BAYESIAN MODEL SELECTION: APPLICATION TO THE ADJUSTMENT OF FUNDAMENTAL PHYSICAL CONSTANTS

BY OLHA BODNAR AND VIKTOR ERIKSSON

1 Unit of Statistics, School of Business, Örebro University, olha.bodnar@oru.se
2 Department of Statistics, Uppsala University, viktor.eriksson@statistik.uu.se

A method originally suggested by Raymond Birge, using what came to be known as the Birge ratio, has been widely used in metrology and physics for the adjustment of fundamental physical constants, particularly in the periodic reevaluation carried out by the Task Group on Fundamental Physical Constants of CODATA (the Committee on Data of the International Science Council). The method involves increasing the reported uncertainties by a multiplicative factor large enough to make the measurement results mutually consistent. An alternative approach, predominant in the meta-analysis of medical studies, involves inflating the reported uncertainties by combining them, using the root sum of squares, with a sufficiently large constant (often dubbed dark uncertainty) that is estimated from the data.

In this contribution, we establish a connection between the method based on the Birge ratio and the location-scale model, which allows one to combine the results of various studies, while the additive adjustment is reviewed in the usual context of random effects models. Framing these alternative approaches as statistical models facilitates a quantitative comparison of them using statistical tools for model comparison. The intrinsic Bayes factor (IBF) is derived for the Berger and Bernardo reference prior, and then it is used to select a model for a set of measurements of the Newtonian constant of gravitation ("Big G") to estimate a consensus value for this constant and to evaluate the associated uncertainty. Our empirical findings support the method based on the Birge ratio. The same conclusion is reached when the IBF corresponding to the Jeffreys prior is used and also when the comparison is based on the Akaike information criterion (AIC). Finally, the results of a simulation study indicate that the suggested procedure for model selection provides clear guidance even when the data comprise only a small number of measurements.

1. Introduction. The values of fundamental constants in the physical sciences, like the Newtonian constant of gravitation, G, and the relative atomic masses (atomic weights) of many elements, are often determined by combining measurement results from studies performed at different times and places [Mohr, Newell and Taylor (2016), Meija and Possolo (2017), Newell et al. (2018), Alighanbari et al. (2020)]. A similar need arises in measurement science: interlaboratory studies are done routinely to ensure the comparability of measurements made by different laboratories working independently and to substantiate the claims that these laboratories make concerning their measurement and calibration capabilities [Mandel and Paule (1970), Ruhkin (2003), Toman, Fischer and Elster (2012), Koepke et al. (2017), Bodnar and Elster (2020)]. However, the vast majority of such data blending exercises occur in the medical sciences, in the form of the meta-analysis of studies that, taken individually, may be inconclusive, but prove decisive once they are pooled with other studies of the same issue [Brockwell and Gordon (2001), Lambert et al. (2005), Higgins, Thompson and Spiegelhalter (2009), Rukhin (2013), Bodnar et al. (2017), Jones et al. (2018)].

Keywords and phrases: Intrinsic Bayes factor, Birge ratio method, location-scale model, random effects model, reference prior, meta-analysis, interlaboratory comparison study, Newtonian constant of gravitation.
Each individual study yields an estimate of the quantity of interest with its associated uncertainty, which is typically expressed in the form of a standard error or confidence interval and captures the evaluated contributions from all identified sources of uncertainty that are deemed to be substantively significant.

Very often, however, the results that one wishes to combine are mutually inconsistent, in the sense that the measured values are appreciably more variable than their associated uncertainties indicate that they should be. Figure 1 depicts measurements of the Newtonian constant of gravitation made between 1982 and 2018. The individual reported uncertainties clearly are overoptimistic, falling short of capturing the contributions from all the sources of uncertainty in play, which are the root cause of the overdispersion of the measured values. Such heterogeneity requires an explanation, and, in the case of $G$, it is an active topic of research [see, e.g., Rothleitner and Schlamminger (2017)]. This “extra” variability is often called *dark uncertainty* [cf. Thompson and Ellison (2011)] because it remains in the dark until the results of different studies are compared.

![Figure 1](image_url)

**Fig 1.** Measured values (diamonds) of the Newtonian constant of gravitation, $G$, and associated standard uncertainties (thick, vertical blue line segments, each representing a measured value plus or minus one standard uncertainty). The measurement results are listed in Table 1 of Merkatas et al. (2019), which also explains the labels of the results.

The Birge ratio method and the random effects meta-analysis seem to be the two most popular approaches to mitigating the presence of dark uncertainty when the results of several studies are pooled. Koepke et al. (2017) provided an overview of the cultural preferences that have led some groups to adopt one approach and other groups to adopt the other.

The Birge ratio method is generally favored by physicists, including those in CODATA [Mohr, Newell and Taylor (2016)] and the Particle Data Group [Particle Data Group and Zyla et al. (2020)]. The procedure was introduced by Birge (1932) and has been widely used, for example, by Weise and Wöger (2000) and Mohr, Newell and Taylor (2016), among many others.

The random effects model is the predominant method used in medical research [Hardy and Thompson (1998), Ades, Lu and Higgins (2005), Turner et al. (2015), Guolo and Varin (2017), Veroniki et al. (2019)]. Mandel and Paule (1970) introduced it into chemistry. Toman and Possolo (2009, 2010), Toman, Fischer and Elster (2012), Hannig et al. (2018), and Possolo et al. (2021) described relevant methods and provided examples of applications. Bodnar,
Link and Elster (2016) proposed objective Bayesian methods for Bayesian inference in the context of generalized random effects models and applied them to compute a consensus value for the Planck constant.

Bodnar et al. (2016) compared the conventional additive random effects model and the Birge ratio method, which was presented in the context of the location-scale model. This comparison focused on the robustness to model misspecification of confidence intervals for the overall mean. However, Bodnar et al. (2016) discussed only one aspect of model selection, and their results do not answer the question of which method is preferable for taking dark uncertainty into account. In this contribution, we address model selection, focusing on how the two models cope with the challenge posed by the presence of dark uncertainty. We do so from the viewpoint of objective Bayesian statistics, relying on the intrinsic Bayes factor (IBF).

Bayesian model selection may be preferable for several reasons [see Berger and Pericchi (2001, p. 138–140)]: (i) the Bayes factor is interpreted as an odds ratio that is easy to understand; (ii) it serves as an automatic Occam’s razor since it tends to favor a simple model over a complex model; (iii) relying on the Bayes factor achieves consistency in the sense that if any of the models being compared is the true model, then this model is selected with a probability that increases as the sample size increases (under mild regularity conditions); and (iv) the approach does not require the models to be nested, thus making it possible to compare different types of models involving different numbers of parameters, possibly with different meanings. Berger and Pericchi (2001, p. 142) also pointed out some limitations of the approach to model selection based on intrinsic Bayes factors. For example, the Bayes factor can be difficult to compute, and it may not be reliable when using vague proper prior distributions.

Traditionally, the Bayes factor is computed by assigning an informative prior to the model parameters. However, in practice, there may be no prior information, or there may be only vague information, about the parameters [see, e.g., Lambert et al. (2005)]. As a result, the choice of the prior distribution may exert a larger influence upon the Bayes factor than on the estimation of a parameter. This happens because the influence of the prior distribution on the Bayes factor does not diminish as the sample size increases [Berger and Pericchi (1996)].

Since the Bayes factor is particularly vulnerable to the misspecification of the prior distribution, we will adopt an objective, noninformative prior, either the Jeffreys prior [Jeffreys (1946)] or a reference prior [Berger and Bernardo (1992)]. The latter is obtained by maximizing the Shannon mutual information between the prior and the posterior, thus giving the data the best chance to drive the posterior away from the prior [Berger and Bernardo (1992), Clarke and Yuan (2004), Berger, Bernardo and Sun (2009), Bodnar and Elster (2014a)].

Different approaches for objective Bayesian model selection have various limitations and implementation difficulties. The traditional Bayes factor corresponding to an informative proper prior cannot be used for objective Bayesian model selection based on a noninformative prior because noninformative priors are usually improper. A modification of the definition of the conventional Bayes factor is required to overcome this difficulty. The modifications proposed by O’Hagan (1995) and Berger and Pericchi (1996) are used most often. O’Hagan (1995) suggested a fractional Bayes factor for model selection when an improper prior is used, and Berger and Pericchi (1996) developed a theory of Bayesian model selection based on the intrinsic Bayes factor.

The benefit of the intrinsic Bayes factor is that it is automatic (and arguably objective) in the sense that it only depends on the observed data and a noninformative prior distribution. It may, however, be computationally intensive and unstable [Berger and Pericchi (1996, p. 120–121)]. The intrinsic Bayes factor makes use of a minimal training sample for parameter estimation. To increase the stability of the result, it is common to average (or to compute
the median of) the intrinsic Bayes factors that correspond to all possible training samples. In this contribution, we will introduce the use of the empirical probability for Bayesian model selection based on the intrinsic Bayes factor, and we will apply it to measurements of the Newtonian constant of gravitation.

In this paper, we derive the intrinsic Bayes factor for both the location-scale model and the random effects model when noninformative priors are assigned to the model parameters. In Section 2, we introduce the concept of Bayesian model selection, with a special emphasis on the case in which the model parameters are endowed with a noninformative prior. The competing models, namely the location-scale model related to the Birge ratio approach and the random effects model, are presented in Section 3. Here, we also derive the quantities needed for the computation of the intrinsic Bayes factor. Section 4 comprises a simulation study used to compare the procedures that, in Section 5, are applied to the measurements of the Newtonian constant of gravitation. Section 6 discusses the results. The derivation of the theoretical results is presented in Supplement A. Additional simulation results concerning the performance of the IBF for model selection are presented in Supplement B for the Jeffreys prior and in Supplement C for the Akaike information criterion (AIC). Finally, the posterior predictive distributions under the location-scale model and the random effects model are derived in Supplement D.

2. Bayesian model selection. In this section we introduce the Bayes factor, provide its interpretation, and discuss how it can be used for model comparison. Let \( M_1, \ldots, M_q \) denote the models we wish to compare. Under model \( M_i \), the data \( x = (x_1, \ldots, x_n) \) follow the probability density function \( f_i(x|\theta_i) \) with \( \theta_i \in \Theta_i \), which, as a function of \( \theta_i \), is the likelihood function. In the Bayesian approach to model selection, we start by assigning a prior probability \( P(M_i) \), which represents the probability that a model is correct, to each model \( M_i \), and we assign a prior distribution \( \pi_i(\theta_i) \) to \( \theta_i \). The true data-generating model need not be among \( \{M_1, \ldots, M_q\} \) [Kass and Raftery (1995, Section 3.2)]. In such a case, we select the model that fits the data best, which is similar to model selection using the Akaike information criterion (AIC) [Held and Bové (2014)].

The posterior probability \( P(M_i|x) \) of the model \( M_i \) being correct when the sample \( x \) is observed is given by

\[
P(M_i|x) = \frac{\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)} B_{ji}(x)}{\sum_{j=1}^{q} \frac{P(M_j)}{P(M_i)}},
\]

where

\[
B_{ji}(x) = \frac{m_j(x)}{m_i(x)} = \frac{\int f_j(x|\theta_j)\pi_j(\theta_j) d\theta_j}{\int f_i(x|\theta_i)\pi_i(\theta_i) d\theta_i},
\]

is the Bayes factor between \( M_j \) and \( M_i \) and

\[
m_i(x) = \int_{\Theta_i} f_i(x|\theta_i)\pi_i(\theta_i) d\theta_i,
\]

is the marginal distribution of the data under model \( M_i \), which describes the probability of observing \( x \) under model \( M_i \).

The posterior probability \( P(M_i|x) \) is dependent on the subjectively chosen prior probabilities \( P(M_i) \) for \( i = 1, 2, \ldots, q \) of the models \( M_i \) being correct. If one wishes the Bayesian model selection procedure to be objective, then one must assign equal prior probabilities to the considered models, that is, \( P(M_1) = \ldots = P(M_q) = 1/q \). In this case, the posterior probability \( P(M_i|x) \) used for the selection of a model will depend on the Bayes factors \( B_{ji}(x) \).
only. In a special case, when \( q = 2 \) and \( P(M_1) = P(M_2) = 0.5 \), the Bayes factor \( B_{ji}(x) \) can be used alone to perform Bayesian model selection. The Bayes factor is interpreted as the odds in favor of \( M_j \) against \( M_i \) based on the evidence in the data \( x \). The inequality \( B_{ji} > 1 \) is supportive of model \( M_j \), while \( B_{ji} < 1 \) is interpreted as evidence against \( M_j \). Finally, it is noted that the prior distributions \( \pi_i(\theta_i) \) of the parameter \( \theta_i \) are considered to be subjective because they are decided before considering the data, but they are objective if one uses a noninformative prior.

The Bayes factor can also be interpreted as an odds ratio if the marginal distribution \( m_i(x) \) is proper, such that \( \int_{x \in \Omega} m_i(x) dx = 1 \) for all \( i = 1, \ldots, q \). However, the latter equality may not always be fulfilled when the model parameters are endowed with an improper prior, denoted by \( \pi_i^N(\theta_i) \), i.e., when objective Bayesian inference is preferable. An improper prior distribution \( \pi_i^N(\theta_i) \) cannot have a normalizing constant and its application leads to an improper marginal distribution of the data \( m_i(x) \). As a result, one can no longer apply conventional Bayesian model selection without modifying the definition of the Bayes factor. In the next section, we describe the intrinsic Bayes factor developed by Berger and Pericchi (1996) for Bayesian model selection with improper noninformative priors.

### 2.1. Intrinsic Bayes factor

The intrinsic Bayes factor (IBF) is a solution to the problem that is created for the Bayes factor by an improper prior distribution, namely that the Bayes factor has no interpretation as an odds factor under an improper prior distribution. The IBF differs from the ordinary Bayes factor in that it makes use of a training sample. The training sample is employed for the purpose of transforming the improper prior distribution into the proper posterior distribution, which is then used as a prior for the rest of the elements in the sample.

The training sample \( x_{(\ell)} \) is a subset of \( x \) of size \( m \). Let \( x_{(\ell)} = x - x_{(\ell)} \) be the data set with the training sample removed. Then, \( x_{(\ell)} \) and \( x_{(\ell)} \) form a partition of \( x \). Let \( m_i^n(x_{(\ell)}) \) be the marginal distribution of \( x_{(\ell)} \) computed under the noninformative prior \( \pi_i^N(\theta_i) \) and let \( \pi_i^N(\theta_i|x_{(\ell)}) \) denote the posterior distribution of \( \theta_i \) computed based on \( \pi_i^N(\theta_i) \) and the sample \( x_{(\ell)}. \) A training sample \( x_{(\ell)} \) is called proper if \( 0 < m_i^n(x_{(\ell)}) < \infty \) for all \( M_i \), and it is called minimal if it is proper and no subset of \( x_{(\ell)} \) is proper. As a result, the training sample is used to transform the improper prior \( \pi_i^N(\theta_i) \) into the proper posterior \( \pi_i^N(\theta_i|x_{(\ell)}) \), which is then used in the computation of the conditional distribution of \( x_{(\ell)} \) given the training sample \( x_{(\ell)}. \) It is noted that a part of the data, which could have been used to compute the Bayes factor, is lost since it is needed for the specification of the proper posterior \( \pi_i^N(\theta_i|x_{(\ell)}) \). We are sacrificing a portion of the data set \( x \), namely the training sample \( x_{(\ell)} \), for parameter estimation instead of using it for model selection. On the other side, we want as much data as possible for the computation of the Bayes factor since the overall goal is to make a comparison between models. This is why the minimal training sample is preferable.

The size of the minimal training sample is usually the same as the number of parameters in the model. However, there are exceptions to this rule: if, for example, the prior distribution is proper for some parameters, then the minimal training sample size may be less than the number of parameters in the model [see Berger and Pericchi (1996)]. As an example, suppose we have a joint prior distribution of two parameters that is improper. Then, just two observations from the data set should suffice as a training sample. When the sample size is \( n \), there are \( n(n - 1)/2 \) combinations of two observations from \( x \) to choose from. One should pick every combination to compute the IBF, which implies \( n(n - 1)/2 \) computations of the IBF. Then, conclusions are drawn based on a summary of the computed IBFs, such as the average IBF or the median IBF.
Let $m^N_i(x_{(ℓ)}|x_ℓ)$ denote the conditional distribution of $x_{(ℓ)}$ given $x_ℓ$. Then, the intrinsic Bayes factor computed for the sample $x_{(ℓ)}$ conditioned on the training sample $x_ℓ$ is given by

$$B^N_{ji}(x_{(ℓ)}|x_ℓ) = \frac{m^N_j(x_{(ℓ)}|x_ℓ)}{m^N_i(x_{(ℓ)})} \frac{m^N_i(x_ℓ)}{m^N_j(x_ℓ)} = \frac{m^N_j(x)}{m^N_i(x)} \frac{m^N_i(x_ℓ)}{m^N_j(x_ℓ)} = B^N_{ji}(x) \times B^N_{ij}(x_ℓ).$$

which can also be rewritten as

$$B^N_{ji}(x_{(ℓ)}|x_ℓ) = \frac{m^N_j(x_{(ℓ)}|x_ℓ)}{m^N_i(x_{(ℓ)})} \frac{m^N_i(x_ℓ)}{m^N_j(x_ℓ)} = \frac{m^N_j(x)}{m^N_i(x)} \frac{m^N_i(x_ℓ)}{m^N_j(x_ℓ)} = B^N_{ji}(x) \times B^N_{ij}(x_ℓ).$$

Recall that the difficulty with the conventional Bayes factor was related to improper prior distributions. For two models $M_j$ and $M_i$ with likelihoods $f_j(x|θ_j)$ and $f_i(x|θ_i)$ and with employed improper priors $π^N_j(θ_j) = c_j h_j(θ_j)$ and $π^N_i(θ_i) = c_i h_i(θ_i)$, the IBF of $M_j$ and $M_i$ is given by

$$B^N_{ji}(x_{(ℓ)}|x_ℓ) = B^N_{ji}(x) \times B^N_{ij}(x_ℓ).$$

The important point is that the arbitrary constants of proportionality $c_j$ and $c_i$ will cancel in the IBF [see O’Hagan (1995)]. Hence, the IBF does not depend on anything arbitrary or subjective, and it can therefore be used for objective Bayesian model selection.

We summarize this subsection by saying that a noninformative prior distribution is used for an objective approach to Bayesian model selection. The IBF is a method in which the noninformative improper prior $π^N_i(θ_i)$ is transformed into the proper posterior $π^N_i(θ_i|x_ℓ)$ using a minimal training sample $x_ℓ$, while the rest of the data $x_{(ℓ)}$ are employed for model selection, resulting in the IBF. The IBF measures the odds in favor of $M_j$ against $M_i$ based on evidence in the data, without the arbitrariness of the constants $c_j$ and $c_i$.

### 3. Models for dark uncertainty

Interlaboratory comparisons and meta-analysis studies provide measured values qualified with uncertainty evaluations. When there are correlations between the measured values, these should also be reported. For example, there are three non-zero correlations for the Newtonian constant data; they are listed in Section 5. Moreover, it is usually assumed that the measured values can be modeled using a normal distribution [see, e.g., Lambert et al. (2005), Turner et al. (2015), Hannig et al. (2018)], that is, $x \sim N(μ, U)$, where $1$ stands for the $n$-dimensional vector of ones. The positive definite matrix $U$ contains the squared reported uncertainties in the main diagonal, and the off-diagonal entries contain the covariances $ρ_{ij} u_i u_j$, where $ρ_{ij}$ is the correlation between the measurements of the $i$th and $j$th studies. The entries of $U$ are often assumed to be known with full certainty.

Any dark uncertainty, however, by definition is not reflected in the reported uncertainties, and the measured values are overdispersed in consequence, as is the case for the measurements of $G$. In physics and metrology applications, this issue is often addressed by increasing the quoted uncertainties by a multiplicative factor, known as the Birge ratio. This approach corresponds to the assumption that $x \sim N(μ, τ^2_{LS} U)$, where $τ^2_{LS}$ models the extra variability observed in the sample (see Section 3.1 for details).

An alternative approach involves an additive adjustment, which corresponds to the random effects model typically favored in medical applications. In this case, the assumption is that
\( \mathbf{x} \sim N(\mu \mathbf{1}, \mathbf{U} + \tau^2_{RE} \mathbf{I}) \), where \( \tau_{RE} \) models the extra variability that is “missing” from the uncertainties reported for the individual studies (see Section 3.2).

Since both approaches can be described as statistical models, one can use statistical techniques for model selection to identify the approach that seems best for the data at hand. This identification offers no guarantee about either model being the true data-generating model, and it serves only to compare the abilities of the alternative models to describe the patterns in the data.

In this section, we will develop the theoretical results needed to compute the IBF for the location-scale model, which performs the multiplicative adjustment of uncertainty, and for the random effects model, which provides an additive adjustment. To the best of our knowledge, no other models have been used for meta-analysis and interlaboratory studies. Other approaches may be proposed in the future, and the derived theoretical results of Sections 3.1–3.3 should be adopted for each new model. We provide the IBF used for model comparison with the Berger and Bernardo reference prior for the location-scale model and the random effects model. The corresponding formulas for the Jeffreys prior and for the Akaike information criterion are summarized in Supplements B and C, respectively. We note that the results in Sections 3.1–3.3 are derived without imposing any assumptions on the elements of the positive definite matrix \( \mathbf{U} \), while in the literature, this matrix is usually restricted to being diagonal.

3.1. Location-scale model. The location-scale model implements the Birge method by multiplying each uncertainty by a common factor larger than 1, which is called the Birge ratio. Bodnar and Elster (2014b) studied the Birge ratio within a Bayesian framework in the context of the location-scale model.

The location-scale model \( M_{LS} \) is defined by

\[
\mathbf{x} = \mu \mathbf{1} + \varepsilon_{LS}, \quad \text{with} \quad \varepsilon \sim N(0, \tau^2_{LS} \mathbf{U}),
\]

where \( \mu \) is the location parameter and \( \tau_{LS} \) is the scale parameter following Definition 4.3 in Held and Bové (2014). The location-scale model assumes that the measurement data are obtained from a multivariate normal distribution with the mean \( \mu \mathbf{1} \) and the covariance matrix \( \tau^2_{LS} \mathbf{U} \), i.e., \( \mathbf{x} | \mu, \tau_{LS}, M_{LS} \sim N(\mu \mathbf{1}, \tau^2_{LS} \mathbf{U}) \). The factor \( \tau_{LS} \) is an unknown quantity that represents the unexplainable interlaboratory heterogeneity. The overall mean \( \mu \) is the target parameter in many applications in physics, chemistry, and medicine, while \( \tau_{LS} \) is a nuisance parameter included in the model to capture the heterogeneity. The motivation behind the application of the model is that the elements in \( \mathbf{U} \) can be understated due to the presence of heterogeneity. By multiplying \( \mathbf{U} \) by \( \tau^2_{LS} \), the elements in \( \mathbf{U} \) then become uniformly adjusted for the heterogeneity. On the other side, the reported correlations \( \rho_{ij} \) are unchanged under this adjustment. The notation \( \mathbf{x} | \mu, \tau_{LS}, M_{LS} \) makes it explicit that the data are conditioned on the two parameters \( \mu \) and \( \tau_{LS} \) under the location-scale model \( M_{LS} \).

The likelihood function of \( \mathbf{X} \), when the location-scale model is assumed, is given by

\[
f(\mathbf{x} | \mu, \tau_{LS}, M_{LS}) = \frac{\tau^{-n}_{LS}}{(2\pi)^{n/2}} \text{exp}\left(-\frac{1}{2\tau^2_{LS}}(\mathbf{x} - \mu \mathbf{1})^T \mathbf{U}^{-1} (\mathbf{x} - \mu \mathbf{1})\right).
\]

The Berger and Bernardo reference prior has been derived for the general location-scale model in Fernández and Steel (1999) and is given by

\[
\pi^N(\tau_{LS}) = \pi^N(\mu, \tau_{LS}) \propto \frac{1}{\tau_{LS}}.
\]

The marginal distribution of \( \mathbf{x} \) and the marginal distribution of the training sample \( \mathbf{x}_f \) are derived in Theorem 3.1; its proof is given in Supplement A.
Theorem 3.1. Let \( n > 2 \) and let \( \mathbf{U} \) be positive definite. Then, under the location-scale model (1) and reference prior (3),

(i) the marginal distribution of the whole sample is given by

\[
m(\mathbf{x}|\mathbf{M}_{LS}) = \frac{\Gamma\left(\frac{n-1}{2}\right)(\mathbf{x}^T \mathbf{Q} \mathbf{x})^{-\frac{n-1}{2}}}{(\det(\mathbf{U}))^{1/2}2\pi^{-\frac{n}{2}}\sqrt{\mathbf{I}^T \mathbf{U}^{-1} \mathbf{I}}},
\]

where

\[
\mathbf{Q} = \mathbf{U}^{-1} - \frac{\mathbf{U}^{-1} \mathbf{I} \mathbf{I}^T \mathbf{U}^{-1}}{\mathbf{I}^T \mathbf{U}^{-1} \mathbf{I}}.
\]

(ii) The size of a minimal training sample \( \mathbf{x}_\ell \) is given by \( m = 2 \), i.e., \( \mathbf{x}_\ell = \{x_i, x_j\} \) where \( x_i, x_j \in \mathbf{x} \) and \( i \neq j \). Moreover, the marginal distribution of the training sample is given by

\[
m(\mathbf{x}_\ell|\mathbf{M}_{LS}) = \frac{1}{2\sqrt{\det(\mathbf{U}_\ell)}}(\mathbf{x}_\ell^T \mathbf{Q}_\ell \mathbf{x}_\ell)(\mathbf{1}_2 \mathbf{U}_\ell^{-1} \mathbf{1}_2),
\]

where \( \mathbf{U}_\ell \) is obtained from \( \mathbf{U} \) by taking the elements lying at the intersections of the \( \ell \) rows and \( \ell \) columns, and

\[
\mathbf{Q}_\ell = \mathbf{U}_\ell^{-1} - \frac{\mathbf{U}_\ell^{-1} \mathbf{1}_2 \mathbf{1}_2^T \mathbf{U}_\ell^{-1}}{\mathbf{1}_2^T \mathbf{U}_\ell^{-1} \mathbf{1}_2} \quad \text{with} \quad \mathbf{1}_2 = (1, 1)^T.
\]

The results of Theorem 3.1 provide all the quantities that are needed for the computation of the IBF, as shown in Section 3.3.

3.2. Random effects model. The random effects model \( \mathbf{M}_{RE} \) is defined by

\[
\mathbf{x} = \mu \mathbf{1} + \lambda_{RE} + \varepsilon_{RE}, \quad \text{with} \quad \lambda_{RE} \sim N(0, \tau_{RE}^2 \mathbf{I}) \quad \text{and} \quad \varepsilon_{RE} \sim N(0, \mathbf{U}).
\]

Based on the above equation, we can see that the random vector \( \mathbf{x} \) is multivariate normally distributed, i.e., \( \mathbf{x}|\mu, \tau_{RE}, \mathbf{M}_{RE} \sim N(\mu \mathbf{1}, \mathbf{U} + \tau_{RE}^2 \mathbf{I}) \). In the random effects model, an adjustment for the heterogeneity is made to \( \mathbf{U} \) by adding a common term \( \tau_{RE}^2 \) to every element in the diagonal of \( \mathbf{U} \), where \( \tau_{RE}^2 \) represents the unexplainable interlaboratory heterogeneity.

Since only the diagonals of \( \mathbf{U} \) are modified by \( \tau_{RE}^2 \), the variances \( u_i^2 \) are adjusted for the heterogeneity, while the covariances in \( \mathbf{U} \) are unaltered. Finally, we note that the definition of the random effects model given here (8) corresponds to how this model is typically used in metrology and medicine: it is used with a single observation for each study. In many other applications, the model is formulated to accommodate multiple, replicated observations made in each study (Pinheiro and Bates, 2000).

For the random effects model, the likelihood function is expressed as

\[
f(\mathbf{x}|\mu, \tau_{RE}, \mathbf{M}_{RE}) = \frac{\left(\det(\mathbf{U} + \tau_{RE}^2 \mathbf{I})\right)^{-\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mu \mathbf{1})^T(\mathbf{U} + \tau_{RE}^2 \mathbf{I})^{-1}(\mathbf{x} - \mu \mathbf{1})\right).
\]

The Berger and Bernardo reference prior has been derived by Bodnar, Link and Elster (2016) and it is given by

\[
\pi^N(\tau_{RE}) = \pi^N(\mu, \tau_{RE}) \propto \sqrt{\tau_{RE}^2 \cdot \text{tr}\left((\mathbf{U} + \tau_{RE}^2 \mathbf{I})^{-2}\right)}.
\]

It has been proven by Bodnar, Link and Elster (2016, p. 32) that \( \pi^N(\mu, \tau_{RE}|\mathbf{x}) \) is proper for the random effects model and the Berger and Bernardo reference prior if \( n \geq 2 \). This leads to the conclusion that the minimal training sample size is equal to \( m = 2 \). Finally, the marginal distributions of \( \mathbf{x} \) and the training sample \( \mathbf{x}_\ell \) are presented in Theorem 3.2.
The proof of Theorem 3.2 is given in Supplement A. The marginal distributions of the whole sample and the training sample are presented as one-dimensional integrals that cannot be derived analytically. However, they can be computed with a high precision through numerical integration, for example, by using the Simpson rule [see Givens and Hoeting (2012)].

Since the range of \( \tau_{RE} \) is from 0 to \(+\infty\), we make a transformation under the integrals in (10) and (12), following Bodnar, Muhumuza and Possolo (2020), defined by \( \tau_{RE} = \tan(\omega) \), with the Jacobian equal to \( |d\omega/\tan(\omega)| = 1/\cos(\omega)^2 \). It is a one-to-one map of the interval \( \omega \in (0, \pi/2) \) to \( \tau_{RE} \in (0, \infty) \). This allows us to compute the integral over the bounded range \( \omega \in (0, \pi/2) \) rather than over the unbounded range \( \tau_{RE} \in (0, \infty) \). In the case of (10), we then get

\[
m(x|M_{RE}) = \int_0^{\pi/2} (2\pi)^{-n/2} (\det(U + \tan(\omega)^2 I))^{-1/2} \exp\left(-\frac{1}{2} x^T Q(\tan^2(\omega)) x \right) \frac{d\omega}{\cos^2(\omega)}
\]

with \( Q(\tan^2(\omega)) \) defined in (11). A similar transformation is also used in the computation of the integral in (12).

3.3. Location-scale model versus random effects model. The difference between the location-scale model and the random effects model lies in how they account for the heterogeneity that arises when the results of several studies are combined. The heterogeneity can be modeled either as a multiplicative correction factor (location-scale model) or as an additive correction term (random effects model). To choose between the location-scale model and the random effects model, we apply the IBF presented in Section 2.1 along with the marginal distributions of the whole sample and the training sample derived in Theorems 3.1 and 3.2.
The IBF for comparing the random effects model ($M_{RE}$) to the location-scale model ($M_{LS}$) using the training sample $x_\ell$ is given by

$$B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell) = B_{M_{RE}M_{LS}}^N(x) \times B_{M_{LS}M_{RE}}^N(x_\ell) = \frac{m(x|M_{RE})}{m(x|M_{LS})} \times \frac{m(x_\ell|M_{LS})}{m(x_\ell|M_{RE})},$$

where $m(x|M_{RE})$, $m(x|M_{LS})$, $m(x_\ell|M_{LS})$, and $m(x_\ell|M_{RE})$ are given in (10), (4), (6), and (12), respectively. A value of $B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell)$ that is larger than one indicates that the random effects model is preferable, while the inequality $B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell) < 1$ suggests that the application of the location-scale model is preferable.

Both the location-scale model and the random effects model have been shown to have a minimal training sample size of $m = 2$ (see Theorems 3.1 and 3.2). With a total sample size of $n$, we have a total of $n(n-1)/2$ combinations of two observations to choose from. Let $X_\ell = (x_1^\ell, x_2^\ell, \ldots, x_{L-1}^\ell, x_L^\ell)$ be the set of training samples; there are $L = n(n-1)/2$ elements in $X_\ell$. The IBF (14) is computed for every possible combination of two observations and is aggregated over the possible training samples by computing the average or the median. The reason for this is that the mean and median of the IBF over $X_\ell$ have increased stability in comparison with the IBF of an arbitrary $x_i^\ell$.

Berger and Pericchi (2001, p. 149) noted that it is recommended to put the more “complex” model in the numerator of (14) when the average of the IBF is computed over the training samples. One of the reasons for this is that large values of the IBF computed for certain training samples might dominate the resulting value of the average IBF, although most of the IBFs are close to zero. To symmetrize the impact of large and small values of the IBF computed for each of the training samples, the logarithmic transformation is used before calculating the average IBF and the median IBF. This leads to the following formulas for the average IBF and the median IBF:

$$aB_{M_{RE}M_{LS}}^I = \frac{1}{L} \sum_{\ell} \log \left( B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell) \right)$$

and

$$mB_{M_{RE}M_{LS}}^I = \text{median} \left( \log \left( B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell) \right) \right),$$

respectively.

We also computed the empirical probability that the random effects model will be preferable, which is expressed as

$$epB_{M_{RE}M_{LS}}^I = \frac{1}{L} \sum_{\ell} 1_{(0,\infty)} \left( \log \left( B_{M_{RE}M_{LS}}^I(x_\ell|x_\ell) \right) \right),$$

where $1_A(\cdot)$ is the indicator function of set $A$.

Based on the computed values in (15), (16), and (17), the random effects model is more suitable than the location-scale model if $aB_{M_{RE}M_{LS}}^I > 0$, $mB_{M_{RE}M_{LS}}^I > 0$, or $epB_{M_{RE}M_{LS}}^I > 0.5$, depending on the selected criterion. Otherwise, the location-scale model is preferable.

4. Simulation study. In this section, we investigate the performance of the Bayesian model selection procedure based on the intrinsic Bayes factors described in Section 3.3 by drawing samples from the location-scale model and the random effects model, respectively. By doing this, we can check to see if the model selection procedure based on the considered IBFs leads to the model from which the data were generated and we can study the impact of the sample size on the decision.

The following two simulation scenarios are considered:
Fig 2. Average of the logarithms of intrinsic Bayes factors (15) used to compare the random effects model and the location-scale model as a function of $\tau_{LS}$ when the reference prior is employed. We set $n \in \{5,10,15,20\}$ and $r \in \{-0.8,-0.4,0,0.4,0.8\}$. The data were drawn from the location-scale model (see Scenario 1).

- **Scenario 1**: draw a sample from the location-scale model $x = \mu 1 + \varepsilon_{LS}$, where $\varepsilon_{LS} \sim N(0, \tau_{LS}^2 U)$;
- **Scenario 2**: draw a sample from the random effects model $x = \mu 1 + \lambda_{RE} + \varepsilon_{RE}$, where $\varepsilon_{RE} \sim N(0, U)$ and $\lambda_{RE} \sim N(0, \tau_{RE}^2 I)$.

In both scenarios, we set $\mu = 0$ and consider several sample sizes $n \in \{5,10,15,20\}$. The square roots of the diagonal elements of the matrix $\mathbf{U} = (u_{ij})_{i,j=1,...,n}$ are drawn from the uniform distribution on $[0,1]$, while the nondiagonal elements of $\mathbf{U}$ are set to $u_{ij} = r^{\lfloor |i-j| \rfloor} \sqrt{u_{ii} u_{jj}}$ for $i,j \in 1,...,n$ and $i \neq j$, with $r \in \{-0.8,-0.4,0,0.4,0.8\}$. The structure of the matrix $\mathbf{U}$ corresponds to the structure of an autoregressive model with an autoregressive parameter equal to $r$. Several values of $\tau_{LS}$ and $\tau_{RE}$ are considered, namely $\tau_{LS} \in \{1.0,1.2,1.5,2.0,2.5,3.0\}$ and $\tau_{RE} \in \{0.0,0.25,0.5,1.0,1.5,2.0\}$. The results in Figures 2–7 are based on 10000 independent repetitions. For $\tau_{LS} = 1$ and $\tau_{RE} = 0$, the two models coincide, and they both correspond to the case in which dark uncertainty is absent.

Figures 2–4 present the values of the average IBF, the median IBF, and the empirical probability IBF defined in (15), (16), and (17), respectively, and computed for the data generated from the location-scale model (Scenario 1). Similar values were obtained under Scenario 2, which involved drawing the data from the random effects model; these values are depicted in Figures 5–7. In almost all of the considered cases, we observe that the average IBF and the median IBF are negative and the empirical probability IBF is smaller than 0.5 when the data are generated from the location-scale model; meanwhile, the average IBF and median
IBF are positive and the empirical probability IBF is larger than 0.5 when the data are drawn from the random effects model, thus supporting the model selection procedure based on the IBFs defined in Section 3.3. Some minor deviations from this observation are present only when \( n = 5 \), the data are drawn from the location-scale model, and the values of \( \tau_{LS} \) are not large for \( r \neq 0 \). Interestingly, when \( n = 5 \) and \( r = 0.0 \), the random effects model is preferred regardless of whether the data were generated following Scenario 1 or Scenario 2.

In general, we conclude that the model selection procedure with the intrinsic Bayes factor detects the random effects model when it is the true model considerably more often than it detects the location-scale model when it is the true model. This finding is in line with the results of Bodnar et al. (2016), who showed that the random effects model is more robust to model misspecification. As the sample size increases, the performance of the three considered model selection criteria based on the IBF improves. The absolute values of the average IBF and the median IBF become larger, while the empirical probability IBF is close to one for Scenario 2 and close to 0 for Scenario 1. For \( n = 10 \), the values of Figure 7 indicate that the random effects model is correctly chosen for almost all of the considered values when \( r \in \{-0.8, -0.4, 0, 0.4, 0.8\} \) and it is correctly chosen 80% of the time for \( r \in \{0, 0.4\} \); meanwhile, the probability of correctly specifying the location-scale model is between 70% and 80%, and it is slightly smaller for \( r = -0.4 \). The results for \( n = 15 \) and \( n = 20 \) are even stronger. For example, the random effects model is chosen with a probability larger than 0.9 for both \( r = 0 \) and \( r = 0.4 \) for all considered values \( \tau_{RE} > 0 \) (cf. Figure 7). Finally, we note that

![Graphs showing the median of log(IBF) for different sample sizes and correlation coefficients](image)

**Fig 3.** Median of the logarithms of intrinsic Bayes factors (16) used to compare the random effects model and the location-scale model as a function of \( \tau_{LS} \) when the reference prior is employed. We set \( n \in \{5, 10, 15, 20\} \) and \( r \in \{-0.8, -0.4, 0, 0.4, 0.8\} \). The data were drawn from the location-scale model (see Scenario 1).
Empirical probability of the logarithms of intrinsic Bayes factors \((17)\) used to compare the random effects model and the location-scale model as a function of \(\tau_{LS}\) when the reference prior is employed. We set \(n \in \{5, 10, 15, 20\}\) and \(r \in \{-0.8, -0.4, 0, 0.4, 0.8\}\). The data were drawn from the location-scale model (see Scenario 1).

the correlation coefficient \(r\) has only a minor impact on the computed values of the intrinsic Bayes factors.

Additional results of the simulation study concerning model selection with the IBF using the Jeffreys prior and the AIC are presented in Supplement B and Supplement C, respectively. The results obtained in both supplements are similar to those depicted in Figures 2–7, except when data were drawn from the location-scale model with values of \(\tau_{LS}\) close to one. In this particular case, independent of \(n\), the IBF under the Jeffreys prior erroneously selects the random effects model.

5. Modeling dark uncertainty in measurements of the Newtonian constant of gravitation. The Newtonian constant of gravitation appears as \(G\) in Newton’s third law, which characterizes the gravitational force between two bodies with masses \(m_1\) and \(m_2\) whose centers of mass are a distance \(r\) apart: \(F = G \frac{m_1 m_2}{r^2}\). If both masses are 1 kg and \(r = 1\) m, the gravitational force between the masses is \(6.67 \times 10^{-11}\) N.

The sixteen measurements of the Newtonian constant of gravitation that we will use for illustration are given in Table 1. We will also use the correlations between some of them, which are listed in Tiesinga et al. (2021, Table XXIX). We use the following correlations: \(\text{cor}(\text{NIST-82, LANL-97}) = 0.351\), \(\text{cor}(\text{HUST-05, HUST-09}) = 0.134\), and \(\text{cor}(\text{HUST-09, HUST-TOS} - 18) = 0.068\).
Both the location-scale model and the random effects model accommodate the dark uncertainty shown in Figure 1. Both models have been used previously for this purpose in Mohr, Newell and Taylor (2016) and Bodnar, Muhumuza and Possolo (2020), among others. We use the Kolmogorov–Smirnov test to ascertain the adequacy of both models for these data. For the location-scale model, the test yields a $p$-value of 0.6, and for the random effects model, it yields 0.12, thus suggesting that both models cannot be rejected at significance level of 10%. However, since there are only sixteen observations and the parameters have been estimated from these data, the Kolmogorov-Smirnov test likely will not have much power to detect the alternative. Even though the Kolmogorov-Smirnov test does not reject any of two models, it does not automatically mean that the models are fit for purpose following the famous quote “All models are wrong, but some are useful” by George Box (see, Box (1976)). As such, the predictive ability of the models should be investigated together with the model selection procedure.

To select one of the two models, we first assign the Berger and Bernardo reference prior to the parameters of the two models and perform Bayesian model selection based on the intrinsic Bayesian factor. The application of a noninformative reference prior is motivated by the absence of information about the parameters of the two models that could be used to determine an informative prior. Using the fact that the size of the minimal training sample is $m = 2$, we compute the intrinsic Bayes factor for all 120 possible specifications of the training sample consisting of two measurement results.
The resulting values of the IBF used to compare the random effects model and the location-scale model are depicted in the plot on the left side of Figure 8. We observe that in most cases, the IBF is negative, meaning that the location-scale model is preferable. Only for 13 training samples out of 120 possible training samples is the random effects model chosen. This leads to the following value of the empirical probability IBF: $\text{epB}_{\text{IMREMLs}} = 0.1083$. The other two measures for Bayesian model selection discussed in Section 3.3 are negative and equal to the following values:

$$aB_{\text{IMREMLs}}^I = -0.6814 \quad \text{and} \quad mB_{\text{IMREMLs}}^I = -0.8777,$$

which supports the application of the location-scale model to fitting the heterogeneity in the measurements of the Newtonian constant of gravitation.

We also obtained results by performing model selection based on the intrinsic Bayes factor under the Jeffreys prior, as described in Supplement B. The values of the IBF obtained when the Jeffreys prior is employed are shown in the plot on the right side of Figure 8. The random effects model is chosen for 24 out of 120 possible training samples, and the empirical probability IBF is equal to the following value: $\text{epB}_{\text{IMREMLs}} = 0.2$. Finally, the other two measures are negative and they have the following values: $aB_{\text{IMREMLs}}^I = -0.4854$ and $mB_{\text{IMREMLs}}^I = -1.0657$. As a result, all considered criteria consistently select the location-scale model.

This conclusion is also obtained when the Akaike information criterion (see Supplement C) is used to compare the location-scale model with the random effects model based on
FIG 7. Empirical probability of the logarithms of intrinsic Bayes factors (17) used to compare the random effects model and the location-scale model as a function of $\tau_{RE}$ when the reference prior is employed. We set $n \in \{5, 10, 15, 20\}$ and $r \in \{-0.8, -0.4, 0, 0.4, 0.8\}$. The data were drawn from the random effects model (see Scenario 2).

Berger and Bernardo reference prior

Jeffreys prior

FIG 8. Logarithm of intrinsic Bayes factors computed for each of the $L = 120$ minimal training samples of size $m = 2$ for the measurements of the Newtonian constant of gravitation.

the likelihoods of the two models. The difference between the AIC values computed for the random effects model and the location-scale model is $dAIC(M_{RE}, M_{LS}) = -45.054$, indicating that the location-scale model is selected by the considered frequentist approach.
Measurements of the Newtonian constant of gravitation $G$ together with their uncertainties, which are expressed as standard errors. These results are listed in Table 1 of Merkatas et al. (2019), which also explains the labels of the results. Tiesinga et al. (2021, Table XXIX) provides the following non-zero correlations between some of the measurement results: $\text{cor}(\text{NIST} - 82, \text{LANL} - 97) = 0.351$, $\text{cor}(\text{HUST} - 05, \text{HUST} - 09) = 0.134$, and $\text{cor}(\text{HUST} - 09, \text{HUST} - \text{TOS} - 18) = 0.068$.

| Study     | Measurement ($G/10^{-11}$) | Uncertainty ($G/10^{-11}$) |
|-----------|----------------------------|----------------------------|
| NIST-82   | 6.67248                    | 0.00043                    |
| TR&D-96   | 6.6729                     | 0.00050                    |
| LANL-97   | 6.67398                    | 0.00070                    |
| UWash-00  | 6.674255                   | 0.00092                    |
| BIPM-01   | 6.67559                    | 0.00027                    |
| UWup-02   | 6.67422                    | 0.00098                    |
| MSL-03    | 6.67387                    | 0.00027                    |
| HUST-05   | 6.67222                    | 0.00087                    |
| UZur-06   | 6.67425                    | 0.00012                    |
| HUST-09   | 6.67349                    | 0.00018                    |
| JILA-10   | 6.67260                    | 0.00025                    |
| BIPM-14   | 6.67554                    | 0.00016                    |
| LENS-14   | 6.67191                    | 0.00099                    |
| UCI-14    | 6.67435                    | 0.00013                    |
| HUST-TOS-18 | 6.674184               | 0.000078                  |
| HUST-AAF-18 | 6.674484              | 0.000078                  |

Finally, the discrepancies between the empirical data and model predictions obtained as point estimators from the posterior predictive distributions, together with the posterior predictive credible intervals, are depicted in Figure 9 for the Berger and Bernardo reference prior (left) and for the Jeffreys prior (right). The formulas used in the computations are presented in Supplement D. The observed values of the measurement results are inside the 95% posterior predictive credible intervals, with the exception of the measurement result BIMP-14. The lengths of the posterior predictive credible intervals derived under the location-scale model vary more than the lengths of the intervals computed for the random effects model. Moreover, the intervals are considerably smaller for laboratories that reported small standard uncertainties. Such a property of the posterior predictive credible intervals derived under the
location-scale model leads to better performance in comparison to the random effects model by allowing laboratories that reported small uncertainties and provided measurement results close to the predicted values to have more influence on the computation of the model selection criteria. In contrast, the random effects model reduces the impact of the reported uncertainties by making the posterior predictive credible intervals relatively large for all laboratories.

6. Summary. In many applications, the variability of individual studies that are pooled together to determine a consensus value cannot be explained by the reported variabilities of each study. This indicates the presence of heterogeneity, also known as dark uncertainty. One can also treat the results of each individual study as an expert witness about the parameter $\mu$ and use the linear pooling method, as discussed in Genest and Schervish (1985).

Endowing the parameters of the location-scale model and the random effects model with the Berger and Bernardo reference prior and the Jeffreys prior, we derive the expression of the intrinsic Bayes factor to compare the random effects model and the location-scale model. In the case of the location-scale model, analytical expressions for the marginal distributions of the whole data set and the training data are derived, while one-dimensional integrals representing the marginal distributions are obtained in the case of the random effects model. These integrals can be computed numerically using Simpson’s rule. Note that even though we apply Bayesian methods, no Monte Carlo Markov Chain approach is needed. The performance of the Bayesian model selection procedures based on the Berger and Bernardo reference prior and the Jeffreys prior, and the performance of the frequentist method based on the Akaike information criterion, were investigated using an extensive simulation study.

The derived theoretical findings are applied to measurements of the Newtonian constant of gravitation. The results of the empirical study support the application of the location-scale model to the computation of the Newtonian constant of gravitation. Both the average IBF and the median IBF are significantly smaller than zero, while the empirical probability IBF is less than 11% under the reference prior and equal to 20% under the Jeffreys prior, indicating that for the majority of the specified training samples, the location-scale model is preferred to the random effects model. The same decision is made when the model selection is performed using the AIC values. As such, our findings support the application of the Birge ratio method to the adjustment of fundamental physical constants, just as it is currently used in the computation of the CODATA 2018 values of fundamental constants [see Tiesinga et al. (2021)].

The developed methodology can be applied in many fields of science. For instance, it can be used for analysing the relative atomic masses in chemistry, for combining the results of interlaboratory studies in metrology, and for performing meta-analysis in medicine.

Acknowledgments. The authors would like to thank Professor Kafadar and two anonymous Referees for their constructive comments, which improved the quality of this paper.

This research was partially supported by the National Institute of Standards and Technology (NIST) Exchange Visitor Program. The first author is grateful to the Statistical Engineering Division of the National Institute of Standards and Technology (NIST) for providing an excellent and inspiring environment for research.

Funding. Olha Bodnar acknowledges valuable support from an internal grant (Rörlig resurs) from Örebro University.

SUPPLEMENTARY MATERIAL

Supplement A: Proofs of the theoretical results
In Supplement A, the proofs of Theorems 3.1 and 3.2 are presented.
Supplement B: Results of the simulation study using the Jeffreys prior
In Supplement B, additional results from the simulation study obtained by employing the Jeffreys prior for both models are presented.

Supplement C: Results of the simulation study based on the Akaike information criterion
In Supplement C, the problem of model selection is approached from the viewpoint of frequentist statistics. Here, we present the results of a simulation study in which the comparison of the two models is based on the Akaike information criterion.

Supplement D: Posterior predictive distribution
In Supplement D, the posterior predictive distributions for the location-scale model and the random effects model are derived by employing the Berger and Bernardo reference prior and the Jeffreys prior.

REFERENCES

ADES, A. E., LU, G. and HIGGINS, J. (2005). The interpretation of random-effects meta-analysis in decision models. *Medical Decision Making* **25** 646–654.

ALIGHANBARI, S., GIRI, G., CONSTANTIN, F. L., KOROBOV, V. and SCHILLER, S. (2020). Precise test of quantum electrodynamics and determination of fundamental constants with HD+ ions. *Nature* **581** 152–158.

BERGER, J. and BERNARDO, J. M. (1992). On the development of reference priors. In *Bayesian Statistics* (J. M. Bernardo, J. Berger, A. P. Dawid and A. F. M. Smith, eds.) **4** 35–60. Oxford University Press.

BERGER, J. O., BERNARDO, J. M. and SUN, D. (2009). The formal definition of reference priors. *The Annals of Statistics* **37** 905–938.

BERGER, J. O. and PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association* **91** 109–122.

BERGER, J. O. and PERICCHI, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In *Model Selection* (P. Lahiri, ed.). *Lecture Notes-Monograph Series* **38** 135–207. Institute of Mathematical Statistics, Hayward, CA.

BIRGE, R. T. (1932). The calculation of errors by the method of the least squares. *Physical Review* **40** 207–227.

BODNAR, O. and ELSTER, C. (2014a). Analytical derivation of the reference prior by sequential maximization of Shannon's mutual information in the multi-group parameter case. *Journal of Statistical Planning and Inference* **147** 106–116.

BODNAR, O. and ELSTER, C. (2014b). On the adjustment of inconsistent data using the Birge ratio. *Metrologia* **51** 516.

BODNAR, O. and ELSTER, C. (2020). Assessing laboratory effects in key comparisons with two transfer standards measured in two petals: A Bayesian approach. In *13th International Workshop on Intelligent Statistical Quality Control 2019, IWISQC 2019, 12 August 2019–14 August 2019* 1–18. City University of Hong Kong.

BODNAR, O., LINK, A. and ELSTER, C. (2016). Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis* **11** 25–45.

BODNAR, O., MUHUMUZA, R. N. and POSSOLO, A. (2020). Bayesian inference for heterogeneity in meta-analysis. *Metrologia* **57** 064004.

BODNAR, O., ELSTER, C., FISCHER, J., POSSOLO, A. and TOMAN, B. (2016). Evaluation of uncertainty in the adjustment of fundamental constants. *Metrologia* **53** 546.

BODNAR, O., LINK, A., ARENDACKÁ, B., POSSOLO, A. and ELSTER, C. (2017). Bayesian estimation in random effects meta-analysis using a non-informative prior. *Statistics in Medicine* **36** 378–399.

BOX, G. E. (1976). Science and statistics. *Journal of the American Statistical Association* **71** 791–799.

BROCKWELL, S. E. and GORDON, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine* **20** 825–840.

CLARKE, B. and YUAN, A. (2004). Partial information reference priors: Derivation and interpretations. *Journal of Statistical Planning and Inference* **123** 313–345.

FERNÁNDEZ, C. and STEEL, M. F. J. (1999). Reference priors for the general location-scale model. *Statistics & Probability Letters* **43** 377–384.

GENEST, C. and SCHERVISH, M. J. (1985). Modeling expert judgments for Bayesian updating. *The Annals of Statistics* **1198–1212.

GIVENS, G. H. and HOETING, J. A. (2012). *Computational Statistics* **710**. John Wiley & Sons.
PARTICLE DATA GROUP and ZYLA ET AL., P. A. (2020). Review of particle physics. Progress of Theoretical and Experimental Physics 2020 083C01.

GUOLO, A. and VARIN, C. (2017). Random-effects meta-analysis: The number of studies matters. Statistical Methods in Medical Research 26 1500–1518.

HANNIG, J., FENG, Q., IYER, H., WANG, C. and LIU, X. (2018). Fusion learning for inter-laboratory comparisons. Journal of Statistical Planning and Inference 195 64–79.

HARDY, R. J. and THOMPSON, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. Statistics in Medicine 17 841–856.

HELD, L. and BOYE, D. S. (2014). Applied Statistical Inference: Likelihood and Bayes. Springer Science & Business Media.

HIGGINS, J., THOMPSON, S. G. and SPIEGELHALTER, D. J. (2009). A re-evaluation of random-effects meta-analysis. Journal of the Royal Statistical Society: Ser. A (Statistics in Society) 172 137–159.

JEFFREYS, H. (1946). An invariant form for the prior probability in estimation problems. Proceedings of the Royal Society A 186 453–461.

JONES, H. E., ADES, A., SUTTON, A. J. and WELTON, N. J. (2018). Use of a random effects meta-analysis in the design and analysis of a new clinical trial. Statistics in Medicine 37 4665–4679.

KASS, R. E. and RAFTERY, A. E. (1995). Bayes factors. Journal of the American Statistical Association 90 773–795.

KOEPEKE, A., LAFARGE, T., POSSOLO, A. and TOMAN, B. (2017). Consensus building for interlaboratory studies, key comparisons, and meta-analyses. Metrologia 54 S34–S62.

LAMBERT, P. C., SUTTON, A. J., BURTON, P. R., ABRAMS, K. R. and JONES, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. Statistics in Medicine 24 2401–2428.

MANDEL, J. and PAULE, R. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. Analytical Chemistry 42 1194–1197.

MENJIA, J. and POSSOLO, A. (2017). Data reduction framework for standard atomic weights and isotopic compositions of the elements. Metrologia 54 229–238.

MERKATAS, C., TOMAN, B., POSSOLO, A. and SCHLAMMINGER, S. (2019). Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation. Metrologia 56 054001.

MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2016). CODATA recommended values of the fundamental physical constants: 2014. Reviews of Modern Physics 88 035009.

NEWELL, D. B., CABIAFI, F., FISCHER, J., FUJII, K., KARSHENBOIM, S. G., MARGOLIS, H. S., DE MIRANDES, E., MOHR, P. J., NEZ, F. and PACHUCKI, K. (2018). The CODATA 2017 values of h, e, k, and N_A for the revision of the SI. Metrologia 55 L13–L16.

O’HAGAN, A. (1995). Fractional Bayes factors for model comparison. Journal of the Royal Statistical Society: Series B (Methodological) 57 99–118.

PINHEIRO, J. C. and BATES, D. M. (2000). Mixed-Effects Models in S and S-Plus. Springer-Verlag, New York.

POSSOLO, A., KOEPEKE, A., NEWTON, D. and WINCHESTER, M. R. (2021). Decision tree for key comparisons. Journal of Research of the National Institute of Standards and Technology 126 126007.

ROTHLEITNER, C. and SCHLAMMINGER, S. (2017). Invited Review Article: Measurements of the Newtonian constant of gravitation, G. Review of Scientific Instruments 88 111101.

RUKHIN, A. (2003). Two procedures of meta-analysis in clinical trials and interlaboratory studies. Tatra Mountains Mathematical Publications 26 155–168.

RUKHIN, A. L. (2013). Estimating heterogeneity variance in meta-analysis. Journal of the Royal Statistical Society: Ser. B 75 451–469.

THOMPSON, M. and ELLISON, S. L. (2011). Dark uncertainty. Accreditation and Quality Assurance 16 483–487.

TIESINGA, E., MOHR, P. J., NEWELL, D. B. and TAYLOR, B. N. (2021). CODATA recommended values of the fundamental physical constants: 2018. Reviews of Modern Physics 93 025010.

TOMAN, B., FISCHER, J. and ELSTER, C. (2012). Alternative analyses of measurements of the Planck constant. Metrologia 49 567–571.

TOMAN, B. and POSSOLO, A. (2009). Laboratory effects models for interlaboratory comparisons. Accreditation and Quality Assurance 14 553–563.

TOMAN, B. and POSSOLO, A. (2010). Erratum to: Laboratory effects models for interlaboratory comparisons. Accreditation and Quality Assurance 15 653–654.

TURNER, R. M., JACKSON, D., WEI, Y., THOMPSON, S. G. and HIGGINS, J. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. Statistics in Medicine 34 984–998.
VERONIKI, A. A., JACKSON, D., BENDER, R., KUSS, O., LANGAN, D., HIGGINS, J. P., KNAPP, G. and SALANTI, G. (2019). Methods to calculate uncertainty in the estimated overall effect size from a random-effects meta-analysis. *Research Synthesis Methods* 10 23–43.

WEISE, K. and WÖGER, W. (2000). Removing model and data non-conformity in measurement evaluation. *Measurement Science and Technology* 11 1649.