More Noise Does Not Mean More Precision: A Review of Aldenberg and Rorije (2013)

David R. Fox[1,2]

[1]Environmetrics Australia, Melbourne, Australia; [2]University of Melbourne, Melbourne, Australia

Summary — This paper provides a critical review of recently published work that suggests that the precision of hazardous concentration estimates from Species Sensitivity Distributions (SSDs) is improved when the uncertainty in the input data is taken into account. Our review confirms that this counter-intuitive result is indeed incorrect.

Key words: *ANOVA, Bayesian predictive distribution, beta-content tolerance intervals, components of variation, species sensitivity distributions, uncertainty estimation.*

Address for correspondence: *David R. Fox, PO Box 7117, Beaumaris, Victoria, Australia 3193.*
E-mail: *david.fox@environmetrics.net.au*

Introduction

In their recent paper, Aldenberg and Rorije (1; hereafter referred to as A&R) investigated the impact on the estimation of an HC_x (a concentration which is hazardous to no more than $x\%$ of a population) by using Species Sensitivity Distributions (SSDs) when 'data uncertainty' is taken into account. This is an important issue that is currently ignored by current ecotoxicological practice, although Fox (2) outlined a strategy based on sampling from the posterior distributions of the No Effect Concentrations (NECs) of species.

The paper by A&R tackles this issue from two perspectives: first through the use of a partitioning of total variation in the set of responses according to the source, as is done in conventional analysis of variance (ANOVA). The second strategy uses the tools of Bayesian inference. On the basis of the ANOVA analysis, A&R concluded that *between* group variability *decreases* with increasing *within* group variability. The results of A&R's Bayesian analysis reinforced this finding, and led them to state what they themselves acknowledged was the counter-intuitive conclusion, that increased noise in the SSD input data yields more precise estimates of an HC_x.

Regrettably, I believe that the ANOVA-based analysis provided in A&R is flawed, and the Bayesian analysis essentially duplicates the ANOVA analysis and thus provides no additional insights. Indeed, the analytical results of A&R's Bayesian analysis are identical to those obtained by using traditional maximum likelihood methods — a point acknowledged by the authors.

With respect to the ANOVA analysis, A&R constructed small, artificial data sets in such a way that the counter-intuitive conclusion regarding *between group* and *within group* variation was supported. Their approach displays aspects of what Fox and Burgman (3) refer to as 'anchoring', i.e. the adoption of a statement that has already been proposed. A&R acknowledged that their colleagues Wout Slob and Ad Ragas "[had] suggested that error-in-data would lead to reduced SSD variance many years ago". The anchoring to this suggestion meant that, rather than closely examining the ANOVA results, they were treated as confirmatory evidence supported by a Bayesian analysis, which, by virtue of simplifying assumptions, mirrored the ANOVA analysis.

As pointed out by Bayarri and Berger (4), there are certain circumstances where a joint Frequentist (e.g. ANOVA)-Bayesian approach is either necessary or preferable, such as the design of experiments or the use of Frequentist methods to assist with the elicitation of priors. However, in the present context we see no need to utilise both frameworks — A&R's thesis is built around the interpretation of ANOVA results. However, increasingly, as in the case of A&R's thesis, simplified Bayesian analyses are used to lend support to a Frequentist interpretation of the data, rather than in a stand-alone context that exploits the richness and power of a full Bayesian analysis. Thus, instead of making cogent arguments supporting the choice of a *single* statistical framework, we see both paradigms being used. The problem is that, in many instances, for reasons of simplicity and analytical tractability, uniform distributions are adopted for the Bayesian priors, which often lead to results that are similar or identical to those obtained by using classical, Frequentist statistics. Such is the case in the A&R paper.

In the remainder of this paper, I present an alternative analysis of the effects of increasing

data error on SSD estimation and inference. I commence with an examination of A&R's ANOVA model, and show that the combination of a misspecified model, together with illegitimate data constructs, were responsible for their findings. I then look at how the use of simplifying assumptions in the Bayesian analysis essentially duplicated the ANOVA analysis, rather than providing an independent, additional line of evidence. Finally, I note the duality between A&R's "new extrapolation constants" based on the Bayesian predictive Species Sensitivity Distribution and the Frequentists' more familiar beta-content tolerance intervals.

Between and Within Species Variation: A&R's ANOVA analogy

In the following development Y_{ij} denotes the log-transformed toxicity for the i^{th} replicate of species j where i = 1,...,m and j = 1,...,n. A simple partitioning of the total variation in the set of Y_{ij}s into the components "between species" and "within species", as was done in the A&R paper, implies the one-way ANOVA model given by Equation 1:

$$Y_{ij} = \mu + \alpha_j + \varepsilon_{ij} \qquad \text{[Equation 1]}$$

where μ is the overall (population) mean, α_j is the 'effect' of the j^{th} species, and the ε_{ij} are stochastic errors assumed to be independently normally distributed as ε_{ij}~N(O,σ^2). A&R define τ "as the standard deviation of SSD of species means", which we interpret to mean the square-root of the variance between species' mean toxicities.

The usual tabular presentation of ANOVA results was provided as "Table 1" in the A&R paper (the formula for SS_B as given in A&R is incorrect: the summation is over j not i and j and the result should be multiplied by m). A&R's table is reproduced here as Table 1. The purpose of this ANOVA table (Table 1) is that it summarises the information needed to simultaneously test the equality of the set of species means $\mu_1,\mu_2,...,\mu_n$ through an assessment of MS_B relative to the magnitude of MS_W. Specifically, the null hypothesis H_0: $\mu_1 = \mu_2 = ... = \mu_n$ is tested against the alternative hypothesis H_1: *at least two means are different* by comparing the ratio $\frac{MS_B}{MS_W}$ with a so-called 'critical value' of the F-statistic having *n–1* and *n(m–1)* degrees of freedom in the numerator and denominator, respectively.

The rationale for this test is straightforward: when H_0 is true, $\tau = 0$ (i.e. mean toxicities are the same for all species) and $Expected[MS_B] = Expected[MS_W] = \sigma^2$, so the ratio $\frac{MS_B}{MS_W}$ is unity. A formal statistical test is based on an assessment of the likelihood of obtaining the value computed from the data (or something greater than it) for this ratio, *assuming the null hypothesis to be correct*. This is the ubiquitous *p-value*.

In their paper, A&R provided the following equations for the estimation of μ, σ and τ:

$$\hat{\mu} = \bar{y}_{\bullet\bullet} \qquad \text{[Equation 2]}$$

$$\hat{\sigma} = \sqrt{MS_W} \qquad \text{[Equation 3]}$$

$$\hat{\tau} = \sqrt{\frac{MS_B - MS_W}{m}} \qquad \text{[Equation 4]}$$

where MS_B and MS_W are given in Table 1, and $\bar{y}_{\bullet\bullet}$ is the mean of all $m \cdot n$ toxicity values. I have no issue with Equations 2 and 3. My concern stems from what I believe to be the inappropriate use of Equation 4, which is linked with the manner in which the artificial data sets used by A&R to motivate the discussion have been constructed.

A&R correctly assert that, whether they be empirical or model estimates of toxicity, the input data to an SSD have error. The SSD is simply a theoretical cumulative distribution function (*cdf*) fitted to the toxicity data, and, as is the case with most distribution-fitting exercises, no account is made of model or measurement error — the data are taken as fixed points having no uncertainty. To their credit, A&R have sought to investigate the implication for quantities derived from the SSD (such as the HC_x), when this uncertainty is explicitly incorporated into the model-fitting process. The process used by A&R to achieve this is to take a small (synthetic) data set having no error, and incrementally introduce variation by manipulating

Table 1: ANOVA table for a one-way ANOVA model, as given in A&R

	df	SS	MS	Expected (MS)
Between species	$n-1$	$SS_B = \sum_{i,j}(\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2$	$MS_B = SS_B/(n-1)$	$\sigma^2 + m \cdot \tau^2$
Within species	$(m-1) \cdot n$	$SS_W = \sum_{i,j}(y_{ij} - \bar{y}_{\bullet j})^2$	$MS_W = SS_W/((m-1) \cdot n)$	σ^2
Total	$m \cdot n - 1$	$SS_T = \sum_{i,j}(y_{ij} - \bar{y}_{\bullet\bullet})^2$	$MS_T = SS_T/(m \cdot n - 1)$	

what they believed to be σ^2. Estimates of τ are obtained for each data set by using ANOVA techniques to decompose the total variation in the set of responses and then use Equation 4. A&R constructed two variants of the base (zero error) data set and found that their estimates of τ *decreased* with the assumed increases in σ. On the basis of this finding, A&R claimed *"SSD shrinkage with increasing data error"* — a predicted phenomenon attributed to Wout Slob and Ad Ragas in the *Acknowledgement* section of A&R's paper. Despite the admission that this seems to be counter-intuitive, A&R stated that *"bigger σ-values would lead to negative point estimates of τ"*.

Below, I demonstrate that intuition is indeed correct, and that A&R's assessment is flawed as a result of the combined effects of: a) the manner in which extra variation is introduced by A&R into their artificial data sets; b) a partitioning of total variation that is inconsistent with the implied statistical model; and c) the inappropriate use of Equation 4.

A&R's artificial data sets

Notwithstanding that the A&R paper and its recommendations are predicated on essentially three variants of a *single* set of extremely limited *artificial* data, I have concerns with the manner in which they have been constructed. The data sets used by A&R are reproduced in Tables 2a–c.

Table 2a is the 'base' table from which data in Tables 2b and 2c have been constructed. In essence, the data in Table 2a have been generated by using Equation 1 with $\alpha_1 = -1.309$; $\alpha_2 = -0.536$; $\alpha_3 = 0$; $\alpha_4 = 0.536$; $\alpha_5 = 1.309$; and $\sigma^2_{ij} = 0$. The analysis given immediately under Table 2a is correct, as is the estimate of $\hat{\tau} = 1.0$. Note that τ is an explicit function of the α terms, namely:

$$\tau^2 = \frac{\sum_{j=1}^{n} \alpha_j^2}{n-1} \qquad \text{[Equation 5]}$$

Using the values of α given above in Equation 5 results in:

$$\tau^2 = \frac{(-1.309)^2 + (-0.536)^2 + (0)^2 + (0.536)^2 + (1.309)^2}{5-1} = \frac{4.00}{4} = 1.0$$

A&R defined τ^2 as the variance between the collection of sample means. From Tables 2a–c (the over bar is missing on $\bar{y}_{\bullet j}$ and $\bar{y}_{\bullet\bullet}$ in Table 2a):

$$\bar{y}_{\bullet 1} = -1.309;\ \bar{y}_{\bullet 2} = -0.536;\ \bar{y}_{\bullet 3} = 0;\ \bar{y}_{\bullet 4} = 0.536;\ \bar{y}_{\bullet 5} = 1.309.$$

Furthermore,

$$\bar{y}_{\bullet\bullet} = \frac{1}{5}\sum_{j=1}^{5} \bar{y}_{\bullet j} = 0.0$$

and therefore *variance* of $\bar{y}_{\bullet j}$ terms is:

$$\frac{\sum_{j=1}^{5} (\bar{y}_{\bullet j} - \bar{y}_{\bullet\bullet})^2}{n-1} =$$

$$\frac{(-1.309)^2 + (-0.536)^2 + (0)^2 + (0.536)^2 + (1.309)^2}{4} = 1.0$$

which is what was obtained by using Equation 5.

So far, I am in agreement with A&R. However, what is fundamentally important to realise from Equation 5 is that, using the *same set of α values, the value of τ will remain constant, irrespective of the σ value and, in this case, always equal 1.0*. This led me to consider A&R's approach to generating data sets in Tables 2b and 2c.

The data in Table 2b have been obtained by adding –0.1 to the elements of the first row and adding +0.1 to the elements of the third row. The second row remains unaltered. Because +0.1 and –0.1 have been added to each column of data, there is no net change in the $\bar{y}_{\bullet j}$ values or $\bar{y}_{\bullet\bullet}$. The only thing that has changed is that, whereas the data in the columns of Table 2a had zero standard deviation, the data in the columns of Table 2b have standard deviation 0.1, but importantly, not (as assumed by A&R) as a result of any change in the *stochastic* variation σ^2. As I shall demonstrate later, the introduced variation in the artificial data sets is entirely *deterministic*. Since both the $\bar{y}_{\bullet j}$ and the α terms are the same, it is reasonable to conclude that τ is the same and equal to 1.0. And herein lies the problem: A&R estimated τ to be 0.998 for the data in Table 2b, *because they assumed σ^2 had increased* while the between group variability remained constant, and their definition of τ was wrong.

The same process was used by A&R to generate the data in Table 2c (Note: the highlighted entries in Table 2c are incorrect; they should be the same as Tables 2a and 2b) — this time by adding and subtracting 1.0, as above — leaving the $\bar{y}_{\bullet j}$ values and $\bar{y}_{\bullet\bullet}$ unchanged from Table 2a. By the same reasoning as above, τ for the data in Table 2c remains unaltered (and equal to 1.0), yet A&R's estimate for τ is now 0.816. On the basis of these incorrect estimates of τ, and through a simple re-arrangement of terms in Equation 4, A&R erroneously concluded *"when the data error is taken to $\sqrt{MS_W} = \sqrt{3.000} = 1.732$ the SSD would shrink to a point probability mass, as $\hat{\sigma} = 1.732$, $\hat{\tau} = 0$" and "bigger σ-values would lead to a negative point estimate of τ, without altering the method"*. But τ can only assume non-negative values; had A&R taken their artificial data construction method one-step further by adding and subtracting 2, for example, they would have immediately seen that something was amiss. I do acknowledge that there are instances where, in practice, the *F*-ratio calculated as $\frac{MS_B}{MS_W}$ is sometimes less than unity implying $MS_B < MS_W$. However, as noted by Meek

et al. (5), this is usually a result of an incorrect model specification. In particular, Meek *et al*. (5) cite the example in Meek and Turner (6), whereby the inappropriate analysis of a two-factor design by using one-way ANOVA resulted in a small *F*-ratio, which they point out "is an indication of a mis-specified model". As I detail in the next section, this is precisely what has occurred with the A&R analysis. Unfortunately, the anomalous results were used by A&R to discredit the ANOVA method and elevate the superiority of Bayesian techniques — even though A&R's Bayesian formulation of the problem was essentially equivalent to the Frequentists' ANOVA.

Table 2: Artificial data sets used by A&R

a) artificial log-toxicity data for five species, each measured without error

	1	**2**	**3**	**4**	**5**
1	–1.309	–0.536	0.000	0.536	1.309
2	–1.309	–0.536	0.000	0.536	1.309
3	–1.309	–0.536	0.000	0.536	1.309
$y_{\bullet j}$	–1.309	–0.536	0.000	0.536	1.309
$y_{\bullet\bullet}$	0.000				

	df	**SS**	**MS**	**Estimates**
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 0.000$	$MS_W = 0.000$	$\hat{\sigma}^2$
Total	14	$SS_T = 12.000$		

b) as in Table 2a, with 0.1 unit standard deviation in the three replicates per species

	1	**2**	**3**	**4**	**5**
1	–1.409	–0.636	–0.100	0.436	1.209
2	–1.309	–0.536	0.000	0.536	1.309
3	–1.209	–0.436	0.100	0.636	1.409
$\bar{y}_{\bullet j}$	–1.309	–0.536	0.000	0.536	1.309
$\bar{y}_{\bullet\bullet}$	0.000				

	df	**SS**	**MS**	**Estimates**
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 0.000$	$MS_W = 0.010$	$\hat{\sigma}^2$
Total	14	$SS_T = 12.100$		

c) as in Tables 2a and 2b, with 1.0 unit standard deviation in the three replicates per species

	1	**2**	**3**	**4**	**5**
1	–2.309	–1.536	–1.000	–0.464	0.309
2	–1.309	–0.536	0.000	0.536	1.309
3	–0.309	0.464	1.000	1.536	2.309
$\bar{y}_{\bullet j}$	–1.282	–0.524	0.000	0.524	1.282
$\bar{y}_{\bullet\bullet}$	0.000				

	df	**SS**	**MS**	**Estimates**
Between species	4	$SS_B = 12.000$	$MS_B = 3.000$	$\hat{\sigma}^2 + 3\,\hat{\tau}^2$
Within species	10	$SS_W = 10.000$	$MS_W = 1.000$	$\hat{\sigma}^2$
Total	14	$SS_T = 22000$		

Shaded entries are incorrect and should be –1.309, –0.536, 0.536, and 1.309 respectively. Other errors: entries in SS column of b) should be respectively 12.0047, 0.100, 12.1047; SS_T in c) should be 22.000.

Correct ANOVA Model

The first problem with the A&R analysis is that the assumed one-way ANOVA is inconsistent with the underlying structure of the data. The correct model for these data is given by Equation 6:

$$Y_{ij} = \mu + \alpha_j + \beta_i + \varepsilon_{ij} \qquad \text{[Equation 6]}$$

where μ, α and ε are as in Equation 1, and β_i is a second 'factor' effect. Equation 6 is typical of a one-way design with blocking, where the blocking factor applies to the rows of data in Tables 2a–c and could represent, for example, different analysts, time, laboratories, etc. In the present context, the blocking factor represents levels of adjustment applied by A&R to manipulate the variation in the data (which they erroneously ascribed to changes in σ). So, for example, the data in Table 2b are generated from Equation 6, by using $\alpha_1 = -1.309$; $\alpha_2 = -0.536$; $\alpha_3 = 0$; $\alpha_4 = 0.536$; $\alpha_5 = 1.309$ and $\beta_1 = -0.1$; $\beta_2 = 0.0$; $\beta_3 = 0.1$, while the data in Table 2c are generated from Equation 6 with the same set of α terms and $\beta_1 = -1$; $\beta_2 = 0.0$; $\beta_3 = 1$. The generic ANOVA table for the model given by Equation 6 is given in Table 3. The column of expected mean squares in Table 3 shows that the EMS associated with the species factor is as before, although there is now a term associated with the blocking factor which has an EMS of $\sigma^2 + n\phi^2$. It is relatively straightforward to show that the term ϕ^2 is given by Equation 7:

$$\phi^2 = \frac{\sum_{i=1}^{m}\beta_i^2}{m-1} \qquad \text{[Equation 7]}$$

Two important considerations emerge from Table 3 and Equation 7. The first is that A&R modified the base data in Table 2a to create data sets having some level of 'data error' (call it b) by setting $\{\beta_1 = -b, \beta_2 = 0, \beta_3 = b\}$, for which Equation 7 gives $\phi = b$. In other words, A&R's method of manipulating the base data was not, as they assumed, changing σ — it was changing ϕ. The fact that this was not apparent was because they analysed the data as a one-way design, when, in fact, it was a single factor block design. Furthermore, and more importantly, because A&R were not really changing σ, there was no stochastic error component, which means that Table 3 reduces to Table 4. This is reinforced by the fact that the sum of squares for the error term in Table 4 is zero. *In other words, the model A&R have used to generate their data is completely deterministic (Equation 8).*

$$Y_{ij} = \alpha_j + \beta_i, \quad i = 1,...,m;\ j = 1,...,n \qquad \text{[Equation 8]}$$

Hence, what A&R were ascribing to error was in fact the sum of squares term associated with blocks, which in turn leads to the reported difficulties with Equation 4.

Estimation of Tau

All the data sets used by A&R were constructed by using Equation 8 with the same α terms and different β terms $\{\beta_1 = -b, \beta_2 = 0, \beta_3 = b\}$, for which MS_B in Table 4 is 3.0 and

$$MS_R = \frac{2nb^2}{m-1}.$$

However, because A&R used a single-factor ANOVA design, they effectively collapsed the 'blocks' and 'error' rows of Table 4 into a single entry, which they identified as within species variation. Thus, A&R's estimate of MS_E is

$$\frac{SS_R + SS_E}{df_R + df_E} = \frac{n\sum_{i=1}^{m}\beta_i^2 + 0}{(m-1)+(n-1)(m-1)} = \frac{\sum_{i=1}^{m}\beta_i^2}{(m-1)},$$

that is ϕ.

Now Equation 4 is based on the square root of the difference between MS_B and MS_R which, using A&R's estimate of MS_E and with $m = 5$ and $n = 3$, is

Table 3: ANOVA layout for single factor block design

Source of variation	Sum of squares	df	Mean square	Expected mean square
Species	$SS_B = m\sum_{j=1}^{n}(\bar{Y}_{\bullet j} - \bar{Y}_{\bullet\bullet})^2$	$n - 1$	$MS_B = \frac{SS_B}{n-1}$	$\sigma^2 + m\tau^2$
Blocks (rows)	$SS_R = n\sum_{i=1}^{m}(\bar{Y}_{i\bullet} - \bar{Y}_{\bullet\bullet})^2$	$m - 1$	$MS_R = \frac{SS_R}{m-1}$	$\sigma^2 + n\phi^2$
Error	$SS_E = \sum_{j=1}^{n}\sum_{i=1}^{m}(\bar{Y}_{ij} - \bar{Y}_{i\bullet} - \bar{Y}_{\bullet j} + \bar{Y}_{\bullet\bullet})^2$	$(n - 1)(m - 1)$	$MS_E = \frac{SS_E}{(n-1)(m-1)}$	σ^2
Total	$SS_T = \sum_{j=1}^{n}\sum_{i=1}^{m}(\bar{Y}_{ij} - \bar{Y}_{\bullet\bullet})^2$	$nm - 1$		

Table 4: Partitioning of total variation in a manner that is consistent with A&R's data construction

Source of variation	Sum of squares	df	Mean square
Species	$SS_B = m\sum_{j=1}^{n}\alpha_j^2$	$n-1$	$MS_B = \frac{m\sum_{j=1}^{n}\alpha_j^2}{n-1}$
Blocks (rows)	$SS_R = n\sum_{i=1}^{m}\beta_i^2$	$m-1$	$MS_R = \frac{n\sum_{i=1}^{m}\beta_i^2}{m-1}$
Error	$SS_E = 0$	$(n-1)(m-1)$	$MS_E = 0$
Total	$SS_T = m\sum_{j=1}^{n}\alpha_j^2 + n\sum_{i=1}^{m}\beta_i^2$	$nm-1$	

$\hat{\tau} = \sqrt{\frac{3-b^2}{3}}$, which, as noted by A&R, becomes negative for $b > \sqrt{3}$.

A&R's Bayesian Analysis

A&R used a Bayesian hierarchical modelling approach described in Gelman *et al.* (7) as an alternative to the ANOVA modelling framework. The posterior density for μ and τ was taken to be:

$$p(\mu,\tau\,|\,y) \propto p(\mu,\tau)\prod Normal\left(\bar{y}_j\,|\,\mu,\sqrt{\sigma_j^2+\tau^2}\right)$$

[Equation 9]

where μ, σ and τ are as previously defined and $\sigma_j^2 = \frac{\sigma^2}{m}$

The derivation of this posterior density is based on the assumption that σ is *known*. By using uniform prior distributions on each of μ and τ, A&R investigated the properties of the joint posterior for various values of σ. There are two problems with this analysis: a) as discussed previously, A&R's artificial data sets have been constructed by using Equation 8, so there are *no stochastic components involved*; and, in any event b), the Bayesian formulation treats τ as stochastic, whereas it is a fixed constant in the ANOVA model. The consequence of (a) is that the Bayesian analysis is attempting to make inference about non-existent terms, so the results are bound to be erroneous. The significance of (b) is that the Bayesian formulation is equivalent to a random effects ANOVA, whereas the *assumed* response-generating model treats the effects as fixed. In a Bayesian context, fixed effects are given independent priors without the hierarchical structure.

Discussion and Conclusions

Bayesian methods *do* provide a credible alternative to traditional Frequentist approaches, and offer a potentially richer and more rewarding framework for applied statistical analysis. However, A&R's analysis appears to have suffered from both anchoring and bias. For example, the negative variance estimates obtained by A&R were viewed as a *failure* of ANOVA, yet the use of non-informative priors in the Bayesian analysis, which "leads to trouble in the hierarchical model" and causes "convergence problems", attracted no such criticism. Indeed, A&R spent some time discussing and justifying their choice of prior distribution,

Table 5: Summary statistics for HC_5 computed from data generated using Equation 1 and given values of sigma

Sigma = 0.1		Sigma = 1.0		Sigma = 2.0		Sigma = 3.0	
Min.:	–1.831	Min.:	–3.281	Min.:	–5.4677	Min.:	–7.8378
1st Qu.:	–1.679	1st Qu.:	–2.213	1st Qu.:	–2.9496	1st Qu.:	–3.7519
Median:	–1.645	Median:	–1.846	Median:	–2.2806	Median:	–2.9101
Mean:	–1.646	Mean:	–1.856	Mean:	–2.3373	Mean:	–2.9719
3rd Qu.:	–1.609	3rd Qu.:	–1.523	3rd Qu.:	–1.6557	3rd Qu.:	–2.0954
Max.:	–1.510	Max.:	–0.283	Max.:	0.2681	Max.:	0.3544

noting that "the error-in-data model will not work with the non-informative prior, but needs the uniform prior". They noted the historical importance of the non-informative prior, but cited a growing dissatisfaction with it: "we are inclined to switch to the uniform prior for the non-hierarchical model as well". This choice seems to have been motivated by analytical tractability, rather than by model integrity.

A&R's basic premise is reflected in their statement "the important message is that taking within-group error into account reduces the between group error". Not only is this counter-intuitive, it is wrong, as a cursory examination of the expected between-group mean square $MS_B = \sigma^2 + m\tau^2$ reveals. Clearly, the expected between-group variability increases with increasing σ. A&R used a completely deterministic manipulation of their artificial data to achieve a result where the *presumed* within group variation increased, while the between group variability remained constant — a situation contradicted by the formula for MS_B. This was achieved by manipulating β in Equation 8, but assigning the increases to σ in $MS_B = \sigma^2 + m\tau^2$. This inevitably led to flawed conclusions, which were reinforced by an equally flawed Bayesian analysis, *viz* "the Bayesian analysis confirms the ANOVA findings: the more noise within the individual species estimates, the less information remains for the SSD variance of expected endpoint values for different species".

Compounding the above errors, is the incorrect definition of τ as "the standard deviation of SSD of species means". As is seen by Equation 5, τ is the non-stochastic component of between-group variation due to the treatment effects. With their incorrect interpretation of τ and treating MS_B as fixed, A&R stated that "data error competes with the SSD standard deviation", and argued that the SSD standard deviation *decreases* as data error *increases*. A&R's Figure 3 showed the *shrinkage* of the so-called Bayesian predictive SSD as σ *increases*, which led to another erroneous claim that, with respect to the estimation of an HC_x, "more data error leads to less conservative estimation" — that is, a *larger* HC_x.

It is a straightforward task to simulate data by using the model given by Equation 1 to demonstrate this point. A normal density fitted to the group means is then used to estimate the HC_5. The results of 1,000 such simulations are summarised in Table 5 and Figure 1. In each case, three replicates were generated for each of five species from normal distributions with the same group means, as in Table 2a, and various values for σ.

The highlighted results in Table 5 show that increasing σ results in a *more conservative* (that is, smaller) HC_5 estimate — *not* a less conservative estimate, as claimed by A&R. Furthermore, the increase in variability of the HC_5 estimates with increasing σ is clearly evident.

Note the similarity between A&R's predictive extrapolation formula $\bar{y} + k_{pred} \cdot s_y$ and a more familiar beta-content tolerance interval (8–10). An extract from Table 5 of the A&R paper, showing "predictive extrapolation constants" for the HC_5, as well as our computed k values for various tolerance intervals, are provided in Table 6. Unlike a conventional tolerance interval, A&R's "extrapolation formula" has no probability attached to it. Given that the result of the extrapolation formula is a *random limit* (by virtue of it being based on the *sample* mean and *sample* standard deviation), then, whether one works in a Bayesian or Frequentist framework, there must be *uncertainty* associated with the limit. Thus, for the HC_5 there is some probability, p, that no more than 5% of all species will be affected by a concentration that is no greater than the limit established by the formula. The value of p corresponding to a tolerance interval using A&R's extrapolation constants, is shown in the last column of Table 6. It is immediately evident that p is not constant and appears to decrease with increasing sample size, n.

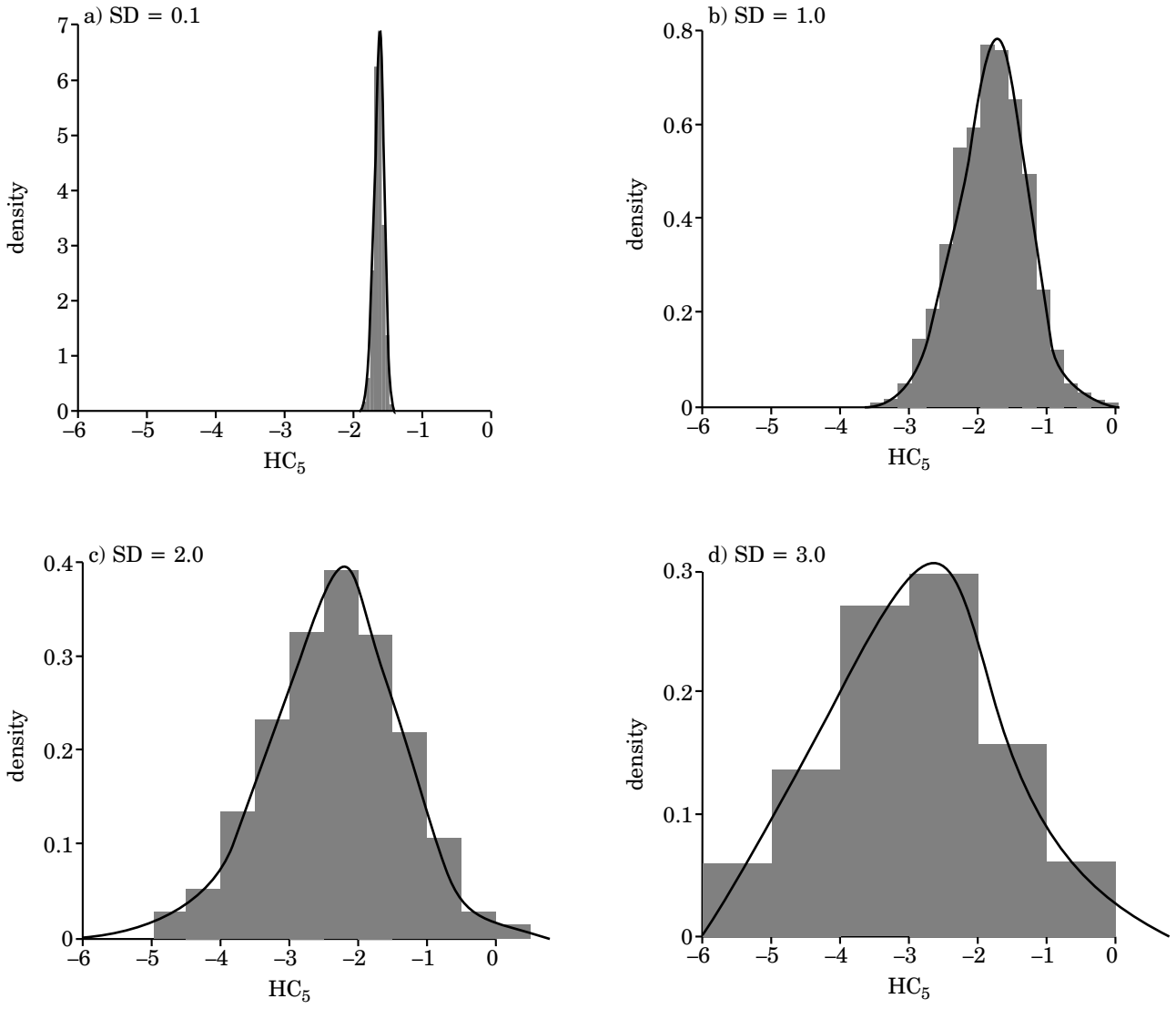
From the analysis above, it is clear that A&R's premise that more noise in the SSD data leads to more precise estimation of the HC_x is fundamentally flawed, as is the alternative method of deriv-

Table 6: A portion of extrapolation constants for HC_5, taken from A&R's Table 5

n	A&R	50% Tolerance	80% Tolerance	90% Tolerance	95% Tolerance	99% Tolerance	Equiv. Tolerance
3	–10.310	–1.938	–3.604	–5.311	–7.656	–17.370	97.2%
4	–3.998	–1.830	–2.968	–3.957	–5.144	–9.083	90.3%
5	–2.977	–1.779	–2.683	–3.400	–4.203	–6.578	85.1%
6	–2.574	–1.750	–2.517	–3.092	–3.708	–5.406	81.4%
7	–2.360	–1.732	–2.407	–2.894	–3.399	–4.728	78.6%

The first entry is from A&R, entries under the next five columns are one-sided beta-content tolerance interval constants k for various levels of 'confidence'. Entries in the last column are the level of confidence attached to a tolerance interval that uses a k-value equal to A&R's extrapolation constant.

Figure 1: Empirical distributions of HC_5 estimates as a function of increasing data variability



ing an HC_x for small samples. According to A&R, "the final recipe is embarrassingly simple: collect your average data or model point estimates, then use the revised predictive extrapolation constants, as if these were error-free, and you're done". Unfortunately, as we have demonstrated, things are not that simple.

We are aware that A&R's method is being presented as a better way of estimating an HC_x for small samples in a regulatory context (11). The errors in A&R's paper must be urgently addressed, before there is widespread uptake and adoption of its recommendations.

Acknowledgements

The author is indebted to the following individuals for their assistance during the preparation of this paper: Professor Murray Aitkin of Melbourne University's Department of Mathematics and Statistics, for his review of the statistical analysis, and Professor Wayne Landis of Western Washington University and Dr Peter Chapman of Golder Associates, who both made many helpful suggestions on style and content.

Received 31.03.15; received in final form 08.06.15; accepted for publication 09.06.15.

References

1. Aldenberg, T. & Rorije, E. (2013). Species sensitivity distribution estimation from uncertain (QSAR-based) effects data. *ATLA* **41**, 19–31.
2. Fox, D.R. (2010). A Bayesian approach for determining the no effect concentration and hazardous

concentration in ecotoxicology. *Ecotoxicology & Environmental Safety* **73**, 123–131.
3. Fox, D.R. & Burgman, M.A. (2008). Ecological risk assessment. In *Encyclopedia of Quantitative Risk Assessment and Analysis* (ed. E. Melnick & B. Everitt), pp. 1600–1603. Chichester, UK: John Wiley & Sons Ltd.
4. Bayarri, M.J. & Berger, J.O. (2004). The interplay of Bayesian and Frequentist analysis. *Statistical Science* **19**, 58–80.
5. Meek, G., Ceyhun, O. & Dunning, K.A. (2007). Small F-ratios: Red flags in the linear model. *Journal of Data Science* **5**, 199–215.
6. Meek, G. & Turner, S. (1983). *Statistical Analysis for Business Decisions*, 783pp. Boston, MA, USA: Houghton Mifflin.
7. Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd edn, 68pp. Boca Raton, FL, USA: Chapman & Hall/CRC.
8. Guenther, W.C. (1972). Tolerance intervals for univariate distributions. *Naval Research Logistics Quarterly* **19**, 309–333.
9. Guenther, W.C., Patil, S.A. & Uppuluri, V.R.R. (1976). One-sided β-content tolerance factors for the two-parameter exponential distribution. *Technometrics* **18**, 333–340.
10. Krishnamoorthy, K. (2009). *Statistical Tolerance Regions: Theory, Applications, and Computation,* 512pp. Chichester, UK: John Wiley & Sons Ltd.
11. Aldenberg, T. (2014). HC_5 estimation in SSDs revisited. *Workshop Programme: Estimating toxicity thresholds for aquatic ecological communities from sensitivity distributions*. 11–13 February 2014, Amsterdam. European Centre for Ecotoxicology and Toxicology of Chemicals.