DOI 10.51558/1840-4561.2021.18.2.5

A STATISTICAL APPROACH FOR RELIABLE MEASUREMENT WITH FAMILIARIZATION TRIALS

Steven Kim[1], Christopher Essert[1]

[1]*Department of Mathematics and Statistics, California State University, Monterey Bay, USA*

Original scientific paper

Abstract
An accurate and reliable measurement is important in exercise science. The measurement tends to be less reliable when subjects are not professional athletes or are unfamiliar with a given task. These subjects need familiarization trials, but determination of the number of familiarization trials is challenging because it may be individual-specific and task-specific. Some participants may be eliminated because their results deviate from arbitrary ad hoc rules. We treat these challenges as a statistical problem, and we propose model-averaging to measure a subject's familiarized performance without fixing the number of familiarization trials in advance. The method of model-averaging accounts for the uncertainty associated with the number of familiarization trials that a subject needs. Simulations show that model-averaging is useful when the familiarization phase is long or when the familiarization occurs at a fast rate relative to the amount of noise in the data. An applet is provided on the internet with a very brief User's Guide included in the appendix to this article.
Keywords: Familiarization; reliability; accuracy; model-averaging; Akaike Information Criterion

INTRODUCTION

An accurate and reliable measurement is important in exercise science and related areas. It is especially challenging with human subjects. When compared to athletes, non-athletes tend to produce less accurate and reliable outcomes because they are not familiar with a given task. Most volunteers who participate in studies are not professional athletes, and these volunteers need practice trials in order to become familiar with the task at hand (Hopkins, 2000). A meta-analysis emphasized the importance of familiarization trials (Hopkins, Schabord & Hawley, 2001). Without a familiarization period, the researchers found that adolescent subjects produced more variable measures of lower-limb electromyograph than adult subjects, indicating that larger sample sizes should be required for adolescents (Waldron, Highton & Gray, 2016). Moreover, when researchers studied children to determine the number of familiarization trials in various fitness tests, the results seemed dependent on the particular test performed (Vrbik et al., 2016).

The necessary number of familiarization trials may be specific to both the individual and the assigned task. The expertise and experience of the researcher(s) may also play an important role because they often study new topics, some in which they may have more or less experience. Furthermore, not all researchers conduct experiments in the same population or under the same conditions. Therefore, researchers have relied on arbitrary decisions within a more or less agreeable scope. For instance, subjects repeated trials until they achieved consecutive measures within an acceptable range, but the choice of the acceptable range varied between studies (Beckham et al., 2019; Stockbrugger & Haennel, 2003). Moreover, in small-sample studies, researchers would not want to lose any portion of a small sample simply because the study participants failed to meet arbitrary criteria.

Different methods and rules have been applied to address the issue of familiarization, but there has always been a common goal: the accurate and precise quantification of the ability (the true unknown state) of a human subject whose performance tends to improve during the first few trials. In this article, for concise notation, we denote τ for the number of trials needed for familiarization and μ for the true ability (measured by a given task) once the familiarization has occurred. For instance, $\tau = 1$ implies a subject is already familiar with a given task, and $\tau = 3$ implies that a subject is familiarized at the third trial. The notations, μ and τ, are graphically depicted in Figure 1. In this article, we assume:

(1) The number of familiarization trials is specific to the subject of the experiment;

(2) A subject is already familiar or needs at least one trial until he or she becomes familiar with the routine ($\tau \geq 1$);

(3) A subject tends to underperform before the familiarization occurs and improves linearly (as an approximation) as he or she gets closer to their true ability;

(4) A subject is asked to perform a sufficient number of trials (denoted by m) so that familiarization occurs before m trials ($\tau < m$); and

(5) m is not too large to cause fatigue or any other factor which negatively affects the measurement once the familiarization has occurred.

Figure 1 represents a subject of $\tau = 3$ who required three trials in order to reach their true ability μ. In practice, researchers observe data points without knowing μ and τ as shown in the right panel of the figure, so τ and μ are to be estimated based on observed data. In this article, we use a statistical approach to the practical challenge of familiarization, and we demonstrate a method of model-averaging which does not require researchers to determine τ before the experiment is performed. It also determines μ by a data-driven weighted average. Section 2 explains the method of model-averaging applied to the familiarization problem. Section 3 provides a numerical example, and Section 4 demonstrates the operating characteristics of model-averaging via simulations under various scenarios. For practitioners, a free applet is provided at https://cessert.shinyapps.io/familiarization/ to apply the method of model-averaging (the applet's User's Guide is furnished in the appendix of this article).

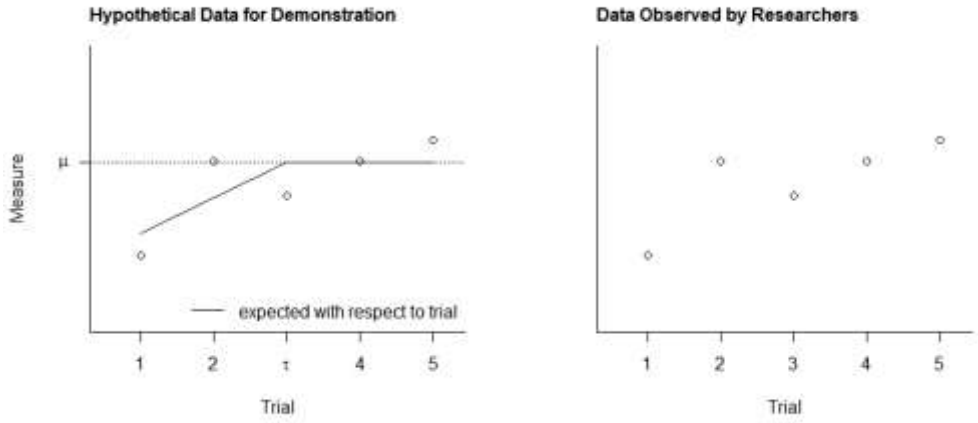

Figure 1: τ denotes the true number of trials needed for familiarization, and μ denotes a subject's true ability after familiarization (left). Researchers observe data in order to estimate unknown μ based on unknown τ (right).

STATISTICAL METHODS

We translate the aforementioned assumptions to a statistical model as follows. For m observed measures (y_1, …, y_m) we assume that data are generated by a normal model $y_t \sim N(\mu_t, \sigma)$ where $\mu_t = \beta_0 + \beta_1 t$ for $t < \tau$ (the expected measure μ_t changes linearly with respect to the number of trials, t, before familiarization occurs) and $\mu_t = \mu$ for $t \geq \tau$ (the true ability μ is maintained once the familiarization occurs). This statistical model is graphically shown in Figure 1. The model intercept β_0, slope β_1 (the rate of change in the expected measure until familiarization occurs), and σ (the standard deviation of the unexplained error around the expectation) are to be estimated given the data (y_1, …, y_m). It is a special case of the two-line model (also known as "changepoint regression") when the location of change τ is known (Julious, 2001). In our context, τ is unknown, but it can be estimated from the subset of natural numbers {1, 2, …, m – 1} which provides the best model fit according to the maximum likelihood estimation. We consider four statistical approaches: sample mean, maximum likelihood estimation (MLE), model-averaging with Akaike Information Criterion (AIC) (Akaike, 1974; Buckland, Burnham & Augustin, 1997; Wagenmakers & Farrell, 2004), and model-averaging with Akaike Information Criterion with finite-sample Correction (AICC) (Burnham & Anderson, 2002; Sugiura, 1978).

SAMPLE MEAN

Among the four statistical methods considered in this article, estimating μ by the sample mean (i.e., the simple average of y_1, …, y_m) may be the simplest approach, and it is the best estimation for μ if a subject is already familiar with a given task (i.e., $\tau = 1$). Note that if $\tau \geq 2$ and the expected

measure increases (i.e., $\beta_1 > 0$), the sample mean will underestimate the population mean μ.

Maximum Likelihood Estimation (MLE)

Under the normal model assumption, the four parameters (τ, β_0, β_1, σ) are to be estimated. The MLE chooses the values of (τ, β_0, β_1, σ) which maximizes the likelihood (a value which quantifies the model fit with the given data). In this article's companion applet, all computations are performed in R (Version 4.0.2).

Model Averaging by AIC Weights

There are $m - 1$ possible choices for τ = 1, 2, …, m – 1. Assuming τ = 2 is true, then the parameters (β_0, β_1, σ) can be estimated by the MLE, and we let L_2 be the maximized likelihood when τ = 2. Similarly, we can obtain the maximized likelihoods L_3, L_4, …, L_{m-1} when assuming τ = 3, 4, …, m – 1, respectively. The Akaike Information Criterion (AIC) is defined as $\text{AIC}_\tau = -2 \cdot \ln(L_\tau) + 2v$, where v is the number of parameters to be estimated (in this case, v is always 3 for our three parameters). For each assumption, we can evaluate AIC_τ for τ = 1, 2, …, m – 1. The AIC-weight is defined for each specific assumption of familiarization occurring at trial τ as

$$w_\tau = \frac{\exp(-0.5 \cdot \text{AIC}_\tau)}{\exp(-0.5 \cdot \text{AIC}_1) + \ldots + \exp(-0.5 \cdot \text{AIC}_{m-1})} .$$

Note that the above definition of AIC-weight is simplified from the original definition given in Wagenmakers & Farrell (2004) and others, yet it is equivalent to the original definition. The AIC-weight w_τ is between zero and one, and it approximates the probability that the assumed value of τ is true after observing data (y_1, …, y_m). The estimation for μ depends on the assumption τ = 1, 2, …, m – 1. After obtaining τ-specific estimates for μ, denoted by $\hat{\mu}_1$, $\hat{\mu}_2$, …, $\hat{\mu}_{m-1}$, the subject's true ability is estimated by the weighted average

$$\hat{\mu} = w_1\hat{\mu}_1 + w_2\hat{\mu}_2 + \cdots + w_{m-1}\hat{\mu}_{m-1} .$$

Model Averaging by AICC Weights

The interpretation of the AIC-weight becomes more accurate as the number of trials per subject increases. For a small number of trials, a corrected AIC (AICC) is recommended (Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). The AICC is defined as

$$\text{AICC}_\tau = \text{AIC}_\tau + \frac{2v(v+1)}{m-v-1} ,$$

where m is the maximum number of trials and v is the number of parameters to be estimated (in this case, v is 3 for our three parameters β_0, β_1 and σ). The definition of AICC-weight is similar to the AIC-weight given in Section 2.3 (replacing AIC_τ by AICC_τ), and the remaining procedure of estimating μ is the same as described in Section 2.3.

NUMERICAL EXAMPLE

As an example, suppose five measurements were observed (1.25, 3.03, 2.37, 3.02, 3.43) from a study participant. The sample mean is 2.62, which also happens to be the final estimate for μ under the assumption of τ = 1 (i.e., the subject did not require any familiarization trials at all; in other words, we assume that the subject was already familiar with the task at hand from the get-go at the very first trial). However, if we assume the familiarization happened at τ = 2, 3, or 4, the respective estimates for μ are 2.96, 3.06, or 3.24 as shown by the dotted lines in their respective plots in Figure 2. According to the method of MLE, μ is estimated as 2.96 because its likelihood value is the highest as shown in the Likelihood column of Table 1. According to the AIC method, μ is estimated by the weighted average

$$2.6200(0.0335) + 2.9625(0.7103) + 3.0569(0.1079) + 3.2447(0.1483) = 3.0031.$$

In this estimation, the assumption τ = 2 was most likely (w_2 = 0.7103 by AIC-weight), and the assumption τ = 1 was least likely (w_1 = 0.0335 by AIC-weight). Using the different method of model-averaging, the method of AICC estimates μ by

$$2.6200(0.9965) + 2.9625(0.0026) + 3.0569(0.0004) + 3.2447(0.0005) = 2.6214.$$

The AICC penalizes an additional parameter (due to the familiarization period) substantially more than the AIC, particularly when the sample size is small. The estimate is very close to the sample mean of 2.62 because the simplest assumption that no familiarization is needed is weighted so much (w_1 = 0.9965 by AICC-weight).

Table 1: Parameter estimates and model weights for each assumption τ = 1, 2, 3, 4.

Assumption	Parameter Estimates				Likelihood and Model Weights		
	β_0	β_1	μ	σ	Likelihood	AIC-weight	AICC-weight
τ = 1	NA	NA	2.6200	0.7647	0.0032	0.0335	0.9965
τ = 2	-0.4625	1.7125	2.9625	0.3398	0.1830	0.7103	0.0026
τ = 3	0.8725	0.7281	3.0569	0.4954	0.0278	0.1079	0.0004
τ = 4	1.1624	0.5206	3.2447	0.4649	0.0382	0.1483	0.0005

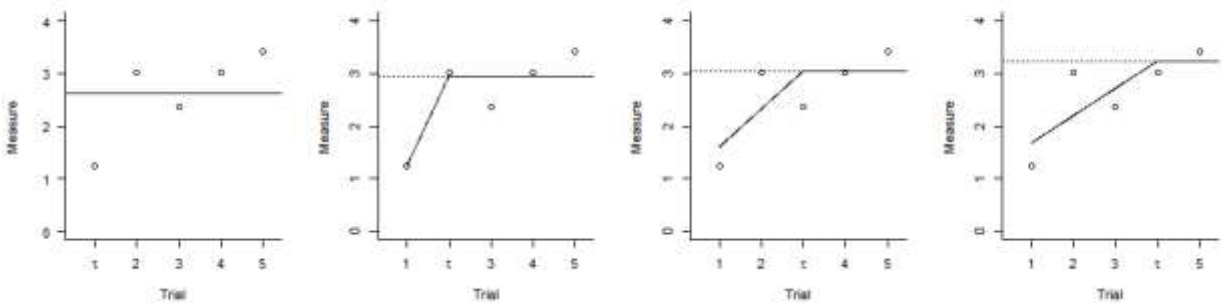

Figure 2: Estimates with respect to trial for each assumption τ = 1, 2, 3, 4 (from left to right).

SIMULATIONS

It is impossible to determine which method of estimation is preferable based on a single numeric example. That one method outperforms another depends on its parameters (particularly τ, β_1, and σ) and the maximum number of trials m. In this section, we compare the above four methods of sample mean, MLE, AIC weights, and AICC weights via simulations.

Simulation Designs

The four methods were compared by simulations using various scenarios: β_0 = 2.5; β_1 = 0.25, 0.5, 0.75, 1; σ = 0.5, 0.25, 0.1; and m = 5, 6, 7, 8, 9, 10. Note that changing the value of β_0 would not affect the relative performance of the four methods because it only determines the height of the two-line model graphically shown in Figure 1. The relative performances are rather sensitive to β_1, σ, and m because it is easier to detect τ (the true number of familiarization trials needed) as β_1 increases (clear trend of systematic changes in observed measures) and σ decreases (less variability of measures around the expected measure with respect to trials). Of course, the amount of statistical information increases as m increases. For each m = 5, 6, …, 10, we set the true values of τ = 2, 3, …, m – 1. Note that the case of τ = 1 is not worth demonstrating via simulations because the sample mean is the best method of estimating μ when a subject does not need any familiarization trials (i.e., a two-line model would be over-fitting in this case). In the simulation studies, the performance of each method is measured by the square-root of the mean squared error (RMSE), which is the average squared distance of estimates from the true value of μ per simulation scenario.

Simulation Results

In summary, (1) the AIC-weight or AICC-weight is the best method (smallest RMSE) in most scenarios, (2) the MLE method is *occasionally* the best when the familiarization trend is very clear (i.e., a large value of β_1, a small value of σ, and/or a large value of τ), and (3) the simple average (mean) is the best when the familiarization trend is *not* clear (i.e., a small value of β_1, a large value of σ, and/or a small value of τ). The simulation results demonstrate the patterns which, upon reflection, make sense when determining whether researchers can benefit from the AIC-weight or AICC-weight.

To focus on the cases when the simple average (mean) outperforms the other methods (MLE, AIC-weight, and AICC-weight), Figure 3 presents the scenarios of

($\beta_1 = 0.25$, $\sigma = 0.5$, $m = 7$),
($\beta_1 = 0.25$, $\sigma = 0.5$, $m = 10$),
($\beta_1 = 0.25$, $\sigma = 0.25$, $m = 7$),
($\beta_1 = 0.25$, $\sigma = 0.25$, $m = 10$),
($\beta_1 = 0.5$, $\sigma = 0.5$, $m = 7$), and
($\beta_1 = 0.5$, $\sigma = 0.5$, $m = 10$).

If we are to choose between the simple average and a model-averaging method in the absence of knowing the true τ and μ, the benefit of using model-averaging seems relatively high when compared using the simple average.

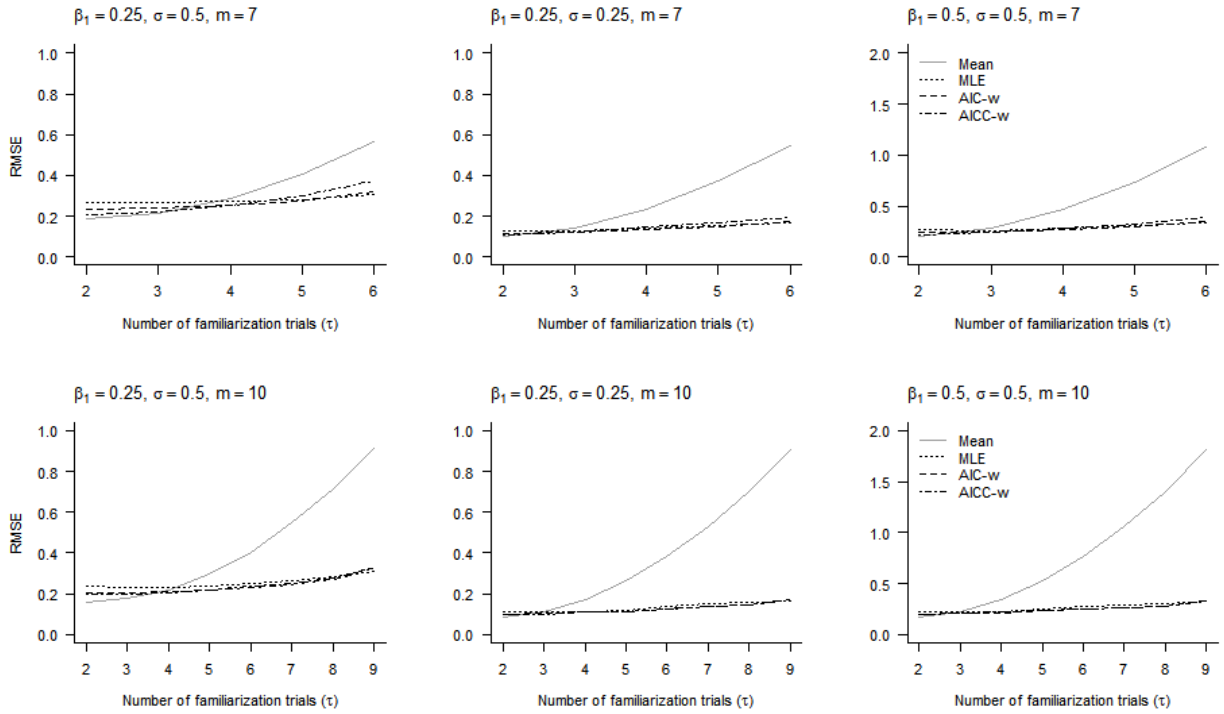

Figure 3: Graphic presentation of the simulation results

DISCUSSION

In many practical situations, from a statistical perspective, the purpose of familiarization trials is to accurately measure a subject's ability or performance, and the purpose of repeated trials is to precisely measure his or her ability by reducing the variance or variability. In the literature, we have seen some effort to pre-specify τ (the number of familiarization trials applied to all subjects) in order to estimate μ, a subject's true ability. In this article, we demonstrate methods of estimating μ without pre-specification of τ. The MLE method is a single-model (best fit) approach, and the model-averaging method is a multiple-model approach. In the scope of our simulation scenarios, it appears that the model-averaging method of using AIC-weight seems reasonable for practical use. The AIC-weight performs worse than the simple average (in terms of mean squared error) under any of the following circumstances:

(1) when researchers let subjects repeat trials a large number of times (i.e., a large m), or
(2) when only one or two familiarization trials are needed (i.e., small τ), or
(3) when unfamiliarized trials and familiarized trials are not clearly distinguishable (i.e., small β_1 and large σ).

Model-averaging methods have been popular in various disciplines. For instance, they have been proposed to determine the maximum tolerable dose in early-phase clinical trials (Yin & Yuan, 2009); to recommend an acceptable benchmark-dose of a toxic agent for public health (Bailer, Noble & Wheeler, 2005; Shao & Small, 2011; Kim, Bartell & Gillen, 2015); to model wind-speed distributions (Gong & Shi, 2010); and in economics and psychological sciences (Steel, 2020;

Hinne et al., 2020), to name a few. It is a popular statistical strategy used to achieve reliable estimation when researchers cannot determine with certainty one single model or whenever model uncertainty occurs. Hopefully, in the future it will also be applied in a wide range of areas in kinesiology and exercise science.

In conclusion, in the absence of knowledge of task-specific τ and/or individual-specific τ, we recommend the model-averaging method for researchers as it only requires a reasonable maximum number of trials without determination of the number of familiarization trials or any other arbitrary rules. For convenience, a free applet is provided at https://cessert.shinyapps.io/familiarization/. It allows the user to enter or to "copy and paste" any number of observations that were recorded and then returns estimates for μ and τ with graphical presentation. See the following Appendix for instructions or clarification if needed.

Appendix: User's Guide

In a web browser, enter "cessert.shinyapps.io/Familiarization." The screen will appear as shown at the top of Figure 4. Data can be separated by commas (e.g., 1.25, 3.03, 2.37, 3.02, 3.43) or separated by spaces (e.g., 1.25 3.03 2.37 3.02 3.43) as shown in Figure 4. Press the Enter key or click on the "Familiarize" button to process the data. It will produce results with graphics. If there are a large number of data points, depending on the screen size, it may be necessary to press the Page Down key to see the full table of results on the bottom of the screen. To clear results and reset the data field, press the function F5 key on the top row of your keyboard. It will allow the next set of data to be entered into the text box.

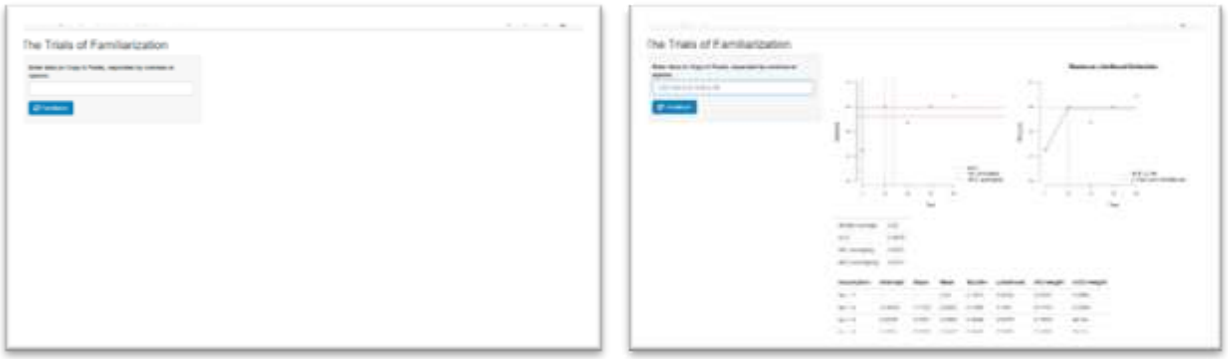

Figure 4: Demonstration of the applet

ACKNOWLEDGMENTS

This research is supported by the Discovery, Creation, and Integration (DCI) Support Program at College of Science, California State University, Monterey Bay.

REFERENCES

1. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716-723.
2. Bailer, A. J., Noble, R. B., & Wheeler, M. W. (2005). Model uncertainty and risk estimation for experimental studies of quantal responses. *Risk Analysis, 25*(2), 291-299.
3. Beckham, G., Lish, S., Disney, C., Keebler, L., DeBeliso, M., & Adams, K. J. (2019). The reliability of the seated medicine ball throw as assessed with accelerometer instrumentation. *Journal of Physical Activity Research, 4*(2), 108-113.
4. Buckland, S. T., Burnham, K. P., & Augustin, N. H. (1997). Model selection: An integral part of inference. *Biometrics, 53*, 603-618.
5. Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.
6. Gong, L., & Jing, A. (2010). Application of Bayesian model averaging in modeling long-term wind speed distributions. *Renewable Energy, 35*(6), 1192-1202.

7. Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*(2), 200-215.
8. Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine, 30*(1), 1-15.
9. Hopkins, W., Schabort, E. J., & Hawley, J. A. (2001). Reliability of power in physical performance tests. *Sports Medicine, 31*, 211-234.
10. Julious S. A. (2001). Inference and estimation in a changepoint regression problem. *The Statistician, 50*(1), 51-61.
11. Kim, S. B., Bartell, S. M., & Gillen, D. L. (2015). Estimation of benchmark dose in the presence or absence of hormesis using posterior averaging. *Risk Analysis, 35*(3), 396-408.
12. R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
13. Shao, K., & Small, M. J. (2011). Potential uncertainty reduction in model-averaged benchmark dose estimates informed by an additional dose study. *Risk Analysis, 31*(10), 1561-1575.
14. Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58*(3), 644-719.
15. Stockbrugger, B.A., & Haennel, R. G. (2003). Contributing factors to performance of a medicine ball explosive power test: a comparison between jump and nonjump athletes. *Journal of Strength Conditioning Research, 17*(4), 768-774.
16. Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory & Methods A7*, 13-26.
17. Vrbik, I., Sporiš, G., Štefan, L., Madić, D., Trajković, N., Valantine, I., & Milanović, Z. (2016). The influence of familiarization on physical fitness test results in primary school-aged children. *Pediatric Exercise Science, 29*(2), 278-284.
18. Wagenmakers, E., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review, 11*(1), 192-196.
19. Waldron, M., Highton, J., & Gray, A. (2016). Effects of familiarization on reliability of muscleactivation and gross efficiency in adolescents and adults. *Cogent Medicine, 3*, 1237606.
20. Yin, G. & Yuan, Y. (2009). Bayesian model averaging continual reassessment method in phase I clinical trials. *Journal of the American Statistical Association, 104*, 954-968.

Correspondence to:
Steven Kim Ph.D. in Statistics
Associate Professor
Department of Mathematics and Statistics
California State University, Monterey Bay
Address: 100 Campus Center Seaside, California 93955 USA
Phone: 831-582-3954
E-mail: stkim@csumb.edu