metrology

MDPI

Article

The Statistical Power and Confidence of Some Key Comparison Analysis Methods to Correctly Identify Participant Bias

Ellie Molloy, Annette Koo *, Blair D. Hall and Rebecca Harding

Measurement Standards Laboratory of New Zealand, Lower Hutt 5010, New Zealand; ellie.molloy@callaghaninnovation.govt.nz (E.M.); blair.hall@measurement.govt.nz (B.D.H.); harding.r@wehi.edu.au (R.H.)

* Correspondence: annette.koo@measurement.govt.nz

Abstract: The validity of calibration and measurement capability (CMC) claims by national metrology institutes is supported by the results of international measurement comparisons. Many methods of comparison analysis are described in the literature and some have been recommended by CIPM Consultative Committees. However, the power of various methods to correctly identify biased results is not well understood. In this work, the statistical power and confidence of some methods of interest to the CIPM Consultative Committees were assessed using synthetic data sets with known properties. Our results show that the common mean model with largest consistent subset delivers the highest statistical power under conditions likely to prevail in mature technical fields, where most participants are in agreement and CMC claims can reasonably be supported by the results of the comparison. Our approach to testing methods is easily applicable to other comparison scenarios or analysis methods and will help the metrology community to choose appropriate analysis methods for comparisons in mature technical fields.

Keywords: key comparison; degrees of equivalence; statistical power

Citation: Molloy, E.; Koo, A.; Hall, B.D.; Harding, R. The Statistical Power and Confidence of Some Key Comparison Analysis Methods to Correctly Identify Participant Bias. *Metrology* **2021**, *1*, 52–73. https://doi.org/10.3390/metrology1010004

Academic Editor: Han Haitjema

Received: 27 July 2021
Accepted: 23 August 2021
Published: 26 August 2021

1. Introduction

The CIPM Mutual Recognition Arrangement (MRA) [1] is the framework through which national metrology institutes (NMIs) demonstrate the equivalence of their measurement standards and the calibration and measurement certificates they issue. The MRA provides reliable quantitative information on the comparability of national metrology services. A database of NMIs' internationally recognised calibration and measurement capabilities (CMCs) is published by the International Bureau of Weights and Measures (BIPM). To maintain or extend CMC entries in this database, NMIs must be able to demonstrate the validity of their claims for measurement capabilities.

The main way that NMIs support CMC claims is to participate in international measurement comparisons. In these "key comparisons", a group of NMIs each measure a common artefact and then apply an agreed analysis method to the results. A consensus reference value for the measured quantity is determined and values of "degrees of equivalence" (DoEs), with associated uncertainties, are calculated for each participant. Each DoE is a measure of the difference between the participant laboratory's measurement and the comparison reference value, which may be expected to reflect the corresponding consistency between participants' national standards. In this way, users of the CMC database may have a reasonable expectation that calibration of an artefact would produce equivalent results, to within the claimed expanded uncertainties, when carried out in different economies.

DoEs are used to assess the equivalence of participants' measurement standards and thereby to provide evidence for the validity of CMC claims. A claim is assessed by comparing the value of a DoE with its uncertainty: if the magnitude of the DoE is less than the expanded uncertainty, then the claim can be accepted, but if not, the evidence is

considered weak and the claim is likely to be rejected. It is therefore vital that the DoE accurately characterises the metrological equivalence of participants. This assessment can be understood as a statistical hypothesis test, with the null hypothesis being that a participant has carried out an adequate uncertainty analysis with no unrecognised effects that would bias the measurements [2,3].

The determination of DoEs and the assessment of equivalence depend on the data produced by a key comparison and the chosen method of data analysis. Ideally, an analysis would correctly distinguish between the participants that have problems with their measurement and the other participants, whose measurements are equivalent. Unfortunately, such a distinction cannot be guaranteed and undesirable outcomes are possible in practice. The rate at which measurements are correctly deemed acceptable is related to statistical confidence—the probability that an unbiased participant measurement is accepted. The rate at which measurements are correctly deemed unacceptable is related to the statistical power of the hypothesis test—the probability that participant measurement bias is detected.

A method of analysis of key comparison data that minimises undesirable outcomes has good balance between high statistical power and good statistical confidence. The likelihood of incorrectly rejecting a participant should be acceptably low because this outcome imposes an unwarranted burden on the laboratory whose measurement is called into question, in terms of needing to generate further evidence to support a CMC claim. Also, the likelihood of incorrectly accepting a participant should be low because the detection of unacknowledged laboratory bias is a primary reason for carrying out comparisons. There has been a lot of discussion about how to determine DoEs and, since the seminal paper by Cox [4], many variants have been proposed to deal with inconsistent data sets [5–7] or to offer statistically distinct approaches or models [3,8–11]. This proliferation of options demonstrates two things: first, that a single analysis method is unlikely to be suitable for all comparisons; and second, that there are few tools available to compare one method with another in any given comparison scenario. In particular, apart from one study that looked at artefact drift during a comparison [12], little attention has been paid to the statistical power of various methods.

The work reported here was undertaken to address this shortfall and give the community of comparison analysts and Consultative Committees an objective means of assessing the fitness of an ever-increasing number of analysis methods in the literature to solve the comparison problem with respect to statistical power. It looks at the determination of 'equivalence' implied by CMC claims; the equivalence of scales, so far as a user is concerned. It also sets out a robust and easily replicated approach for carrying out an assessment of other methods of current interest, methods that are proposed in the future, or any method under various conditions.

We selected a number of comparison analysis methods and used numerical simulation to generate many sets of synthetic comparison data with known properties. This allowed us to calculate long-run success rates to compare the various methods. The approach discriminates clearly between methods and delivers compelling evidence to favour particular methods over others. Understanding the statistical behaviour of selected methods enhances confidence in the information provided under the CIPM MRA. It also allows us to understand how the assumptions underlying various models impact the judgement of CMC claims. Therefore, in this paper, we consider in particular the implications of introducing a 'dark uncertainty' in some methods to account for inconsistency of results with the reference value.

This paper is structured as follows. Section 2 reviews the range of methods recommended by current Consultative Committee guidelines, presents the statistical models associated with comparison data, then gives a brief description of the various methods of comparison analysis that have been selected for testing in this work. In Section 3, we present our method for generating synthetic data sets and the conditions under which our testing and results can be usefully applied. Section 4 presents the results obtained, comparing the power of various methods to detect biased measurements and protect unbi-

ased participants under the range of conditions tested. Section 5 discusses our findings, identifying the strengths and weaknesses of the various methods, examining the impact of some of the assumptions in some methods, addressing the value of this approach to testing methods, and examining the validity of the chosen conditions of test. We give some conclusions in Section 6. There are also two appendices. The first contains further information about one of the methods used and the second contains a full set of simulation results to allow those with interest in a particular method, for example, to look at the detail of its performance under all test conditions.

2. Analysis of Measurement Comparisons

We have selected a number of comparison analysis methods to study, which are representative of those being used or considered by the members of CIPM Consultative Committees. Some Committees make explicit recommendations while others give guidance and options. Again, the disparity within the community is evidence of the difficulty in selecting a method, which this work hopes to alleviate.

In this section, we begin, in Section 2.1, by identifying three statistical models—the consensus, or common mean model [4], the fixed-effects model [6,8,13,14], and the random-effects model [6,15,16]—which underlie particular methods of comparison analysis. Then, in Section 2.2, we describe the methods of analysis that we have chosen to test in this work and identify how each relates to guidance from the various Consultative Committees.

We examined the following methods: the common mean model method [4], the common mean model with largest consistent subset [5], the common mean model with cut-off weighting [17], the common mean model with exclusion of obvious outliers [17], the fixed-effects model with a weighted mean [6] (which gives the same result as the common mean model method), the fixed-effects model with Bayesian model averaging [7], the random-effects model with the method of Mandel and Paule to achieve consistency [16], two other methods for random-effects models implemented by the NIST Consensus Builder, DerSimonian–Laird and Hierarchical Bayesian and the Linear Pool method also implemented by the NIST Consensus Builder [15,18].

The methods chosen are representative but by no means exhaustive; the procedure for testing methods applied here can easily be applied to other methods proposed in the literature or found in Committee guidance.

2.1. Models of Error in Measurement Comparisons

The various comparison analysis methods studied here are each based on one of three commonly used statistical models for comparisons. Before describing the analysis methods in Section 2.2, we summarise these models.

2.1.1. Common Mean Model (CM)

In the common mean model [4,19], all participants measure the same quantity and there is no assumed bias. The model is, in the case of a single artefact and one measurement per participant, expressed as

$$y_i = \mu + e_i \ , \tag{1}$$

where y_i is the value reported by participant i, μ is the unknown true value of the artefact (the *measurand*), and e_i is an unknown measurement error. The best estimate of e_i is zero, but a standard uncertainty u_i is reported by the participant characterising the dispersion of values generated by the measurement.

The best linear unbiased estimate of μ is

$$y_{\mathrm{CM}} = \sum_{i=1}^{P} w_i y_i \ , \tag{2}$$

with

$$w_i = \frac{u_i^{-2}}{\sum_{i=1}^{P} u_i^{-2}}\,, \tag{3}$$

where P is the number of participants and w_i is the weighting of participant i. The DoEs are then the differences between y_i and y_{CM}.

2.1.2. Fixed-Effects Model (FE)

The fixed-effects model [6,8,13,14] includes an unknown systematic effect δ_i biasing each laboratory's measurements,

$$y_i = \mu + \delta_i + e_i\,. \tag{4}$$

This model requires an additional constraint to determine a unique solution. The DoEs, which are the differences between the y_i and the measurand estimate, are considered estimates of the δ_i. The inclusion of the bias parameter distinguishes the fixed-effects model from the common mean model.

2.1.3. Random-Effects Model (RE)

The third comparison model considered is the random-effects model [6,15]:

$$y_i = \mu + b_i + e_i\,, \tag{5}$$

where all the b_i are now random variates drawn from a single Gaussian distribution with a mean of zero and variance τ^2.

When a random-effects model is used to analyse measurement comparison data, the estimate of μ is

$$y_{RE} = \sum_{i=1}^{P} w_{i,\mathrm{RE}}\, y_i\,, \tag{6}$$

with weights adjusted by τ:

$$w_{i,\mathrm{RE}} = \frac{(u_i^2 + \tau^2)^{-1}}{\sum_{i=1}^{P} (u_i^2 + \tau^2)^{-1}}\,. \tag{7}$$

The DoEs, as before, are the differences between the y_i and y_{RE}.

2.2. Comparison Analysis Methods

The common mean model can be used directly as a method and is applicable when the errors e_i are normally distributed and the associated uncertainties u_i correctly characterise their dispersion (i.e., when the analyses of measurement uncertainty producing the u_i were correct). Six of CIPM Consultative Committees giving guidance to their members on the analysis of comparison data recommend the use of the common mean method as a default (the Consultative Committees for Mass and Related Quantities (CCM) [20], Length (CCL) [21], Amount of Substance (CCQM) [22], Electricity and Magnetism (CCEM) [23], Acoustics, Ultrasound, and Vibration (CCAUV) [24], and Photometry and Radiometry (CCPR) [17]). (Note that the Consultative Committee for Amount of Substance (CCQM) also allows the use of the median and the Consultative Committee for Ionizing Radiation recommends the "power-moderate mean" [25]). However, the assumptions that make this method applicable are exactly those a comparison analysis is intended to verify. Given the very real possibility that the assumptions will be violated, several modifications to this basic method have been proposed. We now briefly describe those considered in this study.

2.2.1. Common Mean with Largest Consistent Subset (CM-LCS)

Cox has proposed that only the largest consistent subset of participant measurements be used to calculate the value of the artefact with Equation (2) [5]. All possible subsets of participants are taken into consideration when determining the largest consistent subset. For P participants, all $\binom{P}{P-1}$ subsets of size $P-1$ are considered first, then subsets of size $P-2$, and so on. A consensus value is found for each subset, the DoEs are evaluated, and a value of χ^2 is calculated. This χ^2 is compared to the value $\chi^2_\nu(p)$ expected for normally distributed data with ν degrees of freedom and significance level p. Results are considered consistent if χ^2 is less than $\chi^2_\nu(p)$.

The largest consistent subset is the largest subset for which the value of χ^2 is acceptably low. If two consistent subsets have the same size, the one with the lower value for χ^2 is chosen.

Many of the Consultative Committee guidelines recognise that consistency of results obtained from a simple common mean model cannot be assumed. For example, the guidelines of the CCQM, CCL, CCAUV, CCEM, and CCPR all recommend the use of a chi-squared test or Birge ratio test at the 95% level to check consistency. The CCM and the CCEM account for any inconsistency by recommending the Procedure A proposed by Cox in [4], in which participants are invited to self-exclude or to resubmit results. And the second option given by the CCL recommends an iterative variant with outlier identification and exclusion. While the algorithm for identifying excluded results is different, CM-LCS is analogous to these approaches in that results are excluded until a global consistency statistic at the 95% significance level is satisfied.

2.2.2. Common Mean with Cut-Off (CM-CO)

One of the recommendations of the CCPR is intended to limit the influence of measurements reported with very low values of u_i on the reference value. The recommendation sets a lower limit on the values of uncertainty used to calculate weights in Equation (3). The lower limit is equal to the mean of the u_i values less than or equal to the median of all reported uncertainties,

$$u_{\text{cut}} = \text{mean}(u_i) \quad \text{for all} \quad u_i \leq \text{median}(u_i)\ . \tag{8}$$

The weight attributed to each participant in Equation (3) is then the greater of u_i and u_{cut}. Note, however, the uncertainty of the measurement is not changed.

2.2.3. Common Mean with Exclusion of Obvious Outliers (CM-OO)

Another CCPR recommendation, intended to prevent clearly discrepant results from affecting a comparison analysis, is to exclude outliers from the evaluation of Equation (2). The guidelines consider an 'obvious outlier' to be any measurement result where the magnitude of the DoE is more than 6-fold its associated standard uncertainty i.e., at the coverage factor $k = 6$ level.

We can compare this method with that of CM-LCS which excludes results at the $k = 2$ level and which we also test in this work. Although we do not test it here, we might expect the CCQM-recommended method of outlier exclusion at the 99% significance level ($k = 3$) would give results intermediate between CM-LCS and CM-OO. There is not a clear distinction in this study between a biased result and an outlier because we are only considering the case where there is one result per participant and a single artefact.

2.2.4. Fixed Effects with Weighted Mean (FE-WM)

As mentioned above, a unique solution to Equation (4) can only be found if an additional constraint is added to the model. If Equations (2) and (3) are used as the constraint, we call this method the fixed-effects model with weighted-mean. Under the conditions considered in this paper (a single artefact and one reported measurement per participant), the DoEs, which are the differences between y_i and y_{CM}, are estimates of the δ_i

and are equal to the estimates obtained from the common mean model analysis, so results using either approach will be labelled CM/FE-WM.

2.2.5. Fixed Effects with Bayesian Model Averaging (FE-BMA)

A Bayesian analysis for the fixed-effects model has been proposed by Elster and Toman [7] and demonstrated using the results of the CCPR-K2c.2003 comparison [26]. Although this method was not used to obtain the official results, the report for CCPR-K2c.2003 [26] also presented the result of analysis using FE-BMA.

This method uses a modification of the weighted mean constraint. It assumes there is a subset of at least m unbiased participants and considers every possible subset of m participants. The biases of the participants in each subset are assumed to be zero ($\delta_i = 0$) and a weighted mean of these participants' results is evaluated. For the remaining participants, the difference between this mean and the reported y_i then estimates the bias. The results obtained for different subsets are combined using Bayesian model averaging to account for the probability that a subset is appropriate given the data. This produces posterior probability density functions (PDFs) for each laboratory bias. The means of these PDFs are taken as the DoEs and the symmetric 95% credible intervals as the expanded uncertainties of the DoEs. With this approach, each participant is effectively compared to a different reference value, so the bilateral degrees of equivalence between two participants cannot be found simply from the difference between the corresponding unilateral DoEs as might be expected of other methods [7].

The selection of the order parameter m depends on the data being analysed. One possibility is to evaluate the largest coherent subset; that is, the largest subset in which all members are pairwise equivalent (see [7] for a description of the largest coherent subset and its calculation). In this work, $m = LCHS - 4$, where $LCHS$ is the number of participants in the largest coherent subset. Further detail about the selection of m and the evaluation of expanded uncertainty intervals for the DoEs using Bayesian model averaging is provided in Appendix A.

2.2.6. Random Effects with Mandel–Paule (RE-MP)

A method of determining the value of τ in Equation (7) was proposed by Mandel and Paule [16]. The method evaluates χ^2 for the solution when τ is 0, and, if χ^2 exceeds $\chi^2_\nu(p)$, an iterative process is used to find a value of τ which delivers weights that lead to a sufficiently low value of χ^2.

Both the CCQM [22] and the CCPR [17] recommend the use of RE-MP to account for inconsistency at the 95% level. In the CCQM documents, this is one of many options while in the CCPR it is the preferred option.

2.2.7. Random Effects with DerSimonian and Laird (RE-DL)

The online NIST Consensus Builder implements two random-effects model methods. The first is the DerSimonian and Laird [15,27,28] method, which finds an estimate of τ that satisfies

$$\begin{aligned} \widehat{\tau}^2 &= \max(0, \widehat{\tau}^2_M) \,, \\ \text{where } \widehat{\tau}^2_M &= \frac{(Q - n + 1)}{\sum u_i^{-2} - \frac{\sum u_i^{-4}}{\sum u_i^{-2}}} \,, \\ \text{and } Q &= \sum u_i^{-2} (y_i - \widehat{\mu})^2 \,. \end{aligned} \tag{9}$$

In this study, the method was applied using the NIST Consensus Builder without changing any of the default options, except to obtain a report of the DoEs.

2.2.8. Random Effects with Hierarchical Bayes (RE-HB)

The second random-effects model method implemented in the NIST Consensus Builder is a hierarchical Bayesian procedure [15]. This method specifies probability distributions for all quantities in play and derives estimates of unknown parameters, with uncertainties, using posterior distributions calculated by applying Bayes' rule. As described in the NIST Consensus Builder documentation, the following prior distributions are used:

μ**:** Gaussian with zero mean and standard deviation of 10^5,

τ**:** half-Cauchy with median equal to the median of the absolute differences between measured values and their median,

σ_i**:** half-Cauchy with median equal to median of $\{u_i\}$.

The method was applied using the NIST Consensus Builder without changing any of the default options, except to obtain a report of the DoEs.

2.2.9. Linear Pool (LP)

The third method implemented in the NIST Consensus Builder, and the last method examined in this work, is the Linear Pool method [15]. The method does not use any of the models described in the previous section. It generates a mixture distribution from all submitted measurement results of the form

$$f = \sum w_{i,\mathrm{LP}}\, \phi_i \ , \tag{10}$$

where the ϕ_i are PDFs of Gaussian distributions for each participant, with mean x_i and standard deviation u_i. Note, the weights $w_{i,\mathrm{LP}}$ can be set by the user of the NIST Consensus Builder and are not those used in Equation (2). In this work, these weights were all set to unity. The consensus value obtained by this method is the mean of a sample drawn from this mixture distribution. A corresponding uncertainty interval can be built by selecting percentiles from the sample. The method was applied using the NIST Consensus Builder without changing any of the default options, except to obtain a report of the DoEs.

The NIST Consensus Builder has been discussed at many Consultative Committee meetings in recent years and a recommendation to use it has been included in the CCM guidance document among several options [20].

2.2.10. Leave One Out (LOO)

The evaluation of DoEs can be modified in several of the methods described above. For each participant, a reference value can be calculated without including the participant's result. Then, a DoE can be evaluated by taking the difference between this reference value and the participant's result. We use a suffix to indicate when this option has been used in the analysis, e.g., RE-MP-LOO, RE-HB-LOO, RE-DL-LOO, and LP-LOO.

3. Comparing Various Methods

The purpose of this work is to investigate the ability of various analysis methods to correctly identify biased and unbiased participants in a key comparison (in other words, to investigate the statistical power of various approaches to solving the comparison problem). In the analysis of a comparison, a participant i is deemed to have submitted an 'equivalent' result when

$$|DoE_i| < U(DoE_i) \ , \tag{11}$$

where $U(DoE_i)$ is an expanded uncertainty, obtained using a coverage factor $k = 1.96$ for most methods, except FE-BMA, for which 95% credible intervals are evaluated numerically, and the RE-DL, RE-HB, and LP methods, which used the NIST Consensus Builder to calculate expanded uncertainties directly (see [15] for details). Note that for each of the CM methods and RE-MP, correlations between measurements and the reference value were taken into account when calculating the uncertainty of the DoEs.

Our method of assessment implemented a numerical test-bed on which the long-run performance of various algorithms could be observed. Numerical simulation was used to generate many synthetic sets of comparison data with known properties. The performance of various comparison analysis methods was assessed by observing the relative frequencies of desirable and undesirable assessment outcomes when processing the same sets of data, i.e., the equivalence rates of the various methods. This allowed us to see how various methods behaved in a context representative of BIPM key comparisons in mature fields.

3.1. Testing

We simulated a measurement comparison with 12 laboratories each measuring a single artefact once. Our interest was in the analysis of key comparisons in mature technical fields carried out at the level of CIPM Consultative Committees; i.e., we are not considering pilot studies, regional linked comparisons, situations where no previous comparisons have been carried out, or fields where there are large inconsistencies between results, because comparisons under those conditions are unlikely to be able to support CMC claims. With those considerations in mind, the following conditions were chosen:

- The true value of the artefact is set to zero. There is no loss of generality with this condition.
- Each participant's measurement result has infinite degrees of freedom. This is not always true in practice, but for a key comparison, laboratories tend to put in more effort than usual to improve the confidence in their uncertainty budget, and this usually results in high effective degrees of freedom.
- Each participant's measurement is subject to an error drawn from a Gaussian distribution. The error variance was determined by random variates drawn from a normalised chi-squared distribution with four degrees of freedom. This distribution was chosen as a fair representation of the range of values typically observed in BIPM comparisons—most uncertainties are close to the mean (in this case unity), there are several 'good' laboratories with small uncertainties, and the occasional participant with very large uncertainties. Previous work showed that the simulations are not sensitive to the shape of this distribution [29].
- Only one or two laboratories' measurements are biased (subject to an unacknowledged error). This reflects the situation for key comparisons of mature scales, where the quantity has been compared before—probably by many of the same participants—and most of the sources of error are well known. The results of such comparisons may support new measurement claims.

The validity and implications of these conditions are explored in Section 5.

3.1.1. Input Data Sets

Data were generated by the method described below. We considered comparisons in which there were no biased participants and comparisons in which there were one and two biased participants.

With the number of participants fixed at $P = 12$, a large number (10,000 for all methods except FE-BMA, RE-DL, RE-HB, and LP, where only 1000 sets were processed due to the time required for the calculations) of comparison data sets were generated by a process that attributed a pair of numbers (y_i, u_i) to each participant.

- For a participant i with no bias, these numbers were generated in two steps:
  1. A value of u_i^2 was obtained from a chi-squared random number generator with $\nu = 4$ degrees of freedom, normalised to have a mean of unity;
  2. y_i was set to e_i, where e_i was obtained from a Gaussian random number generator with a mean of zero and variance u_i^2.
- For a biased participant, the generation of (y_i, u_i) was controlled by two parameters f and g, which allowed us to configure various scenarios.
  1. u_i was set to f;

2. y_i was set to $g + e_i$, where e_i was obtained from a Gaussian random number generator with a mean of zero and variance u_i^2.

The parameter f controlled the uncertainty of the biased participants. It was fixed at values between 0.25 and 2.0, corresponding to standard uncertainties between one-quarter and 2-fold the mean of the randomly generated unbiased participant uncertainties. A biased participant with a relatively small uncertainty ($f < 1$) will be more strongly weighted when determining the reference value using Equation (2), while one with a relatively large uncertainty ($f > 1$) will receive a lower weighting. In other words, a biased participant with $f < 1$ makes a 'better' capability claim than most other participants and will have relatively more influence on the reference value. In scenarios with a pair of biased participants, the biased participants were assigned the equal uncertainties.

The parameter g was set to values between 0 and 16 to determine the magnitude of the bias. In scenarios with a pair of biased participants, the biased participants were either both positively biased, in which case the same bias g was assigned to each, or oppositely biased, in which case g and $-g$ were assigned.

4. Results

The methods of comparison analysis described in Section 2.2 were applied to large numbers of synthetic comparison data sets. In Section 4.1, we report on the results obtained when there were no biased participants in the simulated comparisons. In Section 4.2, we investigate what happened when some participants were biased—scenarios with one and two biased participants were considered.

4.1. Unbiased Participants

Even when all participants in a comparison submit results that are free from unacknowledged systematic errors (biases), the measurements are still affected by other errors and the evaluation of equivalence in Equation (11) may incorrectly determine that a participant is biased. So, when the comparison data are bias free, it is of interest to look at the biased participant detection rates for various methods of analysis.

The expanded uncertainty in Equation (11) has a nominal 95% coverage probability, so we expect participants to be classified 'equivalent' on approximately 95% of occasions. In other words, we expect about 5% of unbiased participants to be incorrectly classified as biased.

Table 1 reports our observations of the rates that participants were classified as equivalent by the various methods of analysis. Some variability in the number of participants judged to be equivalent is expected because of the finite number of cases assessed. The standard deviation of observed equivalence rates has been calculated and is reported in brackets. Each judgement may be considered an independent Bernoulli trial. Then, the standard deviation of the equivalence rates p for n trials is $u_{\text{equiv}} = \sqrt{p(1-p)/n}$. So, when $n = 12 \times 10,000$ and $p = 0.95$, the standard deviation is 0.00063 and when $n = 12 \times 1000$ and $p = 0.95$ it is 0.0020.

Practising metrologists would usually expect an equivalence rate close to 0.95 for any method that claims to deliver 95% confidence or credible intervals. The results show the expected 95% success rates for some methods, but significantly higher rates for FE-BMA and the methods of the NIST Consensus Builder. A high success rate is clearly desirable; for unbiased participants, we want the rate of acceptance to be as high as possible and a value of 1 is an ideal result. To be fit for purpose, however, methods must also detect biased measurements and this is examined in the next section.

It must be remembered that the Bayesian definition of probability is not based on the relative frequency of events. So, the probability associated with credibility intervals cannot be automatically assumed to deliver 95% success rates with our testing framework. Nevertheless, it is indeed the relative frequencies of desirable and undesirable outcomes that are of interest when choosing an appropriate method of comparison analysis. Therefore, the observation of significant deviations from what might otherwise be expected is important.

Table 1. Rates that participants were classified as 'equivalent' by the various analysis methods when no biased participants were involved. Section 3.1.1 describes the method used to generate data sets. The standard deviation in the final digits of the repeatability of the numbers is also given in brackets. Methods are grouped by statistical model in the table, so they are not ordered by the observed equivalence rates.

Method	Equivalent
CM/FE-WM	0.94949 (63)
CM-LCS	0.95073 (62)
CM-CO	0.94970 (63)
CM-OO	0.94949 (63)
FE-BMA	0.99558 (61)
RE-MP	0.95111 (62)
RE-MP-LOO	0.95142 (62)
RE-DL	0.9709 (15)
RE-DL-LOO	0.9644 (17)
RE-HB	0.9818 (12)
RE-HB-LOO	0.9732 (15)
LP	0.99988 (10)
LP-LOO	0.99619 (56)

4.2. One or Two Biased Participants

A full set of results is given in Appendix B. In this section, we focus on a few specific examples and explain how to interpret the results.

We studied three scenarios where some of the participants in a comparison are biased:

1. One positively biased participant,
2. Two positively biased participants—with equal biases, and
3. Two biased participants—with biases of equal magnitude and opposite sign.

When there are biased participants, two desirable analysis outcomes are of interest: that a participant is (correctly) judged to be equivalent when the measurement is unbiased, and that a participant is (correctly) judged not to be equivalent when the measurement is biased.

Our results are presented in figures containing pairs of panes with a common abscissa scale. The detection rates of biased participants are shown in the upper pane and the detection rates of unbiased participants are shown in the lower pane. The abscissa scale is the relative bias (g/f) of participants.

Firstly, we consider the default method of comparison analysis, CM/FE-WM. Figure 1 shows the detection rates when only one biased participant is included in the comparison data set. Three traces show the results when the biased participant standard uncertainty (f) is 0.25, 1, and 2 (noting that the mean uncertainty of unbiased participants will be approximately 1).

There is certain behaviour that we expect to see in such a figure. Again, some variability in the numbers observed is expected, but this will be imperceptible on the plots. See Section 4.1 for details. Firstly, all methods should get better at detecting biased results when the magnitude of bias increases relative to the uncertainty. This effect is apparent in Figure 1 pane (i). Secondly, as the relative bias becomes small, the detection rates of unbiased participants should tend towards the results obtained in Section 4.1, which is seen in Figure 1 pane (ii).

The bias of a participant will influence the reference value calculation and hence the determinations of equivalence. For weighted mean methods, the relative importance given to each participant in the calculation is determined by the reported uncertainty, so when a biased participant's result has a relatively small uncertainty, they have a greater influence on the reference value. When the biased participant uncertainty (parameter f) is low, we expect the comparison analysis to be more sensitive to the amount of bias (parameter g) in the data. Figure 1 (ii) shows that an unbiased participant is much more likely to (unfairly)

'fail' the comparison as f decreases, or as g increases. For $f = 0.25$, however, a biased participant is less likely to be detected (correctly). In that case, the higher influence that a biased participant has on the reference value is balanced by the smaller uncertainty in the DoE.

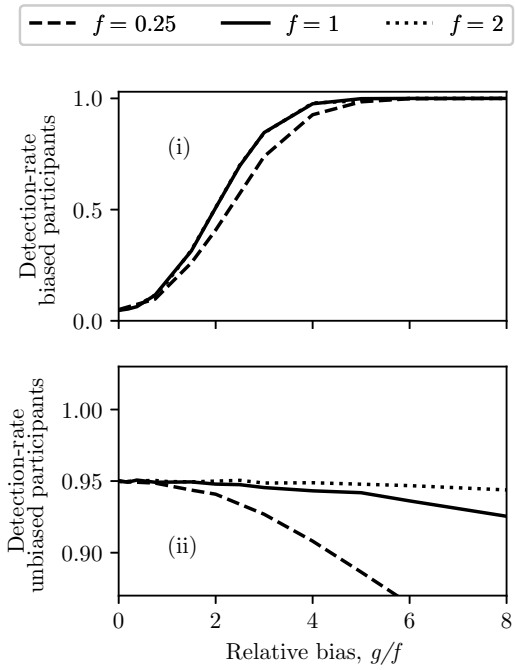

Figure 1. Results obtained with one positively biased participant in each comparison data set. Pane (**i**) shows detection rates of the biased laboratory for three different values of participant uncertainty (the trace for $f = 2$ is coincident with that for $f = 1$); pane (**ii**) shows detection rates of the unbiased laboratories. Data are shown for values of the biased participant uncertainty f of 0.25, 1, and 2. The mean uncertainty of the unbiased participants will be approximately unity. The abscissa scale is relative bias g/f, the bias g divided by the standard uncertainty of the biased participant f.

We now look at how the default approach recommended by most Consultative Committees (CM/FE-WM) compares with some of the other methods. In Figure 2, the CM/FE-WM method detection rates are compared with those of the other CM and FE model methods when biased participants have declared an uncertainty of unity ($f = 1$). Results for the three scenarios (1 biased participant, 2 positively biased participants, and 2 oppositely biased participants) are shown in columns across the figure. The detection rates for the biased participants are indistinguishable for all methods except FE-BMA, which is significantly worse under all of these test conditions. However, the FE-BMA method is correspondingly better at detecting unbiased participants.

When one or two positively biased participants are included in the comparison set, unbiased participants are increasingly penalised by the CM/FE-WM method as the bias increases. This effect is expected and is exacerbated by introducing a cut-off. The CM-CO method is intended to curtail the influence of biased participants that have small uncertainties. In fact, Figure 2 shows that this method only serves to disadvantage unbiased participants when biased participants have a moderate uncertainty; although a small positive effect relative to the CM/FE-WM method, when the biased participant has a small ($f = 0.25$) uncertainty, is seen in Appendix B. The two methods that exclude outliers, CM-OO and CM-LCS, in contrast, deliver significant improvements on the CM/FE-WM method. CM-LCS is remarkably stable in its protection of unbiased laboratories, with no significant loss of power in detecting biased participants. CM-OO 'kicks in' when the relative bias crosses the threshold of approximately 6, which can be understood from the criteria for outlier exclusion.

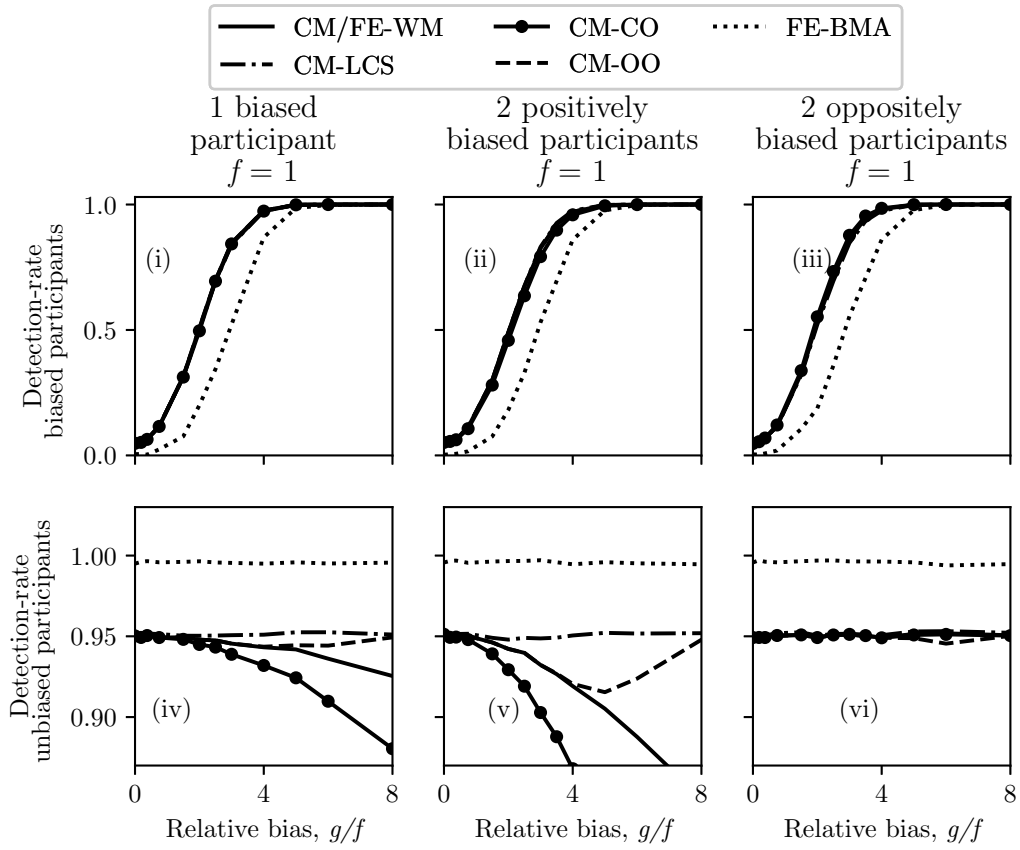

Figure 2. Results obtained with one or two positively biased participant(s) in each comparison data set. Panes (**i–iii**) show detection rates of the biased participant(s) (many of the lines overlap); panes (**iv–vi**) show detection rates of the unbiased laboratories. The biased participant(s) uncertainty is unity ($f = 1$) and the mean uncertainty of the unbiased participants will be approximately unity. The abscissa scale is the relative bias g/f of the biased participant(s) which, in this case, is equal to the bias.

In stark contrast, when equal and opposite biases were used in the generation of the data sets, the unbiased participant detection rates are quite insensitive to increasing relative bias, as seen in Figure 2 (vi). We attribute this to the fact that the combination of biases from both of the biased participants, with equal uncertainties, will have cancelled during the calculation of the reference value.

In Figure 3, we compare CM/FE-WM with the RE model methods and LP, which have been applied both with and without the LOO option. A set of results for two positively biased participants with $f = 1$ is shown in the figure. The detection rate for biased participants is strikingly low for all methods implemented in the NIST Consensus Builder and somewhat low for the RE-MP approach. As expected, the detection rates of unbiased participants is significantly higher than 95%. When LOO is applied, the power of all methods to detect biased participants increases. In the case of LP and RE-HB, the improvement is significant, however not to a level comparable with either RE-MP or CM/FE-WM. The detection rates for unbiased participants decreases when LOO is applied, but not to an untenable level, except for RE-MP, which loses power dramatically. Note that applying LOO to the CM/FE-WM approach does not change the detection rates at all (see the appendix to [30], where it is shown that the ratio of the degree of equivalence to its uncertainty is invariant for a participant whether or not they have been included in the calculation of the reference value for CM/FE-WM).

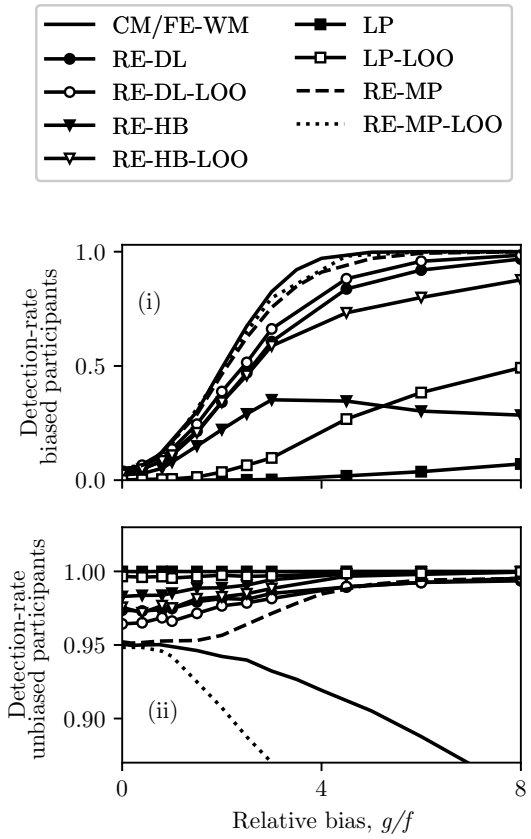

Figure 3. Results obtained with two positively biased participants in each comparison data set and the participant bias $f = 1$. Pane **(i)** shows detection rates of the biased laboratory; pane **(ii)** shows detection rates of the unbiased laboratories. The mean uncertainty of the unbiased participants will be approximately unity. The abscissa scale is equal to the relative bias g/f of the biased participant(s) in this case.

The main trends we observed in our simulations have been demonstrated by the selection of data presented in this section. However, many more detailed observations about the behaviour of the selected methods can be made from the full data set, which is presented in Appendix B.

5. Discussion

This study has looked at whether the performance of various methods of analysis may deviate from what might be generally expected, under conditions considered typical of mature fields where few discrepant results would be expected. That is, whether the detection rates of unbiased comparison participants are affected by including biased results, and whether the detection rates of biased participants are affected by decreasing bias. The results shown in Figures 1, 2 and A3 indicate that there are indeed some significant deviations. The inclusion of biased results has a detrimental effect on the determination of equivalence for unbiased participants. The detection of bias is also method dependent.

5.1. Method Assessment

With the objective of selecting the best method for comparison analysis, we should look for methods that are relatively insensitive to bias. A 'good' method will not be sensitive to changes in the relative bias or the number and polarity of biased results. Our results show that the concerns held by Consultative Committees about the use of the CM/FE-WM approach without any modification are well founded. Although it has good power to detect biased laboratories, unbiased laboratories are severely disadvantaged, especially when a

biased laboratory's result has a small uncertainty. Almost all of the method modifications proposed improve this situation; however, the common mean method with the largest consistent subset (CM-LCS) stands out among the CM and FE methods.

The CM-LCS method is remarkably stable, with a consistent 95% detection rate for unbiased laboratories and, in almost all cases and conditions, has as high a power to detect biased participants as any other approach. This finding would not have been anticipated from theoretical criticisms alone [6]. In all the scenarios considered, the method only under-performs other approaches when there are two oppositely biased participants. There is a small drop in the detection rates, for both biased and unbiased laboratories, when biased participants have small uncertainties and relatively moderate biases (between 0.5 and 1—see Appendix B). This occurs because the CM-LCS method may exclude one biased participant from a consistent subset but retain the other, which will then bias the reference value, lowering the detection rate for unbiased participants and increasing the likelihood of accepting the remaining biased participant.

LCS has been criticised by invoking a principle that no results should be excluded from the calculation of the reference value purely for statistical reasons, i.e., not scientific ones [31]. If the intent of a comparison is to establish a consensus value, of a fundamental constant for example, then this principle is sound. However, if a measurement comparison is used only to establish scale equivalence between participants, this principle is not so important. Comparison results are not used to shift the scale of a participant, rather they are used to assess measurement claims and ensure comparability of measurements made in different economies. That is, the comparison results (of the 'mature' type examined in this work) are used to evaluate precision rather than accuracy. The presence of outliers would certainly trigger an investigation into possible scientific reasons for the disagreement and, if none can be found, will cast doubt on the accuracy of the consensus scale, but without impacting the findings of equivalence between the participants.

Of the other CM and FE methods, CM-CO is marginally effective when a biased result has a low uncertainty, but otherwise has a lower power. The CM-OO method starts to become effective when f/g exceeds the chosen threshold, which can be tuned as necessary. The FE-BMA method was found to be remarkably insensitive to relative bias and the number/polarity of biased results. Unbiased participant detection rates for this method are consistently high: above 99% for all cases and bias conditions. This may be considered desirable. Unfortunately, the detection rates of biased participants by FE-BMA are consistently and significantly lower than the other CM and FE methods, except when the relative bias is large, where the performance of all approaches converges anyway. From a practical point of view, FE-BMA is unlikely to be more successful at detecting biased participants than the eye of a metrologist (although impartiality would not be a concern).

Of the random-effects models studied here, RE-MP appears to have greater statistical power than either of RE-DL or RE-HB. In order to achieve consistency for all data with the reference value (determined differently for each method), RE methods introduce a component of 'dark uncertainty'. In general, RE-MP retains a higher power to detect biased participants because the value of the 'dark uncertainty' τ required to satisfy the model is smaller. However, none of the RE model methods or the LP method outperforms CM-LCS, or even CM/FE-WM, for detection of biased measurements. In fact, most of these methods are significantly weaker. This is perhaps not surprising for the RE methods, because they are intended to provide a consensus value from a set of contributing measurements (such as determining the gravitational constant [18]), not identify differences between those measurements. They are likely to be more suited to pilot studies rather than comparison activity intended to support measurement claims. Among these, however, the statistical power of RE-MP approaches that of CM-LCS for the detection of biased laboratories and often exceeds it in the detection of unbiased laboratories.

The performance of random-effects models in this study highlights the possibility of two distinct purposes for comparison analysis. This work has focussed on using a comparison to determine equivalence. However, a measurement comparison might just

be used to estimate a reference value. Of course the former cannot be achieved without the latter and it can be argued that random-effects models deliver better estimates of the reference value because there is no exclusion of data and the introduction of 'dark uncertainty' achieves consistency with the reference value. On that basis it is no longer statistically meaningful to claim that a participant is biased. The results of this study—the fact that very few 'biased participants' were identified using RE methods—shows this to play out in practice and implies that CMC claims could become less reliable: claims based on an RE method of analysis would have to take account of the additional component of dark uncertainty in order to correctly reflect the evidence for equivalence. It is only by acknowledging this community-wide component of uncertainty, of unknown origin, that equivalence can be established or claimed. This appears to constitute a significantly different approach to the use of comparison results in support of CMC claims, but may be a very valid one. The effect of any outliers will be to reduce the precision of the scale universally, if comparability of measurements made in all participating economies is to be maintained. If an outlier is in fact contributing information about the true value of the measurand, then this approach may improve scale accuracy while sacrificing precision.

The power to detect biased participants tends to increase when the LOO option is used in conjunction with suitable methods. However, this option does mean that each participant's result is compared to a different reference value. This does not preclude the use of such DoEs in support of measurement claims (notwithstanding the requirements of the technical supplement to the MRA [1]). However, an important requirement of degrees of equivalence is that any subset of participants in a primary comparison should be able to provide an unambiguous link from that comparison to a regional measurement comparison. Some careful thought will be required if two linking participants' DoEs have been obtained using different reference values—the same issue may arise with the use of the FE-BMA approach.

5.2. Testing Applicability

Our approach to testing the statistical power of analysis methods, either currently in use or proposed for use by the metrology community, is generally applicable. All candidate methods are required to perform the same task—the analysis of comparison data—so there is no need to distinguish between the different statistical schools of thought and different notions of probability that underpin the methods. As this paper has shown, our approach can identify interesting method behaviours that could not have been anticipated theoretically. The approach is intended to complement rather than substitute for a mathematical analysis of the various methods. It allows the analyst to observe how methods perform under specific conditions of interest, which will help to validate the relevance of more theoretical descriptions and criticisms of various methods. This makes our approach extremely useful to groups, like the Consultative Committees, faced with the task of deciding which method is best suited to a particular type of problem.

While the particular findings of a simulation study may not be expected to hold under significantly different conditions, the approach is generic: data sets typical of conditions prevailing in other technical contexts, if they are well understood, can be produced with little difficulty. More elaborate comparison designs, in which there are multiple measurements per participant and several artefacts, may also be simulated. Furthermore, the relevance of information inferred from simulations may be reviewed once the results of a comparison analysis are available. If the results are compatible with assumptions made in generating simulated data sets, then the insights obtained into the statistical behaviour of the method are pertinent.

5.3. Testing Conditions

It must be understood that our results are particular to the conditions of the test. The low number of biased participants (one or two) means that the simulated comparison data sets are not well represented by random-effects models, which assume that all par-

ticipants' results are subject to an unacknowledged error. Two of the NIST Consensus Builder procedures and the Mandel–Paule method are based on the random-effects model. Nevertheless, a low number of biased participants was chosen for this work because that would seem to be a necessary assumption if the results of a comparison are intended to support measurement capability claims.

If this assumption does not hold, then the determination of a reference value as the weighted mean of results is unreliable and methods of resolving this tend to undermine the purpose of the comparison. For example, when a random-effects model is invoked to account for the variability between participants, the addition of a component of uncertainty to the measurements, referred to as 'dark uncertainty', helps to obtain a meaningful estimate of the artefact/reference value. However, this process inevitably weakens the power of the equivalence hypothesis test—the additional dark uncertainty component will tend to obscure the very non-equivalence that we would like to detect.

Confidence in an assumption of a low number of biased participants can be built on the context of the comparison. If the area is mature, if comparisons have been run previously for the quantity, if uncertainty budgets have been published and accepted by the community, if participants in primary comparisons (i.e., not linked) are chosen largely from a pool of experienced participants, if artefacts are well characterised and stable, then the justification of such an assumption can be made. This list is similar to that of the CCQM guidance document identifying low probability of inconsistent results [22], which even recommends that no reference value or DoEs be calculated or published if there is evidence of 'severe inconsistency' in the comparison results.

If the conditions described for a mature field cannot be met, then it is more likely that a measurement comparison takes the form of a pilot study. The purpose of a pilot study is not to support measurement claims but to, for example, identify whether significant sources of error are being underestimated, to test the suitability of the chosen artefacts, or to compare various methods of realisation of a quantity. A large degree of inconsistency may arise in an otherwise mature field, if measurement uncertainties have been substantially reduced compared to a previous comparison. In these cases, a random-effects model may be more suitable. However, the quantity of interest is now probably the dark uncertainty, τ, which becomes a measure of the readiness of the field to accept substantially reduced uncertainties.

Further assumptions made in this work are that the degrees of freedom are infinite and that the uncertainty budgets reported by unbiased participants are correct. In our experience, for a mature field, the degrees of freedom in comparisons are large, as participants make extra effort to build confidence in their uncertainty budgets. As to the latter assumption, we suspect from experience that participants often report conservative uncertainty budgets. Such cautious behaviour will reduce the statistical power of a method. This might be an interesting topic to pursue in a further study.

6. Conclusions

Trust in the performance of comparison analyses is essential to the proper functioning of the MRA. We have shown that the detection rates of desirable and undesirable outcomes of various analyses vary in ways that are difficult to anticipate. So, this study shows that numerical testing of different methods is a valuable complement to other sources of information about methods when choosing an appropriate method for comparison analysis. This should be considered when developing policy for comparison analysis.

Our testing approach complements theoretical and conceptual analysis, taking into consideration the expectations of a particular metrological community for the performance of comparison analysis tools. It is intended to provide an independent way of building confidence in the safety of CMC decisions made by Consultative Committees assuming the simple evaluation of CMC claims against DoEs described in this paper.

Under conditions expected in a mature technical field, where most participants have made accurate estimates of their measurement uncertainty, no additional information is

available, and the test of the CMC claim against the DoE is as described in Equation (11), the CM-LCS method stands out as having the highest statistical power to detect biased participants and to consistently identify unbiased participants at a 95% level. That this result is perhaps surprising, given the previous critique of CM-LCS, serves to emphasise the value of this type of assessment.

However, random-effects models may deliver a better estimate of the reference value. However, the CMC claim process must be amended to account for the 'dark uncertainty' introduced by this process. A globally consistent, low-risk database of measurement capabilities can still be maintained at the cost of larger uncertainties and uncertainty budgets for primary standards that include a contribution from an unknown source.

The statistical power of new comparison analysis methods proposed in the literature, or existing methods used under differing comparison scenarios, can be assessed in the same way. For a given scenario, the selection of an analysis method can be informed by observing its statistical behaviour. Methods with higher statistical power can be identified and a robust understanding of the statistical nature of comparison analysis, i.e., the probabilities of desirable and undesirable outcomes, can be obtained. This will increase confidence in measurement claims made under the CIPM MRA and better inform decision making based on comparison results.

Author Contributions: Conceptualization, A.K.; methodology, E.M., A.K. and B.D.H.; software, E.M. and R.H.; validation, E.M., R.H. and B.D.H.; investigation, E.M. and R.H.; formal analysis, E.M. and A.K.; writing—original draft preparation, E.M.; writing—review and editing, A.K. and B.D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the New Zealand government.

Data Availability Statement: All of the data is included in the figures in the appendix. The code to calculate DoEs can be found at https://github.com/MSLNZ/Comparison-analysis (accessed on 22 August 2021).

Acknowledgments: The authors are grateful to the NIST Consensus Builder team, in particular Antonio Possolo and Thomas Lafarge who made their code available to us and provided help and advice with troubleshooting. We also thank Rob Willink, Rod White, Lutz Werner, Clemens Elster, and Gerd Wübbeler for helpful discussions of our work and Peter Saunders for advice on numerical integration.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

BIPM	International Bureau of Weights and Measures
CM	Common Mean
CM-LCS	Common Mean with Largest Consistent Subset
CM-CO	Common Mean with Cut-Off
CM-OO	Common Mean with exclusion of Obvious Outliers
CCAUV	Consultative Committee for Acoustics, Ultrasound, and Vibration
CCEM	Consultative Committee for Electricity and Magnetism
CCQM	Consultative Committee for Amount of Substance
CCL	Consultative Committee for Length
CCM	Consultative Committee for Mass and Related Quantities
CCPR	Consultative Committee for Photometry and Radiometry
CIPM	International Committee for Weights and Measures
CMC	Calibration and Measurement Capability
DoE	Degree of Equivalence
FE	Fixed Effects

FE-BMA	Fixed Effects with Bayesian Model Averaging
FE-WM	Fixed Effects with Weighted Mean
LOO	Leave One Out
LP	Linear Pool
MRA	CIPM Mutual Recognition Arrangement
NMI	National Metrology Institute
RE-DL	Random Effects with DerSimonian and Laird
RE-HB	Random Effects with Hierarchical Bayes
RE-MP	Random Effects with Mandel–Paule

Appendix A. Order Parameter Selection and Credible Interval Calculations for FE-BMA

Appendix A.1. Order Parameter Selection

The method of Bayesian model averaging requires an order parameter m to be selected, which is the size of the subset to be averaged for each model. Elster and Toman [7] recommend that m be less than the size of the largest coherent subset. For this work, with one or two biased participants in each set of 12, the size of the largest coherent subset ($LCHS$) was almost always between 9 and 11.

In order to examine the dependence of the simulation results on m, we analysed synthetic comparison data in which there was one biased participant for all values of m. The results are shown in Table A1. Two settings of the parameters for uncertainty and bias were chosen: $f = 0.25, g = 1$ and $f = 1, g = 3$. The data sets were produced using the simulation methods described in Section 3.1.1 and each reported rate was obtained from the analysis of 1000 comparison data sets.

Table A1. FE-BMA detection rates for biased and unbiased participants obtained using different values of the order parameter m and two choices of the uncertainty and bias parameters f and g. Reported values were obtained from 1000 synthetic comparison data sets with one biased participant.

	$f = 0.25, g = 1$		$f = 1, g = 3$	
m	**Biased**	**Unbiased**	**Biased**	**Unbiased**
1	0.000	1.000	0.000	1.000
2	0.001	0.993	0.439	0.995
3	0.121	0.990	0.611	0.992
4	0.251	0.990	0.623	0.992
5	0.348	0.991	0.604	0.993
6	0.408	0.994	0.579	0.995
7	0.438	0.996	0.540	0.996
8	0.442	0.997	0.498	0.997
9	0.430	0.998	0.439	0.999
10	0.389	0.999	0.371	0.999
11	0.320	1.000	0.271	1.000

In the first case, with small uncertainty and moderate bias ($f = 0.25$, $g = 1$), the FE-BMA approach was able to best detect biased measurements when $m = 8$, and it appeared stable for m between 6 and 10. In the second case ($f = 1$, $g = 3$), FE-BMA performed best when $m = 4$ and seemed fairly stable between $m = 3$ and $m = 7$.

Calculations were also carried out using synthetic comparison data with two positively biased participants. In this case, we considered order parameter values $m = 5$, $m = LCHS - 4$ and $m = LCHS - 1$, where $LCHS$ is the size of the largest coherent subset. The results are shown in Figure A1(i), (ii), and (iii). A smaller value of m appears to be best when the uncertainty of the biased participants is large, but a larger value of m produces better detection rates when uncertainty is smaller. These trends are similar to those for comparisons with a single biased participant.

Based on the preceding observations, all the FE-BMA results presented in the body of this paper used $m = LCHS - 4$.

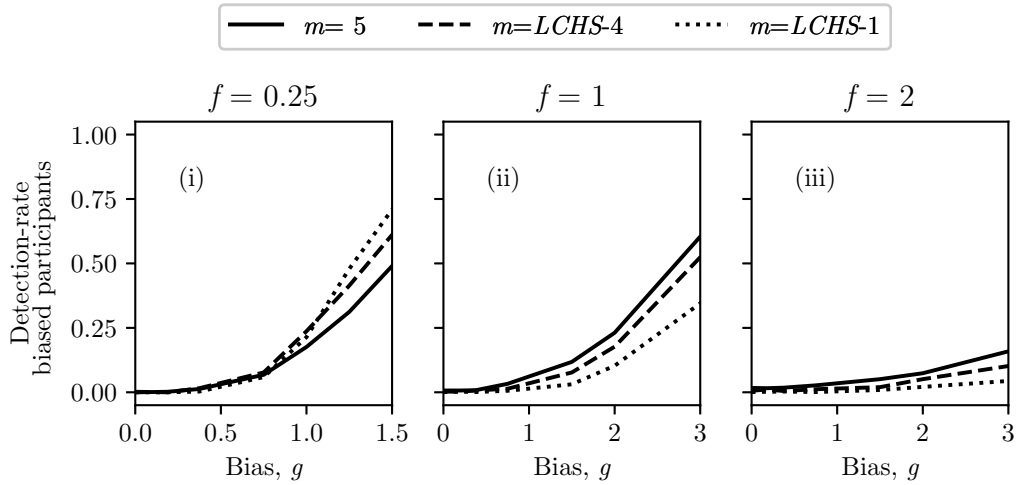

Figure A1. FE-BMA results for various values of order parameter m. Simulations used two positively biased participants with the same uncertainty and bias in each data set. Panes (**i–iii**) show the detection rates for biased laboratories, for the uncertainty parameter values $f = 0.25, 1$ and 2, respectively. Note, $LCHS$ is the size of the largest coherent subset.

Appendix A.2. Credible Interval Calculations

The results presented in this paper were obtained by calculating symmetric 95% credible intervals from posterior distributions for δ_i obtained using FE-BMA. However, these are mixture distributions, with a delta function at $\delta_i = 0$ and a continuous Gaussian component. In some cases, an exact 95% credible interval could not be found, in which case the smallest symmetric credible interval with probability greater than 95% was used. Overall, the average probability of the intervals obtained was always less than 96% across the sets tested.

Appendix B. Full Simulation Results

Several of the methods examined in this work have been used by comparison analysts in the past or have been recommended to particular Consultative Committees. In order to make decisions about change, a more detailed look at the full set of results obtained from our simulations may be desired. The two figures in this appendix show the complete set of simulation results under all of the conditions tested. More detail of the implementation of the methods and the results can be found in [29,30].

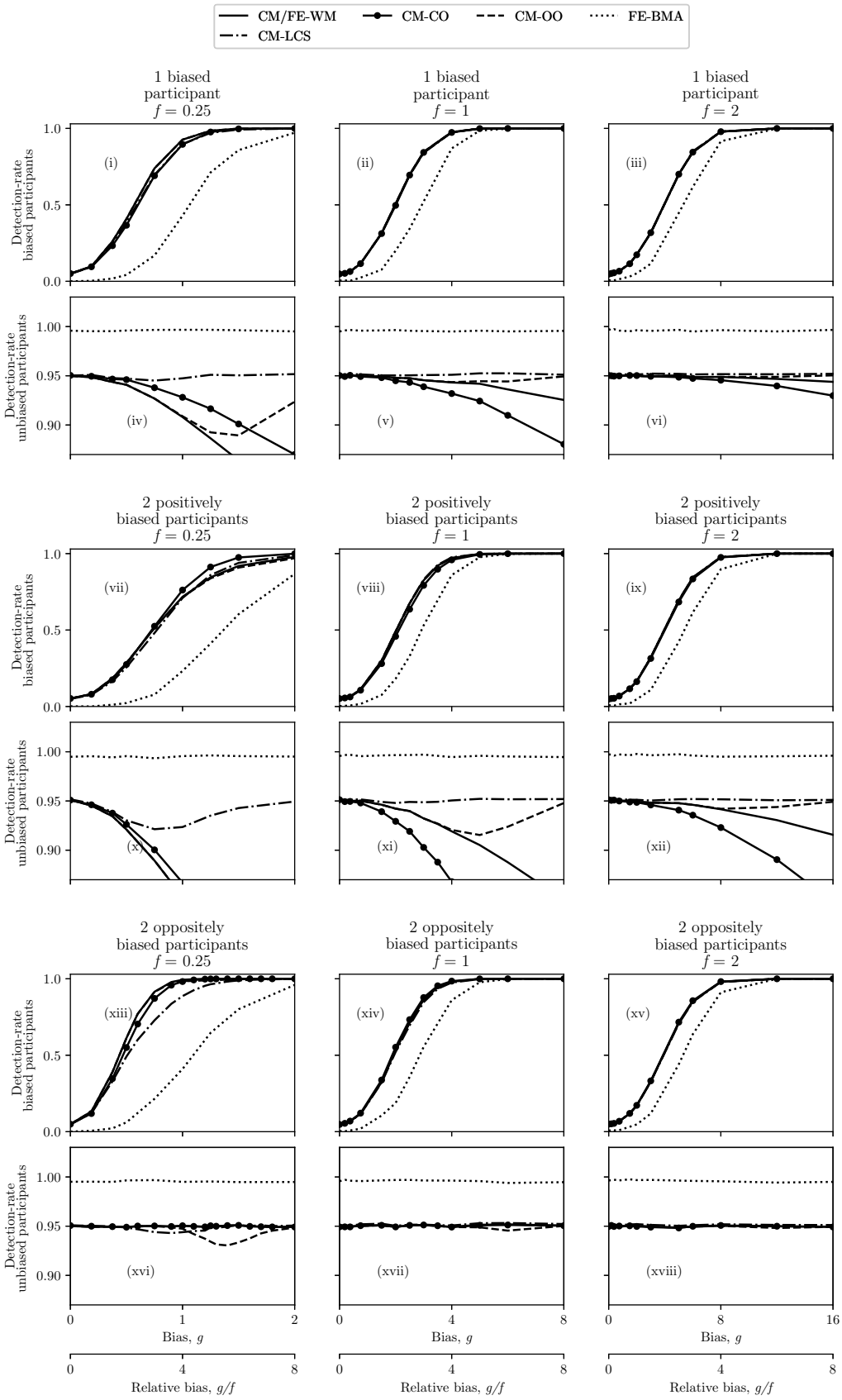

Figure A2. Results obtained with for all test conditions for the CM and FE approaches. Data are shown for each scenario, and for three values of the biased participant uncertainty f. The mean uncertainty of the unbiased participants will be approximately unity.

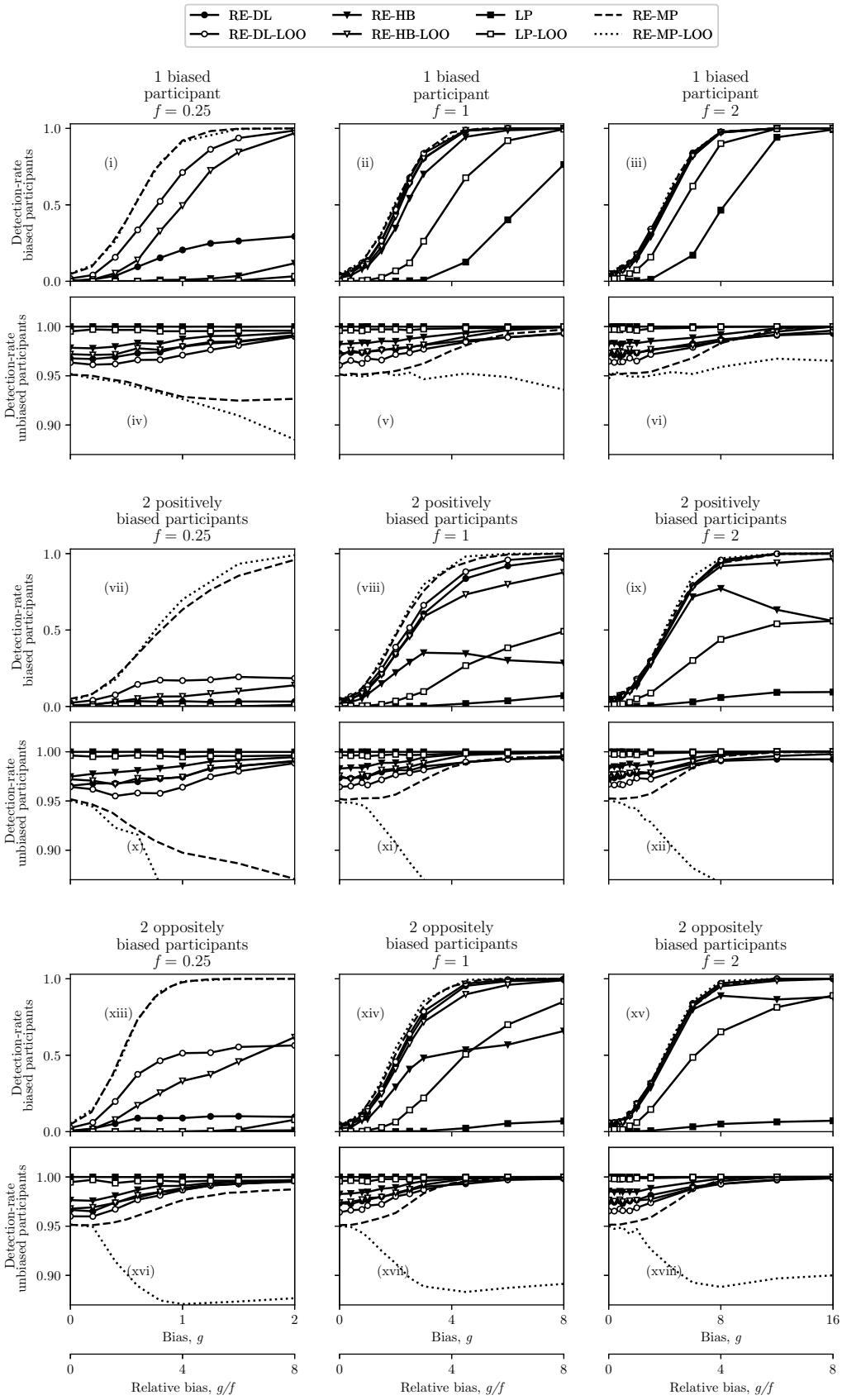

Figure A3. Results obtained with for all test conditions for the RE and LP approaches. Data are shown for each scenario, and for three values of the biased participant uncertainty f. The mean uncertainty of the unbiased participants will be approximately unity.

References

1. CIPM. Technical supplement revised in October 2003. In *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes*; BIPM: Sèvres, France, 2003; pp. 38–41.
2. Willink, R. On the interpretation and analysis of a degree-of-equivalence. *Metrologia* **2003**, *40*, 9–17. [CrossRef]
3. Wübbeler, G.; Bodnar, O.; Elster, C. Bayesian hypothesis testing for key comparisons. *Metrologia* **2016**, *53*, 1131–1138. [CrossRef]
4. Cox, M.G. The evaluation of key comparison data. *Metrologia* **2002**, *39*, 589–595. [CrossRef]
5. Cox, M.G. The evaluation of key comparison data: Determining the largest consistent subset. *Metrologia* **2007**, *44*, 187–200. [CrossRef]
6. Toman, B.; Possolo, A. Laboratory effects models for interlaboratory comparisons. *Accredit. Qual. Assur.* **2009**, *14*, 553–563. [CrossRef]
7. Elster, C.; Toman, B. Analysis of key comparisons: Estimating laboratories' biases by a fixed effects model using Bayesian model averaging. *Metrologia* **2010**, *47*, 113–119. [CrossRef]
8. Elster, C.; Toman, B. Analysis of key comparison data: Critical assessment of elements of current practice with suggested improvements. *Metrologia* **2013**, *50*, 549–555. [CrossRef]
9. Lira, I. Bayesian evaluation of comparison data. *Metrologia* **2006**, *43*, 3–7. [CrossRef]
10. Ballico, M. Calculation of key comparison reference values in the presence of non-zero-mean uncertainty distributions, using the maximum-likelihood technique. *Metrologia* **2001**, *38*, 155–159. [CrossRef]
11. Willink, R. Meaning and models in key comparisons, with measures of operability and interoperability. *Metrologia* **2006**, *43*, S220. [CrossRef]
12. Wübbeler, G.; Bodnar, O.; Mickan, B.; Elster, C. Explanatory power of degrees of equivalence in the presence of a random instability of the common measurand. *Metrologia* **2015**, *52*, 400–406. [CrossRef]
13. Koo, A.; Clare, J. On the equivalence of generalized least-squares approaches to the evaluation of measurement comparisons. *Metrologia* **2012**, *49*, 340. [CrossRef]
14. White, D.R. On the analysis of measurement comparisons. *Metrologia* **2004**, *41*, 122–131. [CrossRef]
15. Koepke, A.; Lafarge, T.; Possolo, A.; Toman, B. NIST Consensus Builder—User's Manual. 2017. Available online: https://consensus.nist.gov/app_{}direct/nicob/NISTConsensusBuilder-UserManual.pdf (accessed 22 August 2021).
16. Paule, R.C.; Mandel, J.; Bureau, N. Consensus Values and Weighting Factors. *J. Res. Natl. Bur. Stand.* **1982**, *87*, 377–385. [CrossRef]
17. CIPM Consultative Committee for Photometry and Radiometry. CCPR-G2 Guidelines for CCPR Key Comparison Report Preparation. 2019. Available online: https://www.bipm.org (accessed 22 August 2021).
18. Koepke, A.; Lafarge, T.; Possolo, A.; Toman, B. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia* **2017**, *54*, S34–S62. [CrossRef]
19. Koo, A. Report on the consultative committee for photometry and radiometry key comparison of regular spectral transmittance 2010 (CCPR-K6.2010). *Metrologia* **2017**, *54*, 02001. [CrossRef]
20. CIPM Consultative Committee for Mass and Related Quantities. Key Comparison Report Template v 1.3. 2019. Available online: https://www.bipm.org (accessed 22 August 2021).
21. CIPM Consultative Committee for Length. CCL-GD-3.2: Key Comparison Report Template. Available online: https://www.bipm.org (accessed 22 August 2021).
22. CIPM Consultative Committee for Amount of Substance. CCQM Guidance Note: Estimation of a Consensus KCRV and Associated Degrees of Equivalence. 2013. Available online: https://www.bipm.org (accessed 22 August 2021).
23. CIPM Consultative Committee for Electricity and Magnetism. CCEM Guidelines for Planning, Organizing, Conducting and Reporting Key, Supplementary and Pilot Comparisons. 2017. Available online: https://www.bipm.org (accessed 22 August 2021).
24. CIPM Consultative Committee for Acoustics Ultrasound and Vibration. Guidance for Carrying out Key Comparisons within the CCAUV. 2015. Available online: https://www.bipm.org (accessed 22 August 2021).
25. CIPM Consultative Committee for Ionizing Radiation. Key Decisions of the CCRI and CCRI Sections. 2019. Available online: https://www.bipm.org (accessed 22 August 2021).
26. Werner, L. Final report on the key comparison CCPR-K2.c-2003: Spectral responsivity in the range of 200 nm to 400 nm. *Metrologia* **2014**, *51*, 02002. [CrossRef]
27. DerSimonian, R.; Laird, N. Meta-analysis in clinical trials revisited. *Contemp. Clin. Trials* **2015**, *45*, 139–145. [CrossRef] [PubMed]
28. Whitehead, A.; Whitehead, J. A general parametric approach to the meta-analysis of randomized clinical trials. *Stat. Med.* **1991**, *10*, 1665–1677. [CrossRef] [PubMed]
29. Molloy, E.; Koo, A. *Methods and Software for Analysing Measurement Comparisons*; Technical Report 0689; Callaghan Innovation: Lower Hutt, New Zealand , 2020. [CrossRef]
30. Koo, A.; Harding, R.; Molloy, E. *A Statistical Power Study of the NIST Consensus Builder Models to Identify Participant Bias in Comparisons*; Technical Report 0805; Callaghan Innovation: Lower Hutt, New Zealand, 2020. [CrossRef]
31. Possolo, A.; Koepke, A.; Newton, D.; Winchester, M.R. Decision tree for key comparisons. *J. Res. Natl. Inst. Stand. Technol.* **2021**, *126*, 1–36. [CrossRef]