3 in all cases. The BIC values for model 2, model 3 (forcing $\beta = 0.1$, and $\gamma = 10$), and model 3 (using the MLs estimates for β and γ) are -5.5, -1.1, and -3.9, respectively. Recall that small BIC values are preferred, so model 2 is preferred over model 3, and with model 3 it is better to use estimates of β and γ from three-dimensional optimization rather than the plug-in values $\beta = 0.1$ and $\gamma = 10$. Alternatively, one could argue that if time and budget permit, it is better to "get under the hood" of each laboratory's experimental approach to see if expert judgment can play a role in improving understanding of total laboratory measurement error.

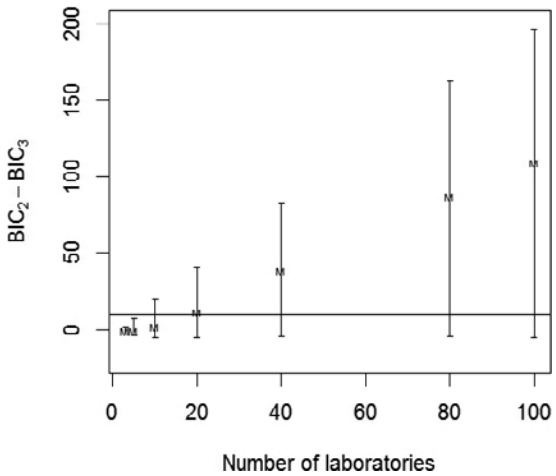

Fig. 3. BIC_2 - BIC_3 for $n = 3, 5, 10, 20, 40, 80$, and 100 laboratories in the case that model 3 is correct, with $\beta = 0.1$ and $\gamma = 10$. The plotting symbol "M" denotes the mean of the BIC_2 - BIC_3 distribution at each value of n.

Aitken[7] provides a good discussion of model selection options, including the BIC and the Bayes factor. The Bayes factor for comparing model 2 to model 3 is the ratio defined as the marginal likelihood of the data for model 2 divided by the marginal likelihood for model 3, where marginal likelihood is the average likelihood with respect to the prior distribution for the parameter values in the model, denoted $L_M = \int L_M(\theta)\pi(\theta)\, d\theta$, where $\pi(\theta)$ is the prior probability density for θ, and $L_M(\theta)$ is the likelihood, such as in Eq. (3), viewed as a function of the parameter $\theta = T$. Alternatively, model selection can be based on the average likelihood with respect to the posterior distribution of θ, which is the basis for the deviance information criterion (DIC), defined as $DIC = p_D + \bar{D}$, where the effective number of model parameters is $p_D = \bar{D} - D(\bar{\theta})$ and where the deviance is defined as $D(\theta) = -2\ln(L_M(\theta)) + C$ with C a constant that does not need to be evaluated because only the difference in deviance matters in comparing two models. Preferred models have smaller DIC values. The notation $\bar{D}$ denotes the average of $D(\theta)$ with respect to the posterior for θ, and $D(\bar{\theta})$ denotes the deviance $D(\theta)$ evaluated at the posterior mean of θ. The DIC is suited for Bayesian model selection problems for which the posterior distribution of θ is obtained by MCMC as we have done here.[7]

For model 2, model 3 (forcing $\beta = 0.1$, and $\gamma = 10$), and model 3 (estimating β and γ), the marginal likelihoods with respect to the prior of θ are 0.31, 0.08, and 0.46 (higher values are preferred), respectively, and the DIC values are -8.5, -4.7, and -8.6, respectively (smaller values are preferred). So, the marginal likelihood with respect to the prior for θ is highest for model 3 with β and γ estimated, but the DIC is approximately the same for model 2 (-8.5) and model 3 (-8.6), again with estimating β and γ being better than forcing $\beta = 0.1$, and $\gamma = 10$. Recall that the BIC indicated a preference for model 2, or if model 3 is used, then again it is better to estimate β and γ than to fix them at 10 and 0.1, respectively. The BIC, marginal likelihood, and the DIC values were evaluated using 10^6 observations from either the estimated posterior or the assumed prior. By

repeating the set of 10^6 observations, we found that 10^6 is sufficiently many observations to have very high confidence in the rankings of the marginal likelihoods, with the estimated marginal likelihoods repeatable across sets of 10^6 observations to within approximately $\pm 1\%$ relative to the given values. In all cases, we used either uniform priors from 0 to a large upper value for the appropriate subset of model parameters T, $\sigma_{S_{lab}}$, β, and γ or used diffuse gamma priors; results are indistinguishable between these two diffuse priors.

V.E. Estimation Performance Assessment Using Another Simulation Study

In another simulation study, we estimate the true value T and report results for ML estimation. Results in this simulation study for MCMC are essentially the same as for ML, provided the posterior maximum is used rather than the posterior mean. We have found that the posterior distribution is almost but not exactly symmetric, so this explains why the maximum posterior estimate is not exactly the same as the mean posterior estimate.

Table I gives the root-mean-squared error (RMSE) across 1000 simulations for estimating the true value T using ML applied to model 2, and ML to model 3 with $\beta = 0.1$ and $\gamma = 10$ or with a three-dimensional maximization allowing the data to fit the values of β and γ. In Table I, the true model is model 2. Table II is the same as Table I, but in Table II, the true model is model 3 with $\beta = 0.1$ and $\gamma = 10$. In both Tables I and II, the BIC is applied to choose which ML estimate to use. Because it is more defensible to allow the data to fit the values of β and γ, the BIC for model 2, BIC_2, was compared to that for model 3, BIC_3 (with fitted values of β and γ), and if $BIC_2 < BIC_3$, then the model 2–based ML estimate was used.

In addition to the RMSE of a method to estimate T, it is important to have a good-quality estimate of the standard deviation of $\hat{T}$. Either ML or MCMC can estimate the standard deviation of $\hat{T}$. For example, for the simulation results in column 2 (which assumes model 2 is correct) of Table I (for which model 2 is correct), the observed standard deviation across the 1000 simulations of $\hat{T}$ for $n = 5$ was 0.046. The corresponding estimated standard deviation was different for each simulation, with an average of 0.045 across simulations, which indicates that the standard ML approximation for the standard deviation of a parameter estimate is adequate in this example when the correct model is used to fit the data. To estimate the standard deviation of the ML estimator, we used the Hessian matrix available from nlm in R. The Hessian matrix is the standard tool used in ML estimation to quantify the curvature of the likelihood function (which determines the estimated standard deviation of the ML estimator). As another check of the adequacy of the standard ML approach, we compared the nominal CI coverage probability to the observed CI coverage probability. Specifically, the interval $\hat{T} \pm 2\hat{\sigma}_{\hat{T}}$ (which varied across simulations) included the true T value (which was assumed to be 2.489 for each simulation) in 941 of 1000 simulations, which is a 94% observed coverage compared to a nominal coverage of 95% assuming that $\hat{T}$ is approximately Gaussian in distribution (so that the interval $\hat{T} \pm 2\hat{\sigma}_{\hat{T}}$ should have approximately a 0.95 probability of including T). This second check of the ML approach also indicates good ML performance. On the other hand, when the wrong model was used to fit the data such as using model 2 when model 3 was correct again for $n = 5$, there was a noticeable difference (repeatable across sets of 1000 simulations) between the observed standard deviation of $\hat{T}$ across the 1000 simulations (0.038) versus the average estimated standard deviation of 0.027. And, using ML with model 3 consistently leads to overestimation of the standard deviation of $\hat{\beta}$ and $\hat{\gamma}$ even when the correct model is assumed.

In general, it is not unusual for an approximation to the standard deviation of a ML parameter to be somewhat inaccurate for small n. For example, again with $n = 5$, but with model 3 being the correct model with $\beta = 0.1$ and $\gamma = 10$, the estimated standard deviations of the estimates $\hat{\beta}$ and $\hat{\gamma}$ (the ML estimates) are both approximately 0.014, but the observed standard deviations across 1000 simulations are both approximately 0.005. However, the observed RMSEs of $\hat{\beta}$ and $\hat{\gamma}$ for estimating β and γ across the 1000 simulations are 0.017 and 0.022, respectively, so there is noticeable bias included in these RMSEs, and this bias is not included in the estimated standard deviations of $\hat{\beta}$ and $\hat{\gamma}$. These findings regarding the RMSE for estimating T and regarding estimating the standard deviation of $\hat{T}$ suggest a need for simulation studies such as this to be a part of these types of meta-analyses.

From Table I we note that as expected, when the true model is model 2, inference using model 2 is preferred. And, the sample size n (the number of laboratory estimates) does not need to be large to choose the correct model using the BIC. This fortunate situation is not a general result. It arises because models 2 and 3 provide a roughly comparable fit to the data as measured by the residual sums of squares, but model 2 has only two parameters while model 3 has three parameters. Table II indicates that the BIC does not lead to as reliable selection of the correct model 3 even for as large as $n = 100$ laboratory results. Nevertheless, using the BIC to choose the model still manages to perform better (have smaller RMSE, see the last column of RMSE values in Table I) than blindly using the wrong model 2, although not as well as the omniscient approach that correctly uses model 3 for all realizations. And, using model 2 is a valid starting point for the analysis even if the BIC suggests a preference for another model such as model 3 in this example.

We note here that the performance of the BIC to distinguish between models 2 and 3 will depend on the

relative parameter sizes, and in a separate simulation study of 10^4 simulations, when $\sigma_{S_{lab}}$ increases relative to the average value of σ_{R_i}, it is simpler to recognize model 2 as the correct model. For example, with $\sigma_{R_i} = 0.027$ for all laboratories as in Fig. 1, when $\sigma_{S_{lab}} = 0.1$ (approximately the example 1 value), 0.5, and 1, there is strong evidence for the correct model 2 over model 3 ($BIC_2 < BIC_3 - 10$), in 10 of 1000 simulations, 480 of 1000 simulations, and 970 of 1000 simulations, respectively.

V.F. Example 2

The 16 measurements of the ^{239}Pu cross section for an incident neutron energy of 14.7 MeV in Hanson[6] were collected from 1954 to 2001. The 16 means in barns are 2.750, 2.580, 2.820, 2.580, 2.520, 2.560, 2.650, 2.390, 2.290, 2.532, 2.440, 2.670, 2.420, 2.449, 2.521, and 2.385. The corresponding reported 16 standard deviations σ_{R_i} are 0.14, 0.09, 0.141, 0.11, 0.088, 0.15, 0.30, 0.076, 0.052, 0.05, 0.092, 0.08, 0.028, 0.027, 0.081, and 0.026. Again, we assume the sample size $n_i = 1$ for each experiment, so the number of observations is $n = 16$.

First, $\chi^2 = \sum_{i=1}^{16} \{\overline{M}_i - \overline{\overline{M}}\}^2 / \sigma_{R_i}^2 = 44.65$, which corresponds to a p value of 0.00005 for model 1, so these data appear inconsistent, and again, we consider models 2 and 3 rather than model 1. Model 2 results in the ML estimate $\hat{\sigma}_{S_{lab}} = 0.09$, which is close in value to the average of the standard deviations σ_{R_i} (0.096) and $\hat{T} = 2.49$, which is the same as the ordinary unweighted average of the 16 results. The estimated standard deviation for $\hat{T}$ using MCMC, ML, or Eq. (4) for a moments-based approach to parameter estimation for Eq. (2) is approximately 0.038, 0.033, and 0.035, respectively.

Using model 3, we obtain the ML estimates for T, β, and γ as 2.44, 0.99, and 1.51, respectively. Or, using metrop in R to implement MCMC, we obtain the maximum posterior estimates for T, β, and γ as 2.40, 0.58, and 1.93, respectively. Alternatively, the mean posterior estimates for T, β, and γ are 2.45, 0.54, and 2.73, respectively. Notice from Fig. 4 that the posterior probability density for T, for β, and for γ is skewed; therefore, the mean of the posterior is somewhat different from the maximum posterior estimate.

Notice that the estimated standard deviation of $\hat{T}$ is larger for model 2 (0.038) than for model 3 (0.027) if MCMC is used and is also larger (0.033 versus 0.024) if ML is used. It is not unusual for ML-based estimates of standard deviation to be different from the standard deviation of the posterior based on MCMC, and provided the data likelihood is appropriate, the MCMC-based estimate is preferred because ML-based estimates of the standard deviation of parameter estimates are typically only correct in the limit as the sample size n approaches infinity. Method of moments approaches based on setting sample variances equal to theoretical variances as we did using Eq. (4) can also give considerably different

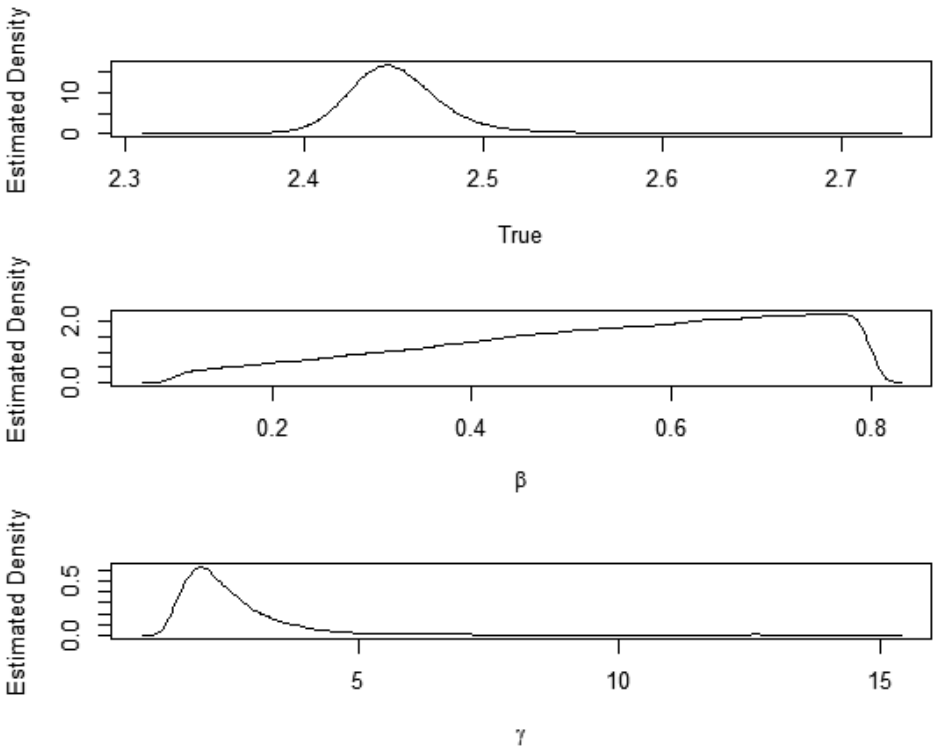

Fig. 4. Estimated posterior probability density for T, β, and γ in example 2.

(and generally slightly inferior[8]) estimates of the true parameter estimate standard deviation.

Next, to assess whether model 2 or model 3 is preferred, the BIC values for model 2, model 3 (forcing $\beta = 0.1$, and $\gamma = 10$), and model 3 (using the ML estimate for β and γ, which are approximately 0.99 and 1.51, respectively) are -13.5, -2.4, and -11.2, respectively. Recall that small BIC values are preferred, so as in example 1, using the BIC, we prefer model 2 over model 3, and with model 3, it is better to use estimates of β and γ from three-dimensional optimization rather than the plug-in values $\beta = 0.1$ and $\gamma = 10$. Notice the very large ML estimate for β of 0.99, which goes against the intent of model 3 to allow for a modest number of optimistic laboratories that underestimate σ_{R_i}. However, the mean posterior estimate for β based on MCMC is 0.54, which is more in line with the intent of model 3.

For example 2, according to either the BIC, the Bayes factor, or the DIC, model 2 appears to be preferred over model 3, and if model 3 is used, it is better to estimate β and γ rather than use some qualitative fitting criteria to choose β and γ. For example 1, the marginal likelihood with respect to the prior for θ is highest for model 3 with β and γ estimated, but the DIC is approximately the same for model 2 (-8.5) and model 3 (-8.6), again with estimating β and γ being better than forcing $\beta = 0.1$, and $\gamma = 10$. Notice also that because the BIC penalizes somewhat for increasing from two to three parameters in going from model 2 to model 3 with β and γ estimated, the BIC will favor model 2 if the maximum likelihod for models 2 and 3 are comparable. Recall that the BIC favored model 2 for both examples 1 and 2.

Also, once a model in addition to model 1 is evaluated, it is meaningful to compare the BIC for model 1 to the BIC for models 2 and 3. Model 1 is strongly rejected by the BIC in examples 1 and 2 [for model 1, BIC = 44.6 in example 1 and BIC = 324 in example 2, which are both much larger (worse) than the BIC values for models 2 and 3 given above].

V.G. Extracting the Distribution of S_i

To this point, we have done the same types of analyses for examples 1 and 2. Because the sample size $n = 16$ is larger for example 2, and because model 2 is preferred over model 3 on the basis of the BIC, the Bayes factor, or the DIC, we further consider recent research that provides a comprehensive assessment of the underlying probability distribution for S_i in Eq. (2).

For example, the package `bspmma` (Ref. 13) in `R`, uses a Dirichlet prior (a prior that makes very few assumptions about the form of the unknown distribution) to suggest how close the distribution of S_i is to Gaussian using the Bayes factor, and either the Bayes factor or the BIC can be used for model selection. In this example, the result of applying MCMC via the `dirichlet.c` function, which implements a conditional Dirichlet distribution (see below), results in $\hat{T} = 2.50$, 2.50, and 2.50, and $\hat{\sigma}_{S_{lab}} = 0.23$, 0.20, or 0.18 for the parameter $M = 5$, 20, or 100, respectively, where M determines the closeness of the distribution of S_i to Gaussian (with larger M implying closer to Gaussian). This estimate of $\hat{\sigma}_{S_{lab}}$ is approximately twice the estimate (0.09) presented above using Eq. (2) assuming a Gaussian distribution. The estimated Bayes factors can be used to choose between the conditional Dirichlet prior distribution and a mixture of Dirichlet process priors as in Fig. 3 of Ref. 13. For this example, the Bayes factor suggests that the conditional Dirichlet is preferred over the mixture of Dirichlet process priors. The conditional Dirichlet distribution as used in Ref. 13 is one in which partial information, such as the approximate value of the median, is known about the otherwise unknown distribution. And, alternate approaches to meta-analyses include emphasis on the sample median as a way to reduce the impact of suspicious individual values.

Four key parameters in the conditional Dirichlet d_1, d_2, d_3, and d_4 characterize the prior for T, with d_1, d_2, and d_4 controlling how diffuse the prior is for T and d_3 determining the prior mean for T. If, for example, d_1, d_2, d_3, and d_4 are changed from the default values of 0.1, 0.1, 0, and 1000, respectively, to 0.00001, 0.00001, 0, and 10 000, then the estimate of $\hat{\sigma}_{S_{lab}}$ changes from approximately 0.20 to 0.083, which is close to the model 2 estimate. The first set of d_1, d_2, d_3, and d_4 values leads to a diffuse prior for T, but the second set of of d_1, d_2, d_3, and d_4 leads to a considerably more diffuse prior for T.

V.H. Summary of Examples 1 and 2

We have evaluated models 1, 2, and 3 for examples 1 and 2, so to summarize, Table III lists the consensus estimate, the estimated standard deviation of the consensus estimate, and an approximate 95% CI for models 1, 2, and 3. Although ML was evaluated as described in Sec. V.C, we recommend using MCMC to obtain samples from the posterior distribution. Therefore, all Table III entries for models 1, 2, and 3 are based on MCMC rather than ML. Table III entries for model 1 are with the σ_{R_i} assumed to be known without error (column 2), or assuming that σ_{R_i} must be estimated (column 3), as an extension to model 1 in which the σ_{R_i} are assumed to be unknown but equal in example 1. Such an extension is not available in example 2 because the σ_{R_i} are assumed to be unequal, so there would be too many unknown parameters; therefore, the cell entry is "Not available."

The most trusted model 3 results are for the case where β and γ are estimated because the BIC suggested it is necessary to estimate β and γ. Because the BIC indicated only a relatively mild preference for model 2 over model 3 (with β and γ estimated), one could argue for reporting both the model 2 and the model 3 consensus values and confidence limits, saying that they depend

on whether model 2 or model 3 is used, with model 3 leading to more optimistic (narrower) confidence limits.

VI. DISCUSSION AND SUMMARY

In the context of meta-analysis for inconsistent measurements, we regard measurements to be inconsistent if a simple model such as $M_{ij} = T + R_{ij}$ (model 1) for measurement j from laboratory i of a measurand having true value T is inadequate, as in both examples 1 and 2.

Combining individual laboratory estimates requires choosing an underlying model for the data. After rejecting model 1, we considered models 2 and 3, where model 2 assumes a mature assay protocol so that all laboratories understand their sources of variation and have high-quality estimates of σ_{R_i}. Further, the modeling assumption $S_i \sim N(0, \sigma^2_{S_{lab}})$ implies that all laboratories have equal understanding of their systematic sources of variation. Model 3 is quite different, assuming that some laboratories understate their true σ_{R_i}. In the example problems, some ability to distinguish between models 2 and 3 was evident, as seen in Tables I and II through the BIC.

We noted that the performance of the BIC to distinguish between models 2 and 3 will depend on the relative parameter sizes. In Sec. V.E, we briefly discussed that in a separate simulation study when $\sigma_{S_{lab}}$ increases relative to the average value of σ_{R_i}, it is simpler to recognize model 2 as the correct model. The safest strategy is to use models 2 and/or 3 to obtain model parameter estimates, do simulation experiments such as ours, and calibrate the ability of the BIC to choose the correct model, as in our Tables I and II. And, goodness-of-fit checks such as residual diagnostics should of course always be applied to the chosen model. Additionally, data evaluators must first decide which candidate models to assess (we assessed models 1, 2, and 3), preferably on the basis of some understanding of laboratory measurement protocols and whether total error is estimated by some or all laboratories.

Potential next steps in future research should include the following. First, allow a vector of true values such as cross sections across a binned energy range, allowing for covariances between pairs of cross-section measurement bins. Second, include physical model fitting to estimate T (as a scalar or vector), allowing for simultaneous model calibration and bias adjustment as in Higdon et al.[20] If possible, allow for a laboratory-wide bias effect B, for example, in Eq. (1) due to effects such as laboratories sharing the same neutron generating sources (usually ^{252}Cf) used to estimate detector efficiency. (However, unless auxiliary information or data are available, $T + B$ is identifiable, but B is not separately identifiable from T.) Fourth, compare model selection options such as the BIC, the Bayes factor, and the DIC for various simulated data sets such as those in Tables I and II.

Regarding the fourth step, model selection criteria, we note that Aitken's[7] suggestion to use the average likelihood with respect to the posterior for θ as a model selection criterion is controversial,[21] partly because the data are used twice, once to estimate the posterior for θ, and again to evaluate the likelihood with respect to the estimated posterior for θ. For that reason, we did not report results for Aitken's posterior likelihood criterion in Sec. V. However, so-called empirical Bayesian methods also use the data twice, and the pros and cons of any such double use can be evaluated empirically using simulation. Conventional wisdom suggests that using the data twice tends to lead to some type of overfitting that would need to be somehow addressed and mitigated. For completeness, we computed Aitken's suggested posterior likelihood criterion (using MCMC observations) for both models 2 and 3. According to Aitken's posterior-likelihood-based model selection criterion, model 2 is strongly favored in example 2, and model 2 is weakly favored in example 1. The same criticism of using the data twice is levied against the DIC, so Tomohiro[22] proposed a Bayesian predictive information criterion as an alternative to the DIC. We believe that most reasonable model selection criteria will have niches for which they work quite well on select examples. A comparison of model selection criteria in the context of meta-analysis for nuclear assay results from multiple laboratories would be valuable.

Both the consensus estimate and its estimated uncertainty depend strongly on the assumed measurement error model, so measurement error model selection options were investigated. Model selection options have a role and can perform well for some sample sizes and parameter ranges. And, we suggest using auxiliary simulations as done here to gauge whether, for example, the BIC model selection option is likely to work well for a given meta-analysis. However, in all cases, there is no substitute for complete reporting of all key information about each experiment in multiexperiment evaluation meta-analysis.

REFERENCES

1. D. G. CACUCI and M. IONESCU-BUJOR, "On the Evaluation of Discrepant Scientific Data with Unrecognized Errors," *Nucl. Sci. Eng.*, **165**, 1 (2010).

2. T. BURR and G. HEMPHILL, "Multi-Component Radiation Measurement Error Models," *Appl. Radiat. Isot.*, **64**, *3*, 379 (2006).

3. T. BURR, T. SAMPSON, and D. VO, "Statistical Evaluation of FRAM Gamma Ray Isotopic Analysis Data," *Appl. Radiat. Isot.*, **62**, 931 (2005).

4. M. VANGEL and A. RUKHIN, "Maximum Likelihood Analysis for Heteroscedastic One-Way Random Effects ANOVA in Interlaboratory Studies," *Biometrics*, **55**, 129 (1999).

5. T. BURR, K. KUHN, L. TANDON, and D. TOMPKINS, "Measurement Performance Assessment of Analytical Chemistry Analysis Methods Using Sample Exchange Data," *Int. J. Chem.*, **3**, *4* (2012).

6. K. HANSON, "Bayesian Analysis of Inconsistent Measurements of Neutron Cross Sections," *AIP Conf. Proc.*, **803**, 431 (2005).

7. M. AITKEN, *Statistical Inference: An Integrated Bayesian/Likelihood Approach*, Chapman and Hall, Boca Raton, Florida (2010).

8. R. MILLER, *Beyond ANOVA*, Chap. 3, Wiley, New York (1986).

9. W. VIECHTBAUER, "Conducting Meta-Analyses in R with the Metaphor Package," *J. Stat. Software*, **36**, *3* (2010).

10. "R: A Language and Environment for Statistical Computing," R Development Core Team (2004).

11. L. BAX, L. YU, N. IKEDA, H. TSURUTA, and K. MOONS, "Development and Validation of MIX: Comprehensive Free Software for Meta-Analysis of Causal Research Data," *BMC Medical Research Methodology*, **6**, *50* (2006).

12. D. BURR and H. DOSS, "A Bayesian Semiparametric Model for Random-Effects Meta-Analysis," *J. Am. Stat. Assoc.*, **100**, *469*, 242 (2005).

13. D. BURR, "bspmma: An R Package for Bayesian Semi-Parametric Models for Meta-Analysis," *J. Stat. Software*, **50**, *4*, 1 (2012).

14. A. JARA, T. HANSON, F. QUINTANA, P. MULLER, and G. ROSNER, "DPpackage: Bayesian Semi- and Nonparametric Modeling in R," *J. Stat. Software*, **40**, *5*, 1 (2011).

15. P. BEVINGTON and D. ROBINSON, *Data Reduction and Error Analysis for the Physical Sciences*, 3rd ed., McGraw Hill, Boston, Massachusetts (2003).

16. T. BURR, T. KAWANO, P. TALOU, F. PEN, N. HENGARTNER, and T. GRAVES, "Alternatives to the Generalized Least Squares Solution to Peele's Pertinent Puzzle," *Algorithms*, **4**, *2*, 115 (2011).

17. M. DeGROOT, *Probability and Statistics*, Chap. 4, Addison-Wesley, Reading, Massachusetts (1986).

18. R. HAIGHT et al., "Accuracies in the Planned Measurements of Prompt Neutron Output in Neutron-Induced Fission of ^{239}Pu: The chi-Matrix," LA-UR-11-05576, Los Alamos National Laboratory (2011).

19. J. GLEASON, "The Scale Contaminated Normal Family," *J. Am. Stat. Assoc.*, **88**, *412*, 327 (1993).

20. D. HIGDON, J. GATTIKER, B. WILLIAMS, and M. RIGHTLEY, "Computer Model Calibration Using High Dimensional Output," *J. Am. Stat. Assoc.*, **103**, 570 (2008).

21. A. GELMAN, C. ROBERT, and J. ROUSSEAU, "Inherent Difficulties in Likelihood-Based Non-Bayesian Inference as Revealed by an Examination of a Recent Book by Aitken": http://www.stat.columbia.edu/~gelman/research/unpublished/GRR16.pdf (2012).

22. A. TOMOHIRO, "Bayesian Predictive Information Criterion for the Evaluation of Hierarchical Bayesian and Empirical Bayes Models," *Biometrika*, **94**, *2*, 443 (2007).