Models and Confidence Intervals for True Values in Interlaboratory Trials

Hari K. IYER, C. M. Jack WANG, and Thomas MATHEW

We consider the one-way random-effects model with unequal sample sizes and heterogeneous variances. Using the method of generalized confidence intervals, we develop a new confidence interval procedure for the mean. Additionally, we investigate two alternative models based on different sets of assumptions regarding between-group variability and derive generalized confidence interval procedures for the mean. These procedures are applicable to small samples. Statistical simulation is used to demonstrate that the coverage probabilities of these procedures are close enough to the nominal value so that they are useful in practice. Although the methods are quite general, the procedures are explained with the backdrop of interlaboratory studies.

KEY WORDS: Coverage probability; Generalized pivotal quantity; Heterogeneous error variances; ISO GUM; One-way random effects model; Type-B assumptions.

1. INTRODUCTION

We consider the situation where measurements of an artifact are made by each of k laboratories (or, in some cases, k different measurement methods). The ith laboratory makes n_i independent measurements, Y_{ij}, $j = 1, \ldots, n_i$. The data from the k laboratories are assumed to follow the model

$$Y_{ij} = \mu_i + e_{ij}, \tag{1}$$

where μ_i is the mean measured value for laboratory i. If μ denotes the true, unknown measurement of interest, then we write $\mu_i - \mu = b_i$ and call b_i the "bias" of laboratory i. The quantity μ is the parameter that we wish to estimate based on combined information from the different laboratories. The quantities e_{ij}, $j = 1, \ldots, n_i$, are random measurement errors associated with the ith laboratory. It is reasonable to assume that e_{ij}, $j = 1, \ldots, n_i$, are independent random variables with mean 0 and variance σ_i^2, $i = 1, \ldots, k$. Generally, this error distribution is assumed to be normal.

The particular statistical approach that is appropriate for the estimation of μ depends on what assumptions are made about the b_i or about the relationship of the μ_i to μ. Different sets of assumptions have been considered by various authors. This in turn has led to different analysis methods. Assumption A encompasses all of the different sets of assumptions that have appeared in the literature in connection with a frequentist analysis of the problem.

Assumption A. For $1 \le i \le k$, b_i is a random variable with cdf F_i whose support is contained in the interval $[m_i, M_i]$, where m_i and M_i are assumed known. (If m_i is negative infinity, then we replace the closed interval at m_i with an open interval; likewise if M_i is positive infinity.) In applications, scientists are often able to specify m_i and M_i based on past experience and/or expert judgment. We now elaborate on various special cases.

Model 1. Suppose that for each i, F_i is a normal distribution with mean 0 and variance σ^2, $m_i = -\infty$, $M_i = \infty$. We then have the one-way random-effects model with unequal sample sizes and heterogeneous error variances. This model has been considered by Rukhin and Vangel (1998), Vangel and Rukhin (1999), Rukhin, Biggerstaff, and Vangel (2000), Paule and Mandel (1971, 1982), and others. Methods for estimating μ and for obtaining an approximate confidence interval for μ have been proposed by these authors.

Model 2. Suppose that for each i, F_i is an unspecified distribution but m_i and M_i are known, finite constants. This case is equivalent to the model considered by Eberhardt, Reeve, and Spiegelman (1989). We call this model a *bounded-bias* model. Eberhardt et al. used the mean squared error of an estimator as the criterion of goodness and derived a minimax estimator for μ in the class of estimators that are linear functions of the individual laboratory means (or method means). They also proposed an associated approximate confidence interval procedure and evaluated its performance using statistical simulation.

Model 3. Suppose that for each i, F_i is a completely specified distribution. This case is equivalent to the model described in the International Organization for Standardization's (ISO) *Guide to the Expression of Uncertainty in Measurement* (GUM) (1995) where the distributions F_i are referred to as *type-B* distributions. One may regard b_i as random variables with known distributions, as we do in this article, or as parameters (measuring bias) with *informative* prior distributions, if one were to pursue a Bayesian solution to this problem. Typically, F_i's are assumed to be normal or uniform on a known interval, but other distributions are also sometimes used. For convenience, we call Model 3 a *GUM-type* model.

Model 4. Suppose that for each i, F_i is a degenerate distribution at b_i, and $\sum_{i=1}^{k} b_i = 0$. This is equivalent to assuming that a one-way "fixed-effects" model holds for the Y_{ij} and that the true value μ is the average of the k laboratory means $\mu + b_i$, $i = 1, \ldots, k$. This is a standard model, and inference about μ is straightforward in this case.

Model 5. In Model 4, suppose that the b_i's are all 0. This is the common-means fixed-effects model that has been studied extensively (see, e.g., Jordan and Krishnamoorthy 1996; Yu, Sun, and Sinha 1999).

Hari Iyer is Professor, Department of Statistics, Colorado State University, Fort Collins, CO 80523 (E-mail: *hari@stat.colostate.edu*). C. M. Jack Wang is Mathematical Statistician, Statistical Engineering Division, National Institute of Standards and Technology, Boulder, CO 80305 (E-mail: *jwang@boulder.nist.gov*). Thomas Mathew is Professor, Department of Mathematics and Statistics, University of Maryland, Baltimore, MD 21250 (E-mail: *mathew@umbc.edu*). The authors thank Andrew Rukhin for his helpful comments and the referees for their valuable suggestions that helped them improve the clarity of presentation as well as correct some errors. The third author acknowledges support of U. S. Army Research Laboratory and U. S. Army Research Office grant DAAD19-01-1-0497. This work is a contribution of the National Institute of Standards and Technology and is not subject to copyright in the United States.

Journal of the American Statistical Association
December 2004, Vol. 99, No. 468, Theory and Methods
DOI 10.1198/016214504000001682

An overview of various approaches for estimating a *consensus value* μ and a discussion of related computational methods, have been provided by Wang and Splett (1997) (see also Schiller and Eberhardt 1991).

Bayesian approaches to the problem of consensus value under different model assumptions may be appropriate for some applications. A particularly interesting Bayesian approach to the problem of laboratory bias or, more important, measurement method bias has been presented by Levenson et al. (2000). These authors considered the case of $k = 2$ measurement methods and assumed μ to be a random variable with a uniform distribution on the interval $[\min\{\mu_1, \mu_2\}, \max\{\mu_1, \mu_2\}]$. They compared their solution to a solution obtained under a Bayesian hierarchical model.

In this article we consider Models 1, 2, and 3 for Y_{ij}, and in each case, develop a confidence interval procedure for μ using the method of generalized confidence intervals (GCI) of Weerahandi (1993) (see also Weerahandi 1995). The procedure is particularly attractive because there is no requirement that samples be large for the procedure to perform adequately. Although the resulting confidence interval has no explicit expression, the required computations are simple and straightforward. Furthermore, numerical results show that the coverage probabilities of these intervals are quite satisfactory.

Under Model 2, the b_i's are assumed to be unknown constants with *known* bounds. In practice, these bounds are elicited from the researcher. It may happen that these bounds fail to yield a valid parameter space for μ. For this reason, in Section 3 we develop a test that is useful for checking the appropriateness of the bounds.

2. A GENERALIZED PIVOTAL QUANTITY FOR μ UNDER MODEL 1

The model that we consider is

$$Y_{ij} = \mu + b_i + e_{ij}, \qquad j = 1, \ldots, n_i, i = 1, \ldots, k, \tag{2}$$

where μ is a fixed unknown constant, b_i are iid normal random variables with mean 0 and variance σ^2, and for each $i = 1, \ldots, k$, e_{ij} are iid normal random variables with mean 0 and variance σ_i^2. Define

$$\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij} \qquad \text{and} \qquad SS_i = \sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2. \tag{3}$$

Note that the $\bar{Y}_i$'s and SS_i's are all independent random variables. Furthermore, the set of statistics $\{\bar{Y}_i, SS_i \mid i = 1, \ldots, k\}$ is minimally sufficient for $\mu, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2$. Except when all the n_i's are equal and the σ_i^2's are all equal, the minimal sufficient statistics are known to be not complete under this model.

A generalized pivotal quantity for μ may be developed by obtaining an expression for μ as a function of observed statistics and pivotal quantities. In the ensuing paragraphs we derive functions of minimal sufficient statistics and related pivotal quantities that facilitate this process. A general procedure for deriving generalized pivotal quantities in a large class of problems was described by Iyer and Patterson (2002).

First, we introduce the following notation for convenience. Given any vector $\mathbf{c} = (c_1, c_2, \ldots, c_k)$ of real numbers with $\sum_{i=1}^k c_i \neq 0$, define

$$\bar{Y}_{\mathbf{c}} = \frac{c_1\bar{Y}_1 + \cdots + c_k\bar{Y}_k}{c_1 + \cdots + c_k}.$$

Let $\mathbf{w} = (w_1, \ldots, w_k)$, where $w_i = 1/(\sigma^2 + \tau_i^2)$, $i = 1, \ldots, k$, and $\tau_i^2 = \sigma_i^2/n_i$. Observing that $\bar{Y}_i \sim \mathrm{N}(\mu, 1/w_i)$, it follows that

$$\hat{\mu} = \bar{Y}_{\mathbf{w}} = \frac{w_1\bar{Y}_1 + \cdots + w_k\bar{Y}_k}{w_1 + \cdots + w_k} \sim \mathrm{N}(\mu, \tau_0^2), \tag{4}$$

where $\tau_0^2 = 1/(w_1 + \cdots + w_k)$. Note that if σ^2 and the σ_i^2's are known, then $\hat{\mu}$ in (4) is the uniformly minimum variance unbiased estimator of μ. Let

$$\mathbf{Y}^t = (\bar{Y}_1, \bar{Y}_2, \ldots, \bar{Y}_k)$$

be the vector of laboratory means. Then $\mathbf{Y} \sim \mathrm{N}(\mu\mathbf{1}_k, \mathbf{G})$, where $\mathbf{1}_k$ is a k-vector of all 1's, and

$$\mathbf{G} = \mathrm{diag}(\sigma^2 + \tau_1^2, \sigma^2 + \tau_2^2, \ldots, \sigma^2 + \tau_k^2). \tag{5}$$

Let $Q = g_{\mathbf{Y},\tau^2}(\sigma^2)$ denote the residual sum of squares under the model $\mathbf{Y} \sim \mathrm{N}(\mu\mathbf{1}_k, \mathbf{G})$, where $\tau^2 = (\tau_1^2, \tau_2^2, \ldots, \tau_k^2)$. That is,

$$\begin{aligned} Q &= g_{\mathbf{Y},\tau^2}(\sigma^2) \\ &= (\mathbf{Y} - \hat{\mu}\mathbf{1}_k)^t\mathbf{G}^{-1}(\mathbf{Y} - \hat{\mu}\mathbf{1}_k) \\ &= \mathbf{Y}^t\left[\mathbf{G}^{-1} - \frac{\mathbf{G}^{-1}\mathbf{1}_k\mathbf{1}_k^t\mathbf{G}^{-1}}{\mathbf{1}_k^t\mathbf{G}^{-1}\mathbf{1}_k}\right]\mathbf{Y} \\ &= \sum_{i=1}^k w_i(\bar{Y}_i - \bar{Y}_{\mathbf{w}})^2. \end{aligned} \tag{6}$$

From standard linear model theory, we know that Q is independent of $\hat{\mu}$. Furthermore, because $\mathbf{Y}$ is independent of the SS_i's, Q is independent of the SS_i's as well. Also, $Q \sim \chi^2$ with $k - 1$ degrees of freedom.

Consider the set of $k + 2$ statistics $\{\bar{Y}_{\mathbf{w}}, Q, SS_1, \ldots, SS_k\}$. Using the relationships given by

$$Z = \frac{(\bar{Y}_{\mathbf{w}} - \mu)}{\tau_0},$$

$$U_i = \frac{SS_i}{\sigma_i^2}, \qquad i = 1, \ldots, k,$$

and

$$Q = \sum_{i=1}^k w_i(\bar{Y}_i - \bar{Y}_{\mathbf{w}})^2,$$

where Z, Q, and U_i, $i = 1, \ldots, k$, are pivotal quantities, and substituting the observed values of the data vector $\mathbf{Y}$, we can obtain an expression for μ involving observed statistics and pivotal quantities only. This gives us a generalized pivotal quantity for μ. As an intermediate step during this process, we obtain an implicit expression for σ^2 as the solution to the equation

$$Q = \sum_{i=1}^k c_i(\bar{Y}_i - \bar{Y}_{\mathbf{c}})^2,$$

with $c_i = 1/(\sigma^2 + \tau_i^2 ss_i/SS_i)$. Lemma 1 gives some observations on the behavior of Q that assure us that σ^2 is uniquely determined by the aforementioned implicit equation. A proof of the lemma is given in the Appendix.

Lemma 1. Let $\mathbf{Y} \sim \mathrm{N}(\mu \mathbf{1}_k, \mathbf{G})$, where $\mathbf{G}$ is given in (5), and let $g_{\mathbf{Y},\tau^2}(\sigma^2)$ be as defined in (6). Then:

(a) With probability 1, $g_{\mathbf{Y},\tau^2}(\sigma^2)$ is a decreasing function of σ^2,

(b) With probability 1, $g_{\mathbf{Y},\tau^2}(\sigma^2)$ is convex in σ^2, and

(c) The maximum value of $g_{\mathbf{Y},\tau^2}(\sigma^2)$ is $g_{\mathbf{Y},\tau^2}(0)$, given by

$$g_{\mathbf{Y},\tau^2}(0) = \mathbf{Y}^t \left[\mathbf{G}_0^{-1} - \frac{\mathbf{G}_0^{-1}\mathbf{1}_k\mathbf{1}_k^t\mathbf{G}_0^{-1}}{\mathbf{1}_k^t\mathbf{G}_0^{-1}\mathbf{1}_k} \right] \mathbf{Y},$$

where $\mathbf{G}_0 = \mathrm{diag}(\tau_1^2, \tau_2^2, \ldots, \tau_k^2)$.

As observed in part (a) of the lemma, $g_{\mathbf{Y},\tau^2}(\sigma^2)$ is a monotonic decreasing function of σ^2, with probability 1. So, given a real number $q \geq 0$, there must exist a unique $a_q \geq 0$ such that $g_{\mathbf{Y},\tau^2}(a_q) = q$. We define the function $h_{\mathbf{Y},\tau^2}(\cdot)$ as

$$h_{\mathbf{Y},\tau^2}(q) = \begin{cases} a_q & \text{if } 0 \leq q \leq g_{\mathbf{Y},\tau^2}(0) \\ 0 & \text{otherwise.} \end{cases}$$

Define $Z = (\bar{Y}_\mathrm{w} - \mu)/\tau_0$ and $Q_i = SS_i/\sigma_i^2$, $i = 1, \ldots, k$. Then $Z \sim \mathrm{N}(0, 1)$, $Q_i \sim \chi^2_{n_i - 1}$, and Z, Q_i, and $Q \sim \chi^2_{k-1}$ are mutually independent, where $Q = g_{\mathbf{Y},\tau^2}(\sigma^2)$ is as defined in (6).

Let $\bar{y}_1, \ldots, \bar{y}_k$, $ss_1, \ldots, ss_k$ denote the observed values of $\bar{Y}_1, \ldots, \bar{Y}_k$, $SS_1, \ldots, SS_k$, and let $\mathbf{y} = (\bar{y}_1, \ldots, \bar{y}_k)^t$. Further, let

$$\mathbf{D} = (\bar{Y}_1, \ldots, \bar{Y}_k, SS_1, \ldots, SS_k)$$

be the vector of sufficient statistics and let

$$\mathbf{d} = (\bar{y}_1, \ldots, \bar{y}_k, ss_1, \ldots, ss_k)$$

be the vector of corresponding observed values. Observe that the vector $\mathbf{T}$, defined by

$$\begin{aligned} \mathbf{T} &= (ss_1\tau_1^2/SS_1, \ldots, ss_k\tau_k^2/SS_k) \\ &= \left(ss_1/(n_1Q_1), \ldots, ss_k/(n_kQ_k)\right) \\ &= (T_1, \ldots, T_k) \end{aligned} \tag{7}$$

is a random vector whose distribution is free of any unknown parameters. Also note that the observed value of $\mathbf{T}$ (obtained by replacing the SS_i's with the corresponding ss_i's) is $\tau^2 = (\tau_1^2, \tau_2^2, \ldots, \tau_k^2)$. Define

$$W_i = \frac{1}{h_{\mathbf{y},\mathbf{T}}(g_{\mathbf{Y},\tau^2}(\sigma^2)) + \tau_i^2 ss_i/SS_i}. \tag{8}$$

Like the T_i, note that W_i, $i = 1, \ldots, k$, are also random variables whose distribution is free of any model parameters. Furthermore, when the observed statistics $\mathbf{d}$ are substituted in $h_{\mathbf{y},\mathbf{T}}(g_{\mathbf{Y},\tau^2}(\sigma^2))$, it reduces to σ^2. Thus, when the observed data values $\mathbf{d}$ are substituted for $\mathbf{D}$ in W_i, they reduce to $w_i = 1/(\sigma^2 + \tau_i^2)$. Let $\mathbf{W} = (W_1, \ldots, W_k)$. Using θ to denote the vector $(\mu, \sigma^2, \sigma_1^2, \ldots, \sigma_k^2)$ of model parameters, we now define a generalized pivotal quantity (GPQ) for μ, denoted by $R(\mathbf{D}; \mathbf{d}, \theta)$, as follows:

$$\begin{aligned} &R(\mathbf{D}; \mathbf{d}, \theta) \\ &\quad = \bar{y}_\mathrm{W} - \left(\frac{\bar{Y}_\mathrm{W} - \mu}{\tau_0}\right)\left(\sum_{i=1}^k W_i\right)^{-1/2} \\ &\quad = \bar{y}_\mathrm{W} - Z \times \left(\sum_{i=1}^k \frac{1}{h_{\mathbf{y},\mathbf{T}}(g_{\mathbf{Y},\tau^2}(\sigma^2)) + T_i}\right)^{-1/2}. \end{aligned} \tag{9}$$

From (9), it is clear that $R(\mathbf{D}; \mathbf{d}, \theta)$ has a distribution free of model parameters. Also, note that when the observed data $\mathbf{d}$ are substituted for $\mathbf{D}$ in $R(\mathbf{D}; \mathbf{d}, \theta)$, it reduces to $R(\mathbf{d}; \mathbf{d}, \theta) = \mu$, which is free of all nuisance parameters. Hence the requirements for $R(\mathbf{D}; \mathbf{d}, \theta)$ to be a GPQ given by Weerahandi (1993) are satisfied. Therefore, a $(1 - \alpha)$ GCI for μ is obtained as

$$L \leq \mu \leq U,$$

where $L = R_{\alpha/2}$ is the $\alpha/2$ percentile and $U = R_{1-\alpha/2}$ is the $1 - \alpha/2$ percentile of the distribution of $R(\mathbf{D}; \mathbf{d}, \theta)$. In actual applications, when closed-form expressions for the required percentiles are unavailable, they may be estimated by simulating the distribution of $R(\mathbf{D}; \mathbf{d}, \theta)$. A single realization of $R(\mathbf{D}; \mathbf{d}, \theta)$ may be generated as follows:

1. Generate a standard normal random deviate Z.
2. For $1 \leq i \leq k$, generate Q_i distributed as a chi-squared random variable with $n_i - 1$ degrees of freedom.
3. Generate Q, a chi-squared random deviate with $k - 1$ degrees of freedom.
4. Calculate T_i, $i = 1, \ldots, k$, as in (7).
5. Calculate

$$g_{\mathbf{y},\mathbf{T}}(0) = \mathbf{y}^t \left[\mathbf{G}_\mathbf{T}^{-1} - \frac{\mathbf{G}_\mathbf{T}^{-1}\mathbf{1}_k\mathbf{1}_k^t\mathbf{G}_\mathbf{T}^{-1}}{\mathbf{1}_k^t\mathbf{G}_\mathbf{T}^{-1}\mathbf{1}_k} \right] \mathbf{y},$$

where $\mathbf{G}_\mathbf{T} = \mathrm{diag}(T_1, T_2, \ldots, T_k)$. A convenient computational form for $g_{\mathbf{y},\mathbf{T}}(0)$ is given by

$$\sum_{i=1}^k \frac{(\bar{y}_i - \bar{y})^2}{ss_i/(n_iQ_i)} - \frac{(\sum_{i=1}^k n_iQ_i(\bar{y}_i - \bar{y})/ss_i)^2}{\sum_{i=1}^k n_iQ_i/ss_i},$$

where $\bar{y} = \sum_{i=1}^k \bar{y}_i/k$.

6. If $0 \leq Q \leq g_{\mathbf{y},\mathbf{T}}(0)$, then find a_Q such that $g_{\mathbf{y},\mathbf{T}}(a_Q) = Q$; otherwise, let $a_Q = 0$. Set $h_{\mathbf{y},\mathbf{T}}(Q) = a_Q$. For computational purposes, a convenient expression for $g_{\mathbf{y},\mathbf{T}}(a_Q)$ is given by

$$\sum_{i=1}^k \frac{(\bar{y}_i - \bar{y})^2}{a_Q + ss_i/(n_iQ_i)} - \frac{1}{\sum_{i=1}^k 1/(a_Q + ss_i/(n_iQ_i))} \times \left(\sum_{i=1}^k \frac{\bar{y}_i - \bar{y}}{a_Q + ss_i/(n_iQ_i)}\right)^2.$$

7. Calculate $W_i = 1/(a_Q + ss_i/(n_iQ_i))$, $i = 1, \ldots, k$.
8. Calculate $\bar{y}_\mathrm{W} = (\sum_{i=1}^k W_i\bar{Y}_i)/(\sum_{i=1}^k W_i)$.
9. Calculate $R(\mathbf{D}; \mathbf{d}, \theta)$ as in (9).

By generating K (with K a large positive integer) independent realizations of $R(\mathbf{D}; \mathbf{d}, \theta)$, we can estimate the required percentiles of the distribution of $R(\mathbf{D}; \mathbf{d}, \theta)$. Let R_i denote the ith-order statistic of $R(\mathbf{D}; \mathbf{d}, \theta)$. We then take $L = R_{\lfloor K\alpha/2 \rfloor}$ and $U = R_{\lceil K(1-\alpha/2) \rceil}$, where $\lfloor \cdot \rfloor$ is the *floor* function and $\lceil \cdot \rceil$ is the *ceiling* function. In our examples we use $K = 10{,}000$ and $\alpha = .05$, so that $L = R_{250}$ and $U = R_{9750}$. The required GCI for μ is $[L, U]$.

Special Cases. If it is known that σ_i^2, $i = 1, \ldots, k$, are all equal, then some simple modifications need to be made to the GPQ. Let σ_e^2 denote the common value of σ_i^2. Define $SS_e = \sum_{i=1}^k SS_i$, $Q_e = SS_e/\sigma_e^2$, and $n_e = \sum_{i=1}^k (n_i - 1)$. Then $Q_e \sim \chi^2_{n_e}$ and Q, Q_e, and Z are mutually independent. The quantities T_i defined in (7) and W_i defined in (8) need to be redefined, as follows. For $i = 1, \ldots, k$, let

$$T_i = \frac{\sigma_e^2}{n_i} \frac{ss_e}{SS_e}$$

and

$$W_i = \frac{1}{h_{\mathbf{y},\mathbf{T}}(g_{\mathbf{Y},\tau^2}(\sigma^2)) + T_i}, \tag{10}$$

where $\tau^2 = (\sigma_e^2/n_1, \ldots, \sigma_e^2/n_k)$ and ss_e is the observed value of SS_e. In particular, if the n_i's are all equal in (10), so that we have the usual homoscedastic one-way random model with balanced data, then the W_i's are all equal and the τ_i^2's are all equal with $\tau_i^2 = \sigma_e^2/n$, where n is the common value of the n_i's. In this case a confidence interval for μ can be obtained using a pivot statistic that has a t-distribution with $k-1$ degrees of freedom. To see this, let $\bar{Y} = \sum_{i=1}^k \bar{Y}_i/k = \sum_{i=1}^k \sum_{j=1}^n Y_{ij}/(nk)$ and $SS_b = \sum_{i=1}^k (\bar{Y}_i - \bar{Y})^2$. It is readily verified that

$$g_{\mathbf{Y},\tau^2}(\sigma^2) = (\sigma^2 + \sigma_e^2/n)^{-1} SS_b$$

and

$$R(\mathbf{D}; \mathbf{d}, \theta) = \bar{y} - (\bar{Y} - \mu)\sqrt{ss_b}/\sqrt{SS_b},$$

where $\bar{y}$ and ss_b denote the observed values of the corresponding random variables. It is clear that a confidence interval for μ can now be obtained using the percentiles of the pivot statistic $\sqrt{k(k-1)}(\bar{Y} - \mu)/\sqrt{SS_b}$. Because $\bar{Y} \sim \mathrm{N}(\mu, (\sigma^2 + \sigma_e^2/n)/k)$ and, independently, $SS_b/(\sigma^2 + \sigma_e^2/n) \sim \chi^2$ with $k-1$ degrees of freedom, we conclude that $\sqrt{k(k-1)}(\bar{Y} - \mu)/\sqrt{SS_b}$ follows a t-distribution with $k-1$ degrees of freedom.

If it is known that various subsets of σ_i^2 are equal, then appropriately pooled estimates of σ_i^2 should be used in the definition of the GPQ. We omit the details.

2.1 Example

For illustration, we use the data reported by Rukhin and Vangel (1998) from an interlaboratory study involving the measurement of trace metal concentrations in oyster tissue samples (see Willie and Berman 1995). In that study, homogeneous samples of oyster tissue were distributed to 28 laboratories, and each laboratory made replicate measurements of several trace metals. Rukhin and Vangel (1998) used only arsenic concentration (mg/kg) data to demonstrate their approach to computing a confidence interval for μ, the true mean arsenic concentration. Twenty-seven of the laboratories made five replicate measurements of the tissue sample, whereas the remaining laboratory made only two replicate measurements. The set of sufficient statistics for these data were reported in table 1 of Rukhin and Vangel (1998).

The large-sample maximum likelihood (ML) procedure of Rukhin and Vangel [1998, eq. (19)], for a nominal coverage probability of 95%, yields the interval $(12.709, 13.741)$ for μ. The corresponding GCI, using 10,000 Monte Carlo runs to estimate the percentiles of the GPQ, is found to be $(12.683, 13.772)$. Rukhin and Vangel estimated the actual coverage probability of their interval to be about 93%. The coverage probability for the GCI, based on simulation results reported in Section 2.3, is estimated to be almost exactly 95%. Therefore, it is not surprising that the GCI is somewhat wider.

2.2 Coverage Probability of the Generalized Confidence Interval

Two distinct probability spaces are needed in the following discussion. We use $\mathcal{P}_{\mathbf{Q},Q,Z}$ (or $\mathcal{P}$, for short) to denote probabilities calculated with respect to the joint distribution of $(Q_1, \ldots, Q_k, Q, Z)$ (regarding the sufficient statistics as fixed constants) and use $P_{\mathbf{d}}$ (P for short) to denote probabilities calculated with respect to the joint distribution of the data values $\mathbf{d}$ or, equivalently, with respect to the joint distribution of the sufficient statistics $\{\bar{y}_1, \ldots, \bar{y}_k, ss_1, \ldots, ss_k\}$, now regarding them as random variables. With this convention, we observe that $L \le \mu \le U \Leftrightarrow \alpha/2 \le \mathcal{P}(R(\mathbf{D}, \mathbf{d}, \theta) \le \mu) \le 1 - \alpha/2$. From this, it is easily deduced that

$$P[L \le \mu \le U] = P\left[\frac{\alpha}{2} \le 1 - E_{Q,Q_1,\ldots,Q_k}\left[\Phi\left((\bar{y}_{\mathbf{W}} - \mu)V^{1/2}\right)\right] \le 1 - \frac{\alpha}{2}\right],$$

where $V = \sum_{i=1}^k 1/(h_{\mathbf{y},\mathbf{T}}(Q) + ss_i/(n_i Q_i))$ and $\Phi(\cdot)$ is the standard normal cdf. An analytical evaluation of this probability is not tractable, and hence we resort to its evaluation using Monte Carlo methods.

2.3 Simulation Study for Model 1

We examined the frequentist coverage probability of the proposed GCI using statistical simulation. It is easy to see that the coverage probability of a GCI depends not on the value of μ, but only on the ratios $\sigma_i/\sigma_{\min}$ and $\sigma/\sigma_{\min}$, where $\sigma_{\min} = \min\{\sigma_i | 1 \le i \le k\}$. Hence, without loss of generality, we assumed that $\mu = 0$ and $\sigma_{\min} = 1$ in the simulations. With this convention, we used the following grid of values for the unknown parameters in the simulation study:

1. $\mu = 0$.
2. $k = \{2, 5, 11, 21\}$.
3. $\sigma_1^2 = \sigma_{\min}^2 = 1$.
4. $\sigma_k^2 = \sigma_{\max}^2 = \max\{\sigma_i^2 | 1 \le i \le k\} = \{1, 2, 3, 4\}$.
5. Given that $\sigma_1^2 = 1$ and knowing the value of σ_k^2, the remaining σ_i^2, $i = 2, \ldots, k-1$ (if any), were chosen by taking equally spaced values in the interval $[1, \sigma_k^2]$.
6. The values for σ^2 were taken to be 0, 1/4, 1/2, 1, $(1 + \sigma_k^2)/2$, σ_k^2, $2\sigma_k^2$, and $4\sigma_k^2$.

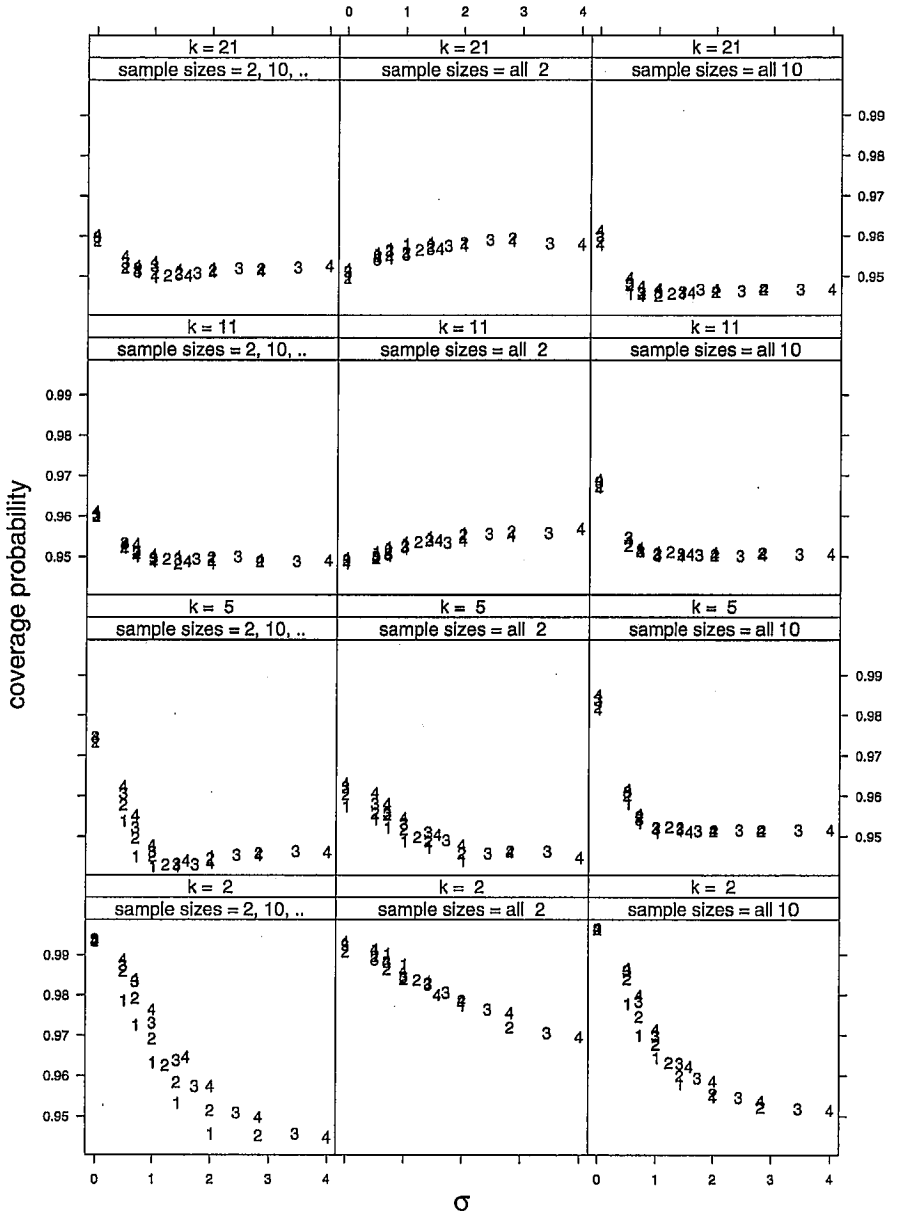

Figure 1. Plots of the Coverage Probabilities of the GCI for Model 1.

7. Given k, three patterns of n_i were used: $n_i = 10$ for all i, $n_i = 2$ for all i, and $n_i = \{2, 10, 2, 10, \ldots\}$.

In total, there were 360 combinations of simulation parameters. For each given combination of model parameters considered in the simulation study, we generated the vector of observed values $\mathbf{d} = (\bar{y}_1, \ldots, \bar{y}_k, ss_1, \ldots, ss_k)$ according to $\bar{y}_i \sim \mathrm{N}(0, \sigma^2 + \sigma_i^2/n_i)$ and $ss_i \sim \sigma_i^2 \chi^2_{n_i-1}$. Based on this $\mathbf{d}$, we generated 10,000 independent realizations of the GPQ R using the procedures given in Section 2, then obtained the 95% GCI for μ. We repeated this process 5,000 times, and recorded the percentage of times that the GCI contained 0. The simulation results are best displayed in graphical form, due to the large quantity of data. Figure 1 shows these results.

Each panel in Figure 1 plots the empirical simulated coverage probability of the interval (y axis) versus the value of σ (x axis) for a specific combination of k and n_i pattern. The plotting symbols "1," "2," "3," and "4" are designated for cases with $\sigma_k^2 = 1$, 2, 3, and 4. An examination of Figure 1 shows that the coverage probability of the GCI is very satisfactory; for most cases, it is very close to the nominal value of .95. When σ^2 is very small, the interval is quite conservative. For comparison purposes, we also obtained the coverage of approximate confidence interval of Rukhin and Vangel (1998); see their equation (19). Their interval is a large-sample (i.e., large k) interval. Not surprisingly, the Rukhin and Vangel (1998) interval is quite liberal when k is small. Specifically, the ranges of the coverage for the parameter combinations considered here are .45–.61, .71–.84, .86–.91, and .91–.93 for $k = 2$, 5, 11, and 21. Because of this, we did not perform any confidence interval width comparisons.

3. A GENERALIZED CONFIDENCE INTERVAL FOR μ UNDER MODEL 2

As in Section 2, the measurement model is

$$Y_{ij} = \mu + b_i + e_{ij}, \qquad j = 1, \ldots, n_i, i = 1, \ldots, k,$$

but now a different set of assumptions are made about the biases b_i. Based on a detailed knowledge of the measurement methods, it is assumed that for $i = 1, \ldots, k$, the magnitude of the bias b_i is bounded by a positive constant M_i. This model, discussed by Eberhardt et al. (1989), is a special case of a more general class of *approximately linear models* presented by Sacks and Ylvisaker (1978), who used a minimax approach to estimate linear functions of model parameters. Following the suggestion of Sacks and Ylvisaker (1978), Eberhardt et al. (1989) considered an estimate of μ of the form

$$\sum_{i=1}^{k} c_i \bar{Y}_i,$$

where the c_i's are *weights* that minimize the objective function,

$$\Psi(\mathbf{c}, \boldsymbol{\tau}) = \sum_{i=1}^{k} c_i^2 \tau_i^2 + \left(\sum_{i=1}^{k} |c_i| M_i \right)^2. \tag{11}$$

This objective function arises from the use of the minimax criterion; that is, the c_i are to be determined so that the resulting consensus value has the smallest maximum mean squared error over the range of the b_i. Here $\mathbf{c} = (c_1, \ldots, c_k)$ and $\boldsymbol{\tau}^2 = (\tau_1^2, \ldots, \tau_k^2)$. Eberhardt et al. (1989) provided an algorithm for computing the optimum weights c_i, which depend on the values of τ_i. Consequently, they suggested using $\hat{\tau}_i^2 = s_i^2/n_i$ in place of τ_i^2 in the objective function. Along with a point estimate of μ, they also proposed a confidence interval procedure that they derived using heuristic arguments.

Here we focus on an interval estimate for μ under Model 2, using the method of GCIs. But first we point out some interesting features of Model 2 that must be taken into account when attempting to make inferences about μ. For the record, we list these features in the following proposition, whose proof is elementary.

Proposition 1. Under the assumptions of Model 2, the following statements hold:

1. The parameter μ is not identifiable.
2. Let $\lambda = \max\{\mu_i - M_i | 1 \le i \le k\}$ and $\omega = \min\{\mu_i + M_i | 1 \le i \le k\}$. Then λ and ω are identifiable.
3. $\lambda \le \mu \le \omega$.
4. $\omega - \lambda \le 2 \min\{M_i | 1 \le i \le k\}$.

One consequence of the foregoing proposition is that even if μ_i $(i = 1, \ldots, k)$ are known without error, the best possible statement that one can make about μ under Model 2 is that μ is between λ and ω. Therefore, when μ_i are estimated based on data, it makes sense to seek a lower confidence bound L for λ and an upper confidence bound U for ω, so that the interval $[L, U]$ may be used as a confidence interval for μ with a nominal confidence level equal to $1 - \alpha$. In this section we develop such an interval using the method of GPQs. We investigate the frequentist coverage properties of this interval via statistical simulation in Section 3.5.

Proposition 1 has other implications as well. For instance, although the model assumptions imply that λ is less than or equal to ω, it is possible that the bias bounds M_i associated with the different laboratories (or methods) are inconsistent with one another and that this may result in a situation where $\lambda > \omega$; that is, the model assumptions fail to define a valid parameter space for μ.

Consider a simple example. Suppose that we have estimates of μ from two laboratories. Suppose also that the true mean for the first laboratory is $\mu_1 = 5$ and the bias bound M_1 is 1. Then μ must lie in the interval between 4 and 6. Now suppose that the second laboratory has true mean μ_2 equal to 7 and the bias bound M_2 is .5. This implies that μ must lie between 6.5 and 7.5. Clearly, the bias bounds provided by the two laboratories are in conflict with one another.

It is thus clear that under Model 2, before proceeding with making inferences about μ, it is advisable to examine whether or not the data are consistent with the requirement that $\lambda \le \omega$. The method of generalized p values (Tsui and Weerahandi 1989) may be applied to develop a test of the null hypothesis $H_0 : \lambda \le \omega$ versus the alternative $H_a : \lambda > \omega$ (or, alternatively, $H_0 : \lambda \ge \omega$ versus $H_a : \lambda < \omega$, if appropriate). We discuss this test in Section 3.3. Eberhardt et al. (1989) appeared to not address this issue.

The minimax-estimator-based confidence intervals of Eberhardt et al. (1989) have certain undesirable features that are worth noting. If we denote the minimax estimator of μ by $\tilde{\mu}$, then the expected value ξ of $\tilde{\mu}$ is

$$\xi = \mu + \sum_{i=1}^{k} c_i b_i,$$

where b_i is the bias of laboratory i. In general, therefore, $\tilde{\mu}$ is a biased estimator of μ. As the variances τ_i^2 of $\bar{Y}_i$ approach 0, $\tilde{\mu}$ converges to ξ, and so the minimax estimator $\tilde{\mu}$ is inconsistent. This is to be expected, because μ is not identifiable under Model 2. However, it is interesting to note that the minimax estimator may not even converge to a value in the interval $[\lambda, \omega]$, although the assumptions of Model 2 imply $\lambda \le \mu \le \omega$. We illustrate this by considering the case where $k = 2$.

Let ψ and δ $(\delta > 0)$ be fixed real numbers. Let $\mu_1 = \psi$, $\mu_2 = \psi + 4\delta$, $\tau_1 = 2\delta$, $\tau_2 = \delta$, $M_1 = 3\delta$, and $M_2 = 2\delta$. The optimal weights for the minimax estimator are found to be $c_1 = 0$ and $c_2 = 1$. Thus $\tilde{\mu} = \bar{Y}_2$ is the minimax estimator, and as the sample size for method 2 approaches infinity, it converges to μ_2. Here we have $\lambda = \psi + 2\delta$ and $\omega = \psi + 3\delta$; thus μ is in the interval $[\psi + 2\delta, \psi + 3\delta]$. However, the minimax estimator converges to $\mu_2 = \psi + 4\delta$. Moreover, it is easy to see that the Eberhardt et al. confidence interval converges to the interval $[\mu_2 - M_2, \mu_2 + M_2] = [\psi + 2\delta, \psi + 6\delta]$.

Now consider the case where $\mu_2 = \psi + 6\delta$ but the remaining parameters are as given before. This is a situation where $\lambda > \omega$; that is, the bias bounds M_1 and M_2 are inconsistent with one another. Nevertheless, the minimax estimator will converge to $\mu_2 = \psi + 6\delta$, and the associated interval will converge to the interval $[\psi + 4\delta, \psi + 8\delta]$!

We now present generalized pivotal quantities for λ and ω.

3.1 Generalized Pivots for λ and ω

We define

$$R_\lambda^* = \max\left\{ \bar{y}_i - (\bar{Y}_i - \mu_i)\sqrt{ss_i/SS_i} - M_i | i = 1, \ldots, k \right\} \tag{12}$$

and

$$R_\omega^* = \min\left\{ \bar{y}_i - (\bar{Y}_i - \mu_i)\sqrt{ss_i/SS_i} + M_i | i = 1, \ldots, k \right\}. \tag{13}$$

We further define

$$R_\lambda = \begin{cases} R_\lambda^* & \text{if } R_\lambda^* \le R_\omega^* \\ (R_\omega^* + R_\lambda^*)/2 & \text{otherwise} \end{cases} \quad (14)$$

and

$$R_\omega = \begin{cases} R_\omega^* & \text{if } R_\lambda^* \le R_\omega^* \\ (R_\omega^* + R_\lambda^*)/2 & \text{otherwise.} \end{cases} \quad (15)$$

It is easily verified that R_λ and R_ω satisfy the conditions required for them to be GPQs for λ and ω. In addition to the fact that, by definition, $R_\omega - R_\lambda \ge 0$, it is easily verified that

$$R_\omega - R_\lambda \le 2\min\{M_i \mid 1 \le i \le k\}$$

and that there is in fact a positive probability that equality is achieved.

Let $R_{\lambda,\gamma}$ and $R_{\omega,\gamma}$ denote the γth percentile of the distribution of R_λ and R_ω. Then $L = R_{\lambda,\alpha/2}$ is a generalized $1-\alpha/2$ lower confidence bound for λ, and $U = R_{\omega,1-\alpha/2}$ is a generalized $1-\alpha/2$ upper confidence bound for ω. In particular, $[L, U]$ is a GCI for μ with approximate coverage probability $1-\alpha$. The percentiles of R_λ and R_ω may be obtained using a Monte Carlo approach, as explained next.

3.2 Estimation of $R_{\lambda,\gamma}$ and $R_{\omega,\gamma}$

A single realization of (R_λ, R_ω) may be generated as follows:

1. Generate independent realizations $t_1, \ldots, t_k$, where t_i, $1 \le i \le k$, is distributed as a Student' t random variable with $n_i - 1$ degrees of freedom.
2. Compute
$$R_\lambda^* = \max\left\{\bar{y}_i - t_i\sqrt{ss_i/n_i(n_i-1)} - M_i \mid i = 1, \ldots, k\right\}.$$
3. Compute
$$R_\omega^* = \min\left\{\bar{y}_i - t_i\sqrt{ss_i/n_i(n_i-1)} + M_i \mid i = 1, \ldots, k\right\}.$$
4. If $R_\lambda^* \le R_\omega^*$, then set $R_\lambda = R_\lambda^*$ and $R_\omega = R_\omega^*$. Otherwise, set $R_\lambda = R_\omega = (R_\omega^* + R_\lambda^*)/2$.

By generating K (with K a large positive integer) independent realizations of (R_λ, R_ω), we can estimate the required percentiles of the distribution of R_λ and R_ω. Let $R_{\lambda,i}$ and $R_{\omega,i}$ denote the ith-order statistic of R_λ and R_ω. We then take $L = R_{\lambda,\lfloor K\alpha/2\rfloor}$ and $U = R_{\omega,\lceil K(1-\alpha/2)\rceil}$. In our examples we use $K = 10{,}000$ and $\alpha = .05$, so that $L = R_{\lambda,250}$ and $U = R_{\omega,9750}$. The GCI for μ is taken to be $[L, U]$.

Remark. It is clear that R_λ^* itself satisfies the requirements for it to be a GPQ for λ. Likewise, R_ω^* satisfies the requirements for it to be a GPQ for ω. However, the joint distribution of $(R_\lambda^*, R_\omega^*)$ allows for the possibility $R_\lambda^* > R_\omega^*$ which is inconsistent with the theoretical requirement that $\lambda \le \omega$. This possibility is eliminated by the use of R_λ and R_ω in place of R_λ^* and R_ω^*. Consider the reparameterization given by $\phi_1 = (\omega + \lambda)/2$ and $\phi_2 = (\omega - \lambda)/2$, and the corresponding GPQs $R_{\phi_1} = (R_\omega + R_\lambda)/2$ and $R_{\phi_2} = (R_\omega - R_\lambda)/2$. Our proposed definitions for R_λ and R_ω follow if we agree to truncate the distribution of R_{ϕ_2} to the positive real axis, because $\omega - \lambda$ is required to be nonnegative.

3.3 A Generalized Test of $H_0 : \lambda \le \omega$

A generalized test (Tsui and Weerahandi 1989) of the hypothesis $H_0 : \lambda \le \omega$ versus $H_a : \lambda > \omega$ may be constructed as follows. Let $\Delta = \omega - \lambda$ and $R_\Delta = R_\omega^* - R_\lambda^*$. It is easily verified that R_Δ is a GPQ for Δ. Let $U = R_{\Delta,1-\alpha}$ denote the $(1-\alpha)$th percentile of the distribution of R_Δ. Then U is an upper $(1-\alpha)$-level confidence bound for Δ. Then reject the null hypothesis $H_0 : \lambda \le \omega$ in favor of $H_a : \lambda > \omega$ if $U < 0$. As usual, $R_{\Delta,1-\alpha}$ may be estimated using a Monte Carlo approach.

Remark. Under Model 2, it was pointed out earlier that the true value μ is not identifiable. However, λ and ω are identifiable. Consequently, it is reasonable to consider the ML estimators of λ and ω and confidence intervals for these parameters obtained by inverting the appropriate generalized likelihood ratio tests. The ML estimates of μ_i, $i = 1, \ldots, k$, are obtained by minimizing

$$\sum_{i=1}^{k} n_i \log\left(SS_i + n_i(\bar{Y}_i - \mu_i)^2\right) + C',$$

where C' is a constant free of μ_i and σ_i^2. The minimization is to be carried out under the constraints $|\mu_i - \mu_j| \le M_{ij}$ (where $M_{ij} = M_i + M_j$). An analytical solution to this problem appears to be intractable except when $k = 2$ or 3. We do not pursue this further in this article.

3.4 Examples

We illustrate the methods discussed in this section using the examples analyzed by Eberhardt et al. (1989). The data are chemical measurements for certification of NIST Standard Reference Material (SRM) 1549.

In the first example, the data are determinations of selenium in nonfat milk powder made by four different analytical methods. The summary statistics and the bias bounds M_i for each method are listed in Table 1.

For this example, we conducted a test of $H_0 : \lambda \le \omega$ versus $H_a : \lambda > \omega$. With $\Delta = \omega - \lambda$, we obtain an upper 95% generalized confidence bound for Δ to be $U = -.824$. We used 1,000,000 realizations of the GPQ R_Δ to obtain U. Based on this result, we reject the null hypothesis and conclude that the specification of the M_i, $i = 1, \ldots, 4$, is inconsistent. One would question the attempt to make inferences about μ based on this inconsistent specification of the M_i.

The data for the second example are determinations of zinc levels in nonfat milk powder made by four different analytical methods. The summary statistics and the bias bound for each method are given in Table 2.

Unlike in the first example, here the intersection of the four intervals $\bar{y}_i \pm M_i$ is not empty. The nominal 95% GCI for μ, using 10,000 simulation runs, is (46.04, 47.56). The minimax

Table 1. Selenium (ng/g) in Nonfat Milk Powder

Method	n_i	$\bar{y}_i$	s_i	M_i
1	8	105.0	9.258	2.1
2	12	109.75	4.555	1.1
3	14	109.5	1.652	1.1
4	8	113.25	5.8	.6

Table 2. Zinc ($\mu g/g$) in Nonfat Milk Powder

Method	n_i	$\bar{y}_i$	s_i	M_i
1	8	45.21	1.68	5.880
2	12	46.63	.47	.466
3	22	46.26	.82	.927
4	8	47.05	1.44	.230

solution of Eberhardt et al. (1989) assigns zero weights to both methods 1 and 3 and produces the interval (45.94, 47.64) for μ. The GCI, using data only from methods 2 and 4, is (46.02, 47.58), which is almost identical to the confidence interval obtained with full data. Note that in this example, $\bar{y}_1 \pm M_1$ contains the remaining three $\bar{y}_i \pm M_i$, and $\bar{y}_3 \pm M_3$ contains $\bar{y}_2 \pm M_2$. That is, with the specified M_i, methods 1 and 3 appear to offer very little additional information for determining μ over that provided by methods 2 and 4.

3.5 Simulation Study for Model 2

We use a setup similar to that for Model 1 in Section 2.3 to determine the grid of unknown parameters for use in the simulation study for Model 2. Specifically, we use the following:

1. $\mu = 0$.
2. $k = \{3, 6, 11\}$.
3. $\sigma_1 = \sigma_{\min} = 1$.
4. $\sigma_k = \sigma_{\max} = \{1, 2, 4\}$.
5. Given that $\sigma_1 = 1$ and knowing the value of σ_k, the remaining σ_i, $i = 2, \ldots, k-1$ (if any) were chosen by taking equally spaced values between 1 and σ_k.
6. The values of μ_i were generated using a uniform distribution between $-\delta$ and δ with $\delta = \{.5, 1, 2\}\sigma_{\max}$.
7. The values of M_i were generated using a uniform distribution between $1.1\mu_i$ and $3.6\mu_i$.
8. Given k, three patterns of n_i were used: $n_i = 10$ for all i, $n_i = 5$ for all i, and $n_i = \{10, 5, 10, 5, \ldots\}$.

Because the simulation results depend on the specific configuration of intervals $\mu_i \pm M_i$, and because there is no convenient canonical representation for such parameter specifications, we used only a small number of configurations in the simulation. Figure 2 plots the nine configurations of $\mu_i \pm M_i$ and values of σ_i used in the simulation for $k = 6$. The vertical dotted lines in each panel indicate the values of λ and ω, and the numbers above the interval are the values of σ_i.

In total, there were 81 combinations of simulation parameters. For each combination of simulation parameters, we generated 10,000 intervals with a 95% nominal confidence level using the GCI and the minimax methods. Because $\lambda \le \mu \le \omega$,

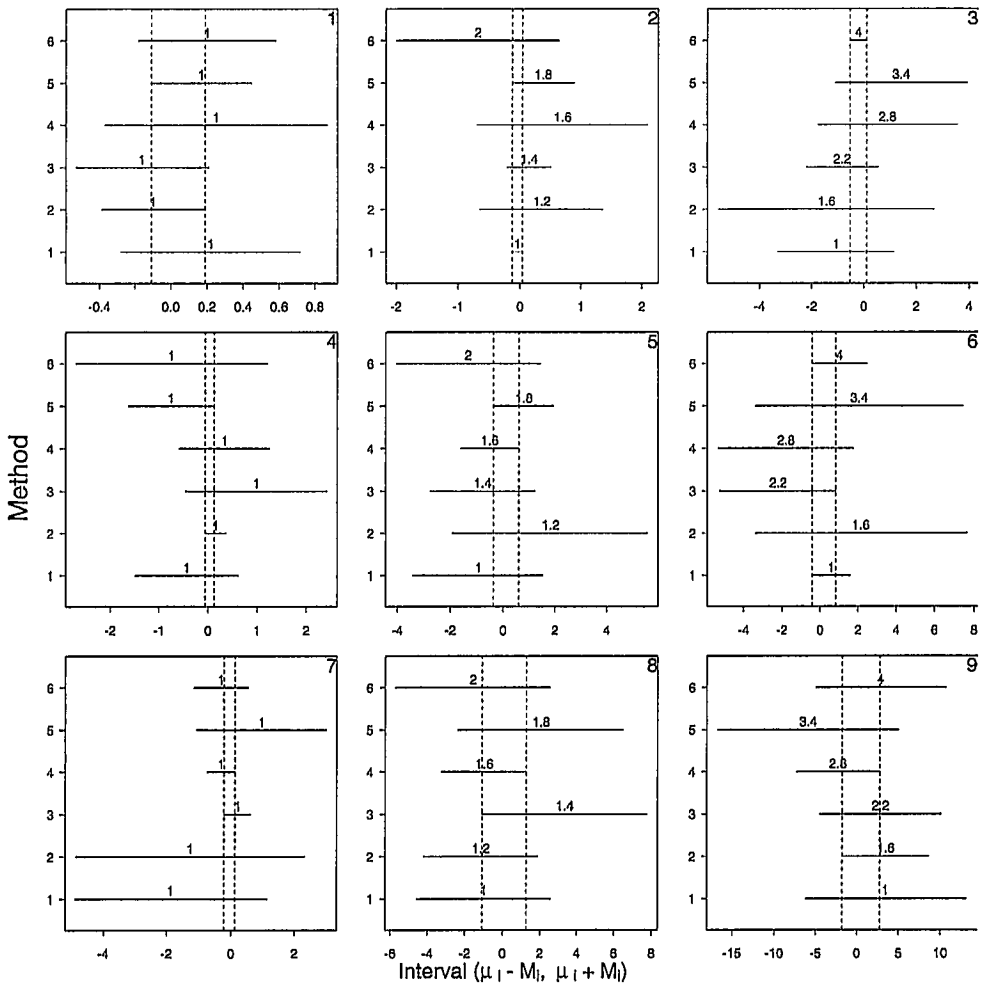

Figure 2. Plots of Intervals $\mu_i \pm M_i$ and σ_i Used in the Simulation ($k = 6$).

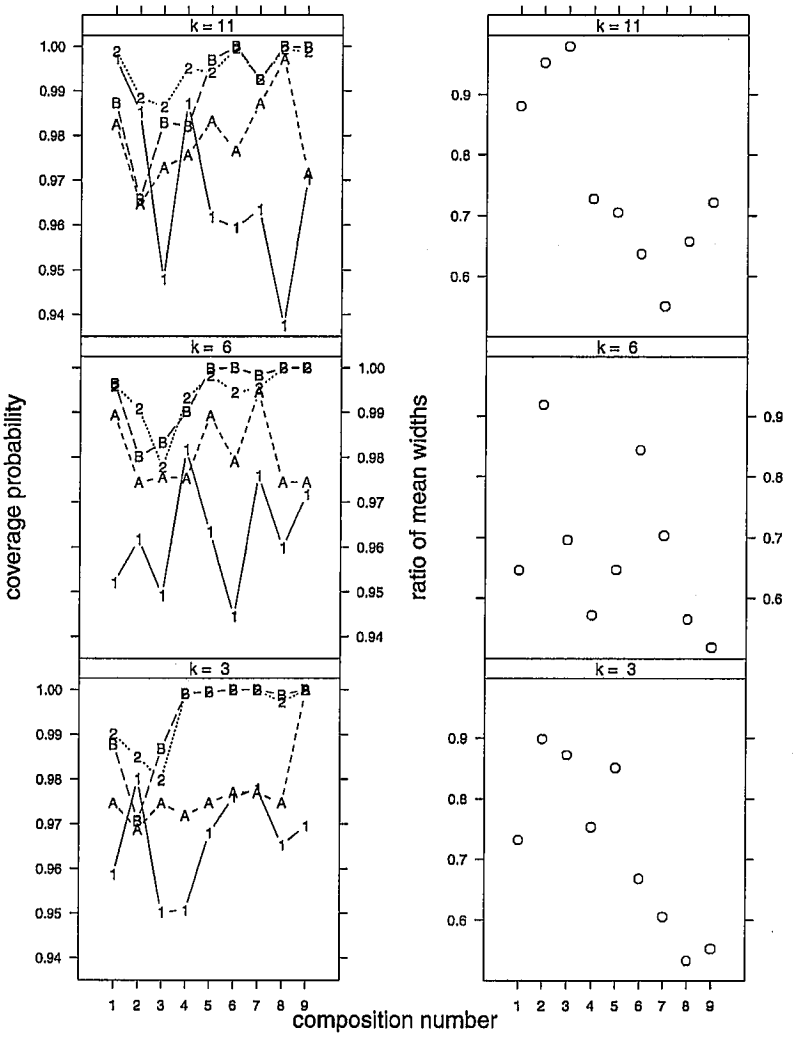

Figure 3. Plots of the Minimum and Maximum Empirical Two-Sided Coverage Probabilities of the Bounds for 101 Equally Spaced Values of μ in the Interval $[\lambda, \omega]$ From a Generalized Confidence Interval (minimum = "1," maximum = "2") and a Minimax Interval (minimum = "A," maximum = "B") Under Model 2, Corresponding to the 9 Configurations Shown in Figure 2.

we considered 101 equally spaced values of μ_0 in the interval $[\lambda, \omega]$. We calculated the proportion of the time that $[L, U]$ contained μ_0 for each method, and also obtained the mean widths of the intervals. We show the results only for $n_i = 10$ in Figure 3; the results for other two n_i patterns are similar.

The left three panels in Figure 3 plot the empirical simulated minimum (plotting symbol "1") and the maximum (plotting symbol "2") of the 101 empirical coverage probabilities (y axis) of the GCI for μ_0 versus the configuration number (x axis) for different values of k. For the minimax interval, these are denoted by the plotting symbols "A" and "B." The right three panels plot the ratio of the mean width of the GCI to the mean width of the minimax interval.

An examination of Figure 3 shows that the coverage probabilities of the lower and upper bounds of the GCI under Model 2, although conservative on some occasions, are generally adequate. It also indicates that, whenever the minimax interval is obtained by using mostly information from a single method and the bias bound for this method is close to 0, the resulting intervals are less conservative relative to other situations. When $k = 6$, configuration 2 (see Fig. 2) is an instance where this phenomenon occurs. The minimax interval derives most of the information from method 1 (with an average weight of 78% over the simulations) which also has the smallest bias bound. The resulting intervals are noticeably less conservative. This same phenomenon is also seen for configuration 3, where method 6, which has a small bias bound, contributes most of the information to the minimax interval.

Putting aside the drawbacks of the minimax intervals described earlier, we also note that the GCI appears to be less conservative than the minimax interval and, in most cases examined, has shorter expected widths. Additionally, when used in the context of interlaboratory trials, minimax intervals tend to not use information from all of the participating laboratories. For example, for $k = 11$, the numbers of laboratories/methods that received nonzero weights ranged from one to six in the nine configurations that we used in the simulation. Using zero weights in estimation of the mean is equivalent to deleting measurements from laboratories or methods that are less reliable. This is quite appropriate in preparing certifications of standard reference materials. However, this may not be practical for analyzing data from interlaboratory trials, because every laboratory wants its data used.

4. A GENERALIZED CONFIDENCE INTERVAL FOR μ UNDER MODEL 3

Under Model 3, the b_i are assumed to have known distributions F_i. These may be considered informative prior distributions on the bias constants b_i that are postulated based on scientific judgment. This is in fact required by policy established by the ISO GUM and is followed by most international standards laboratories. Such informative prior distributions are referred to as *type-B* distributions in the metrology literature. Here we assume that b_i are independent random variables with known distributions F_i. Generally, F_i is assumed to be either normal or uniform, although some other distributions are also discussed in the GUM. In our discussion we do not restrict F_i in any way other than that they be fully specified. Although in practice, it is generally assumed that the b_i are mutually independent, all we need for our procedure to be implementable is that the joint distribution of $(b_1, \ldots, b_k)$ be fully specified.

Under the circumstances, conditional on the b_i, we have that $U_i = \bar{Y}_i - b_i$ are mutually independent, with U_i having a normal distribution with mean μ and variance $\tau_i^2 = \sigma_i^2/n_i$. This is also the unconditional distribution of the U_i. It can be shown that $U_1, \ldots, U_k, SS_1, \ldots, SS_k$ form a set of minimal sufficient statistics for $\{\mu, \sigma_1^2, \ldots, \sigma_k^2\}$. For $i = 1, \ldots, k$, let $W_i = SS_i/(\tau_i^2 ss_i)$ and $w_i = 1/\tau_i^2$. Further, let $\mathbf{W} = (W_1, \ldots, W_k)^t$ and $\mathbf{w} = (w_1, \ldots, w_k)^t$. Now consider the quantity $R^*(\mathbf{D}; \mathbf{d}, \theta)$, defined by

$$R^*(\mathbf{D}; \mathbf{d}, \theta) = \bar{y}_W - \bar{b}_W - Z^* \left(\sum_{i=1}^{k} \frac{n_i Q_i}{ss_i} \right)^{-1/2}, \qquad (16)$$

where

$$\bar{y}_W = \sum_{i=1}^{k} W_i \bar{y}_i \Big/ \sum_{i=1}^{k} W_i$$

$$\bar{b}_W = \sum_{i=1}^{k} W_i b_i \Big/ \sum_{i=1}^{k} W_i, \qquad (17)$$

$Z^* = (\bar{U}_W - \mu)/\tau_0 \sim \mathrm{N}(0, 1)$, and $Q_i = SS_i/\sigma_i^2 \sim \chi^2_{n_i - 1}$, $i = 1, \ldots, k$.

Clearly, the distribution of $R^*(\mathbf{D}; \mathbf{d}, \theta)$ is free of model parameters (conditionally on the b_i and hence unconditionally as well). Furthermore, when the data vector $\mathbf{d}$ is substituted for $\mathbf{D}$ in $R^*(\mathbf{D}, \mathbf{d}, \theta)$, it reduces to $R^*(\mathbf{d}; \mathbf{d}, \theta) = \mu$, which is free of all nuisance parameters. Hence, $R^*(\mathbf{D}; \mathbf{d}, \theta)$ is a GPQ for μ. As usual, a $(1 - \alpha)$ GCI for μ is obtained as

$$L \le \mu \le U,$$

where $L = R^*_{\alpha/2}$ and $U = R^*_{1-\alpha/2}$, the $\alpha/2$th percentile and the $(1 - \alpha/2)$th percentile of the distribution of R^*. These may be conveniently estimated using a Monte Carlo approach. This involves Monte Carlo estimation of L and U whereby a large number of realizations from the distribution of R^* are generated and the empirical $\alpha/2$ and $1 - \alpha/2$ percentiles are used in place of $R^*_{\alpha/2}$ and $R^*_{1-\alpha/2}$ when computing L and U.

A single realization of R^* may be generated as follows:

1. For $i = 1, \ldots, k$, generate mutually independent $Q_i \sim \chi^2_{n_i - 1}$. Compute $W_i = n_i Q_i / ss_i$.
2. Generate $(b_1, \ldots, b_k)$ according to its fully specified joint distribution, independently of $(Q_1, \ldots, Q_k)$. In the case where b_i's are assumed to be mutually independent, generate independent realizations of $b_i \sim F_i$, $i = 1, \ldots, k$, independently of Q_i.
3. Compute $\bar{y}_W$ and $\bar{b}_W$ as in (17).
4. Generate a realization of $Z^* \sim \mathrm{N}(0, 1)$ independently of $Q_i, b_i, i = 1, \ldots, k$.
5. Calculate R^* as in (16).

We remark that although the GPQ R^* does involve the quantities b_i, under Model 3 assumptions they have known distributions, and hence the distribution of R^* is free of the parameters μ and σ_i^2, $i = 1, \ldots, k$. Thus we see that the approach of GCIs may be adapted to situations where informative prior distributions are available for a subset of the model parameters. This approach may be useful in more general situations as well.

4.1 Example

We use the data given in Table 2 to illustrate the GCI method for constructing confidence intervals on μ under Model 3. If we assume that b_i are uniformly distributed over the interval $[-M_i, M_i]$, then the nominal 95% GCI for μ, using 10,000 simulation runs, is (45.85, 47.05). The resulting interval is shorter than the GCI under Model 2, which assumes only $|b_i| \le M_i$. If we assume that M_i is the three-sigma limit of the normally distributed biases; that is, the b_i's are normally distributed with mean 0 and standard deviation $M_i/3$, then the nominal 95% GCI for μ, based on 10,000 simulation runs, is (46.03, 46.86).

4.2 Simulation Study for Model 3

Here we use the same settings as in Section 3.5 for the parameters μ, k, σ_i, and n_i in the simulation study. Without loss of generality, we assume that the mean of b_i is 0. The values of σ_{b_i}, the standard deviation of b_i, were generated using a uniform distribution between .2 and $1.5\sigma_{\max}$. Three types of distributions were used for b_i: uniform, normal, and gamma. The gamma random deviates were generated using $G^* = \sigma^2_{b_i} - G_{\sigma^2_{b_i}}$, where G_s has a pdf of $x^{s-1}e^{-x}/\Gamma(s)$, $x > 0$, so that the mean of G^* is 0 and the variance is $\sigma^2_{b_i}$. Again, there is no convenient canonical way to specify σ_{b_i}, and we used only a small number of combinations in the simulation.

In total, we considered 81 combinations of parameter values in the simulation. For each combination of parameter values, we generated 1,000 intervals with 95% nominal confidence level using the GCI method, and obtained the proportions of the time that the interval contained μ. Figure 4 displays the simulation results.

Each panel in Figure 4 plots the simulated coverage probability of the interval (y axis) versus the distribution type—"U" for uniform, "N" for normal, "G" for gamma—of b_i (x axis) for a specific combination of k and n_i configuration. The plotting symbols "1," "2," and "4" designate the cases with σ_k = 1, 2, and 4. Figure 4 shows that the GCI for μ under Model 3 maintains its coverage probability at or above the nominal value in all cases examined, although it is often conservative.

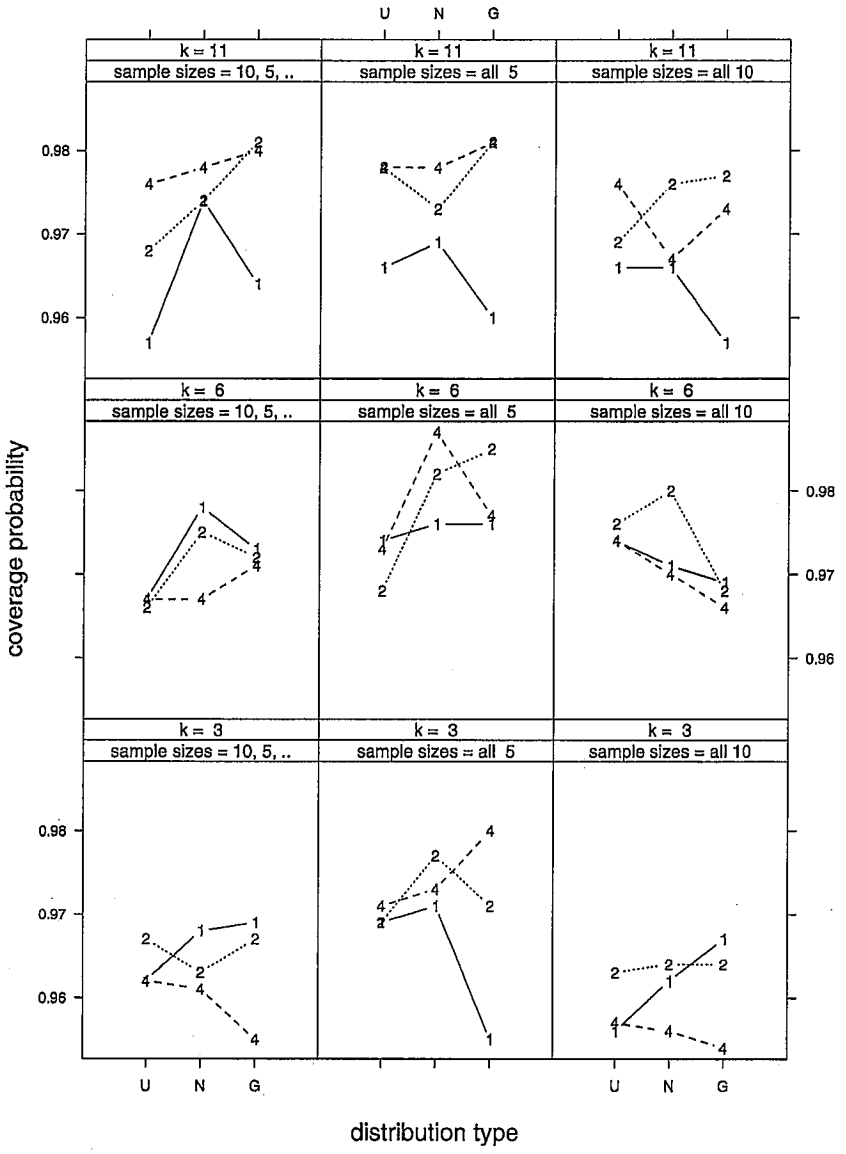

Figure 4. Plots of the Empirical Coverage Probabilities of the Generalized Confidence Interval Under Model 3 for Three Different Choices of Distribution for the b_i (U, uniform; N, normal; G, mean shifted gamma) and Three Different Values of σ_{max} (1, 2, or 4). The nine panels are arranged in a 3 × 3 grid with the rows representing different numbers of methods ($k = 3, 6, 11$) and the columns representing three different sample size patterns.

5. CONCLUSIONS

We have considered important classes of statistical models that have been used to analyze data from interlaboratory trials. They are the random-effects one-way classification model with unequal subsample sizes and heterogeneous variances, the bounded-bias model with known bias bounds, and the ISO GUM type model. For these three models, we have proposed confidence interval procedures for μ, the unknown true value of an artifact, based on the idea of generalized pivotal quantities. We have examined the performance of the proposed interval procedures using statistical simulation and have found that the coverage probabilities are sufficiently close to the nominal values for the proposed methods to be useful in practical applications.

For the one-way random-effects model, there appears to be no other satisfactory frequentist confidence interval for μ when k is small. For large values of k, the approach discussed by Rukhin and Vangel (1998), Vangel and Rukhin (1999), and Rukhin, Biggerstaff, and Vangel (2000) is an alternative satisfactory method. For this model, one might also consider comparing generalized confidence intervals for μ with highest posterior density (HPD) intervals using a Bayesian approach. Box and Tiao (1973) discussed a Bayesian approach for estimating a common mean μ and showed that the posterior distribution, using Jeffreys's prior, turns out to be a product of multivariate t-distributions. A generalization of this to a Bayesian one-way model was discussed by Vangel and Rukhin (1999). Although we did not consider Bayesian approaches in our simulation study, we hope to do so in future work.

For the bounded-bias model, we compare the GCI approach with the minimax approach of Eberhardt et al. (1989). The GCI performed satisfactorily and often better than the minimax approach. Furthermore, it was noted that the interval procedure

of Eberhardt et al. (1989) had some theoretical shortcomings. We also noted that for the bounded-bias model, it may happen that the bias bounds are specified in an inconsistent manner. To statistically test for this possibility, we have proposed a test procedure based on generalized confidence intervals.

For the ISO GUM type models with type-B uncertainties, the GCI was generally conservative, but in no instance in the simulated cases did the empirical coverage probability creep below the nominal rate of 95%.

Finally, we have given real data examples to illustrate the applications of each of the proposed methods. The methods presented here have a much wider range of applications, although they have been presented strictly in the context of interlaboratory trials.

It appears that our methods can be adapted for deriving confidence regions in more general regression models, for example, in general regression models under heteroscedasticity and in some semi-parametric regressions models, such as the approximately linear model due to Sacks and Ylvisaker (1978) and a similar semiparametric model due to Heckman (1988). These authors have addressed only the point estimation problem. The computation of confidence regions in such models is currently under investigation and will be reported in a future article.

APPENDIX: PROOF OF LEMMA 1

Refer to Rukhin et al. (2000) for a proof of part (a) of the lemma. For a proof of part (b), we calculate

$$\frac{d^2Q}{d(\sigma^2)^2} = 2\sum_{i=1}^{k} w_i^3(\bar{Y}_i - \bar{Y}_w)^2 - \frac{2[\sum_{i=1}^{k} w_i^2(\bar{Y}_i - \bar{Y}_w)]^2}{\sum_{i=1}^{k} w_i},$$

which is seen to be positive (almost surely) by an application of the Cauchy–Schwartz inequality. Hence it follows that Q is strictly convex in σ^2 with probability 1. Part (c) is an easy consequence of part (a) and equation (6).

Note that an earlier proof of Lemma 1 used matrix theory, but the proof given by Rukhin et al. (2000), using calculus arguments, is much simpler. This was pointed out to us by Professor Rukhin and also by one of the reviewers. This same reviewer also pointed out that Q is convex.

[Received October 2002. Revised March 2004.]

REFERENCES

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, New York: Wiley.

Eberhardt, K. R., Reeve, C. P., and Spiegelman, C. H. (1989), "A Minimax Approach to Combining Means, With Practical Examples," *Chemometrics and Intelligent Laboratory Systems*, 5, 129–148.

Heckman, N. E. (1988), "Minimax Estimates in a Semiparametric Model," *Journal of the American Statistical Association*, 83, 1090–1096.

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

Iyer, H. K., and Patterson, P. (2002), "A Recipe for Constructing Generalized Pivotal Quantities," Technical Report 10/02, Colorado State University, Dept. of Statistics.

Jordan, S. M., and Krishnamoorthy, K. (1996), "Exact Confidence Intervals for the Common Mean of Several Normal Populations," *Biometrics*, 52, 77–86.

Levenson, M. S., Banks, D. L., Gill, L. M., Guthrie, W. F., Liu, H. K., Vangel, M. G., Yen, J. H., and Zhang, N. F. (2000), "An ISO GUM Approach to Combining Results From Multiple Methods," *Journal of Research of the National Institute of Standards and Technology*, 105, 571–579.

Paule, R. C., and Mandel, J. (1971), *Analysis of Interlaboratory Measurements on the Vapor Pressure of Cadmium and Silver*, Special Publication 260-21, Gaithersburg, MD: National Institute of Standards and Technology.

——— (1982), "Consensus Values and Weighting Factors," *Journal of Research of the National Bureau of Standards*, 87, 377–385.

Rukhin, A. L., Biggerstaff, B. J., and Vangel, M. G. (2000), "Restricted Maximum Likelihood Estimation of a Common Mean and the Mandel–Paule Algorithm," *Journal of Statistical Planning and Inference*, 83, 319–330.

Rukhin, A. L., and Vangel, M. G. (1998), "Estimation of a Common Mean and Weighted Mean Statistics," *Journal of the American Statistical Association*, 93, 303–309.

Sacks, J., and Ylvisaker, D. (1978), "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6, 1122–1137.

Schiller, S. B., and Eberhardt, K. R. (1991), "Combining Data From Independent Chemical Analysis Methods," *Spectrochimica Acta*, 46B, 1607–1613.

Tsui, K. W., and Weerahandi, S. (1989), "Generalized P Values in Significance Testing of Hypotheses in the Presence of Nuisance Parameters," *Journal of the American Statistical Association*, 84, 602–607.

Vangel, M. G., and Rukhin, A. L. (1999), "Maximum-Likelihood Analysis for Heteroscedastic One-Way Random Effects ANOVA in Interlaboratory Studies," *Biometrics*, 55, 302–313.

Wang, C. M., and Splett, J. D. (1997), "Consensus Values and Reference Values Illustrated by the Charpy Machine Certification Program," *Journal of Testing and Evaluation*, 25, 308–314.

Weerahandi, S. (1993), "Generalized Confidence Intervals," *Journal of the American Statistical Association*, 88, 899–905.

——— (1995), *Exact Statistical Methods for Data Analysis*, New York: Springer-Verlag.

Willie, S., and Berman, S. (1995), "NOAA National Status and Trends Program Ninth Round Intercomparison Exercise Results for Trace Metals in Marine Sediments and Biological Tissues," NOAA Technical Memorandum NOS ORCA 93, U. S. Department of Commerce.

Yu, L. H., Sun, Y., and Sinha, B. K. (1999), "On Exact Confidence Intervals for the Common Mean of Several Normal Populations," *Journal of Statistical Planning and Inference*, 81, 263–277.