Research

Author for correspondence:
R. Kenna
e-mail: r.kenna@coventry.ac.uk

In memory of Prof. Sir David John Cameron MacKay (22 April 1967–14 April 2016).

THE ROYAL SOCIETY PUBLISHING

Calibration with confidence: a principled method for panel assessment

R. S. MacKay[1], R. Kenna[2], R. J. Low[2] and S. Parker[1]

[1]Mathematics Institute and Centre for Complexity Science, University of Warwick, Coventry CV4 7AL, UK

[2]Applied Mathematics Research Centre, Coventry University, Coventry CV1 5FB, UK

RK, 0000-0001-9990-4277

Frequently, a set of objects has to be evaluated by a panel of assessors, but not every object is assessed by every assessor. A problem facing such panels is how to take into account different standards among panel members and varying levels of confidence in their scores. Here, a mathematically based algorithm is developed to calibrate the scores of such assessors, addressing both of these issues. The algorithm is based on the connectivity of the graph of assessors and objects evaluated, incorporating declared confidences as weights on its edges. If the graph is sufficiently well connected, relative standards can be inferred by comparing how assessors rate objects they assess in common, weighted by the levels of confidence of each assessment. By removing these biases, 'true' values are inferred for all the objects. Reliability estimates for the resulting values are obtained. The algorithm is tested in two case studies: one by computer simulation and another based on realistic evaluation data. The process is compared to the simple averaging procedure in widespread use, and to Fisher's additive incomplete block analysis. It is anticipated that the algorithm will prove useful in a wide variety of situations such as evaluation of the quality of research submitted to national assessment exercises; appraisal of grant proposals submitted to funding panels; ranking of job applicants; and judgement of performances on degree courses wherein candidates can choose from lists of options.

1. Introduction

This paper addresses the widespread problem of how to take into account differences in standards, confidence and bias in assessment panels, such as those evaluating research quality or grant proposals, employment or promotion applications, and classification of university degree courses, in situations where it is not feasible for every assessor to evaluate every object to be assessed.

Table 1. Panel Assessment Methods. The matrix of four approaches according to use of calibration and/or confidences. Simple averaging (SA) is the base for comparisons. Fisher's IBA does not deal with varying degrees of confidence and the confidence-weighted averaging does not achieve calibration. The method proposed herein (CWC) accommodates both calibration and confidences.

	without confidences	with confidences
without calibration	simple averaging (SA)	confidence-weighted averaging (CWA)
with calibration	incomplete block analysis (IBA)	calibration with confidence (CWC)

A common approach to assessment of a range of objects by such a panel is to assign to each object the average of the scores awarded by the assessors who evaluate that object. This approach is represented by the cell labelled 'simple averaging' (SA) in the top left of a matrix of approaches listed in table 1, but it ignores the likely possibility that different assessors have different levels of stringency, expertise and bias [1]. Some panels shift the scores for each assessor to make the average of each take a normalized value, but this ignores the possibility that the set of objects assigned to one assessor may be of a genuinely different standard from that assigned to another. For an experimental scientist, the issue is obvious: *calibration*.

One solution is to seek to calibrate the assessors beforehand on a common subset of objects, perhaps disjoint from the set to be evaluated [2]. This means that they each evaluate all the objects in the subset and then some rescaling is agreed to bring the assessors into line as far as possible. This would not work well, however, in a situation where the range of objects is broader than the expertise of a single assessor. Also, regardless of how well the assessors are trained, differences between individuals' assessments of objects remain in such ad hoc approaches [3].

If the expertise of two assessors overlap on some subject, however, any discrepancy between their evaluations can be used to infer information about their relative standards. Thus, if the graph Γ_A on the set of assessors, formed by linking two whenever they assess a common object, is sufficiently well connected, one can expect to be able to infer a robust calibration of the assessors and hence robust scores for the objects. The construction of this graph is illustrated in figure 1, beginning from the graph Γ showing which objects are assessed by which assessors.

One approach to achieving such calibration was developed by Fisher [4], in the context of trials of crop treatments. Denoting the score from assessor a for object o by s_{ao}, Fisher's approach is based on fitting a model of the form $s_{ao} = v_o + b_a + \varepsilon_{ao}$ with ε_{ao} independent identically distributed random variables of mean zero. Then b_a is the bias inferred for assessor a and v_o is the value inferred for object o. Fisher's approach is known as *additive incomplete block analysis* (IBA) and a body of associated literature and applications has since been developed (e.g. [5]), though its use in panel assessment seems rare. It is represented as the bottom left entry of table 1.

Another ingredient that is important in many panel assessments, however, is different weights that may be put on different assessments. We refer to these weights as 'confidences'. Fisher's IBA does not take different levels of confidence into account.

If the assessors express confidences in the assessments, for example, by some pre-determined weights assigned to types of assessment or by the assessors declaring confidences in each of their scores, then it is natural to replace SA by confidence-weighted averaging. This is represented as the top right entry of table 1, but it does not address the calibration issue so we do not consider it further.

In this paper, we present and test a method to calibrate scores taking into account confidences; that is, we complete the matrix of approaches represented in table 1, where our method is termed *calibration with confidence* (CWC). We demonstrate that the method can achieve a greater degree of accuracy with fewer assessors than either SA or IBA, and we derive robustness estimates taking the confidences into account.

We are aware of two other schemes that incorporate confidences into a calibration process. One is the abstract-review method for the SIGKDD'09 conference (section 4 of [6]; see also [7]). The other is the abstract-review method used for the NIPS2013 conference (building on [8] and described in [9]). Our method has the advantages of simplicity of implementation and a straightforward robustness analysis. We leave detailed comparison with methods such as these for future publication.

2. The model

Let us suppose that each assessor is assigned a subset of the objects to evaluate. Denote the resulting set of (assessor, object) pairs by E. Let us further suppose that the score s_{ao} that assessor a assigns to object o
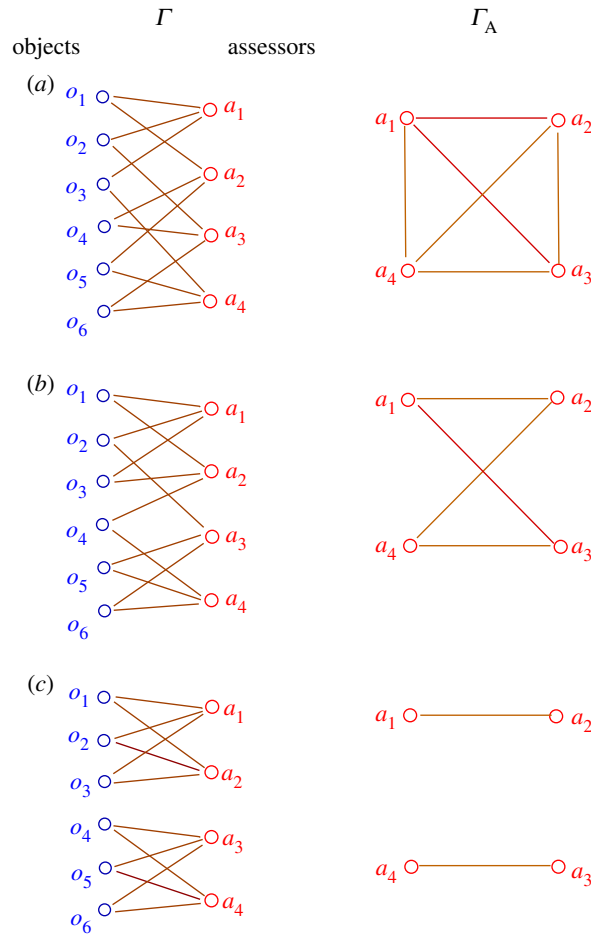
Figure 1. Three examples of assessment graphs Γ showing which object o_j is assessed by which assessor a_k, and the resulting graphs Γ_A on the set of assessors where two assessors are linked if they assess an object in common. Case (a) produces a fully connected assessor graph, (b) a moderately connected graph, whereas case (c) is disconnected.

is a real number related to a 'true' value v_o for the object by

$$s_{ao} = v_o + b_a + \varepsilon_{ao}, \tag{2.1}$$

where b_a can be called the *bias* of assessor a and ε_{ao} are independent zero-mean random variables. Such a model forms the basis for additive IBA. This was also proposed in ref. [10] (see eqn (8.2b) therein) but without a method to estimate the true values. Here we will achieve this and make a significant improvement, namely the incorporation of varying confidences in the scores.

To take into account the varying expertise of the assessors with respect to the objects, we propose that in addition to the score s_{ao}, each assessor is asked to specify a level of confidence for that evaluation. This could be in the form of a rating such as 'high', 'medium', 'low', as requested by some funding agencies, but we propose to allow something more general and akin to experimental science. Confidence can be estimated by asking assessors to specify an uncertainty $\sigma_{ao} > 0$ for their score and then the confidence level (or 'precision') is taken to be

$$c_{ao} = \frac{1}{\sigma_{ao}^2}. \tag{2.2}$$

The instructions to the assessors can be to choose s_{ao} and σ_{ao} so that $\frac{2}{3}$ of their probability distribution for the score lies in $[s_{ao} - \sigma_{ao}, s_{ao} + \sigma_{ao}]$, $\frac{1}{6}$ above this interval and $\frac{1}{6}$ below it. Methods for training assessors to estimate uncertainties are presented in [11]. There are also methods for training assessors on the assessment criteria to improve their accuracy [12], which could also be expected to have the beneficial effect of reducing their uncertainties.

So let us suppose that

$$\varepsilon_{ao} = \sigma_{ao}\eta_{ao}, \tag{2.3}$$

with η_{ao} independent zero-mean, random variables of common variance w. For the moment, we set $w = 1$; extensions to other values of w are considered in appendix A, and in particular are necessary if confidence is expressed only qualitatively. In the case that confidences are reported as only high, medium or low, they can be converted into quantitative ones by for example choosing $\lambda \approx 2$ and setting $c_{ao} = \lambda^2, 1, \lambda^{-2}$, respectively. The interpretation of λ is the ratio of the uncertainty for a low confidence evaluation to that for a medium one, and for a medium one to a high one. Then w is unspecified but can be fit from the data, as in appendix A.

Thus, our basic model is

$$s_{ao} = v_o + b_a + \sigma_{ao}\eta_{ao}. \tag{2.4}$$

3. Solution of the model

Given the data $\{(s_{ao}, \sigma_{ao}) : (a, o) \in E\}$ for all assigned assessor–object pairs, we wish to extract the true values v_o and assessor biases b_a. The simplest procedure is to minimize the sum of squares

$$\sum_{(a,o)\in E} \eta_{ao}^2 = \sum_{(a,o)\in E} c_{ao}(s_{ao} - v_o - b_a)^2, \tag{3.1}$$

where the confidence level c_{ao} was defined in equation (2.2). This procedure can be justified if the η_{ao} are assumed to be normally distributed, because then it gives the maximum-likelihood values for v_o and b_a. It can also be viewed as orthogonal projection of the vector s of scores s_{ao} to the subspace of the form $s_{ao} = v_o + b_a$ in the Riemannian metric given by $|s| = \sqrt{\sum_{ao} c_{ao} s_{ao}^2}$.

Now expression (3.1) is minimized with respect to v_o iff

$$\sum_{a:(a,o)\in E} c_{ao}(s_{ao} - v_o - b_a) = 0,$$

and with respect to b_a iff

$$\sum_{o:(a,o)\in E} c_{ao}(s_{ao} - v_o - b_a) = 0.$$

It is notationally convenient to extend the sums to all assessors (respectively objects) by assigning the value $c_{ao} = 0$ to any assessor–object pair that is not in E (i.e. for which a score was not returned). Then these conditions can be written as

$$C_o v_o + \sum_a b_a c_{ao} = V_o \tag{3.2}$$

and

$$\sum_o c_{ao} v_o + C_a' b_a = B_a. \tag{3.3}$$

Here,

$$V_o = \sum_a c_{ao} s_{ao} \tag{3.4}$$

is the confidence-weighted total score for object o and

$$B_a = \sum_o c_{ao} s_{ao} \tag{3.5}$$

is that for assessor a,

$$C_o = \sum_a c_{ao} \tag{3.6}$$

is the total confidence in the assessment of object o and

$$C_a' = \sum_o c_{ao} \tag{3.7}$$

is the total confidence expressed by assessor a.

Equations (3.2) and (3.3) form a linear system of equations for the v_o and b_a. It has an obvious degeneracy in that one could add a constant k to all the v_o and subtract k from all the b_a and obtain another solution. One can remove this degeneracy by, for example, imposing the condition

$$\sum_a b_a = 0. \tag{3.8}$$

This is the simplest possibility and corresponds to a translation (shift) that brings the average bias over assessors to zero. Alternatives are discussed in appendix B.

Define a graph Γ linking assessor a to object o if and only if $(a, o) \in E$, as illustrated in the left column of figure 1. The edges in the graph are weighted by the confidences c_{ao}. Whether the set of equations (3.2) and (3.3) has a unique solution after breaking the degeneracy depends on the connectivity of Γ. Define a linear operator L by writing equations (3.2) and (3.3) as

$$L \begin{bmatrix} v \\ b \end{bmatrix} = \begin{bmatrix} V \\ B \end{bmatrix}, \tag{3.9}$$

where v, b, V and B denote the column vectors formed by the v_o, b_a, V_o and B_a, respectively. The operator L has null space of dimension equal to the number of connected components of Γ (this follows from Perron–Frobenius theory, e.g. [13]). Thus, if Γ is connected, the null space of L has dimension one, so corresponds precisely to the null vectors $v_o = k \ \forall o, b_a = -k \ \forall a$, that we already noticed and dealt with. Connectedness of Γ ensures that if (3.9) has a solution then there is a unique one satisfying (3.8).

It remains to check that the right-hand side of equation (3.9) lies in the range of L, thus ensuring that a solution exists. This is true if all null forms of the adjoint operator $L^\dagger$ send the right-hand side to zero. The null space of $L^\dagger$ has the same dimension as that of L, because L is square, and an obvious non-zero null form α is given by

$$\alpha(v, b) = \sum_o v_o - \sum_a b_a. \tag{3.10}$$

It follows from the definitions of V and B that $\alpha(V, B) = 0$. So a solution exists.

Thus, under the assumption that the assessor–object graph Γ is connected, equations (3.2) and (3.3) have a unique solution (v, b) satisfying equation (3.8). Note that connectedness of Γ is necessary for uniqueness, otherwise one could follow an analogous procedure, adding and subtracting constants independently in each connected component of Γ, and thereby produce more solutions.

Equations (3.2) and (3.3) have a special structure, due to the bipartite nature of Γ, that can be worth exploiting. The first equation (3.2) can be written as

$$v_o = \frac{V_o - \sum_a b_a c_{ao}}{C_o}. \tag{3.11}$$

This can be substituted into the second equation (3.3) to obtain

$$\sum_{a'} C_{aa'} b_{a'} - C'_a b_a = \sum_o \frac{c_{ao} V_o}{C_o} - B_a, \tag{3.12}$$

where

$$C_{aa'} = \sum_o \frac{c_{ao} c_{a'o}}{C_o} \tag{3.13}$$

can be considered as weights on the edges of the graph Γ_A on assessors illustrated in the right column of figure 1. The dimension of the reduced system (3.12) is the number N_A of assessors (rather than the sum of the numbers of assessors and objects), which gives some computational savings. Replacing one of the equations in (3.12), say that for the 'last' assessor, by equation (3.8) gives a system with a unique solution that can be solved for b by any method of numerical linear algebra, e.g. LUP decomposition [14]. Then v can be obtained from equation (3.11).

A slightly more sophisticated approach to incorporating a degeneracy-breaking condition into equations (3.12) is described in appendix B.

A key question with any black-box solution like the one presented here is how robust is the outcome? We propose two ways of quantifying the robustness. One is to bound how much the outcome would change if some of the scores were changed (e.g. representing mistakes or anomalous judgements). We treat this in appendix C. The other is to evaluate the posterior uncertainty of the outcomes, assuming normal distribution of the η_{ao}. This is treated in appendix D.

4. Case studies

We have tested the approach in three contexts. We report in detail on two case studies here. In the first case study, we use a computer-generated set of data containing true values of assessed items, assessor biases and confidences for the assessments and resulting scores. This has the advantage of allowing us to compare the values obtained by the new approach with the true underlying value of each item. The second case study is an evaluation of grant proposals using realistic data based on a university's internal competition. In this test, of course, there is no possibility to access 'true' values, so instead we compare the evidence for the models using a Bayesian approach (appendix E), and we compare their posterior uncertainties (appendix D). The third context in which we tested our method was assessment of students; we report briefly on this at the end of the section.

4.1. Case study 1: simulation

In the simulation, $N_O = 3000$ objects are assessed by a panel of $N_A = 15$ assessors. This choice was motivated by the number of outputs and reviewers in the applied mathematics unit of assessment at the UK's 2008 research assessment exercise. The simulation was carried out using Matlab, and the system of equations was solved using its built-in procedure, which computed the LU decomposition of L (with the last row replaced by the degeneracy-breaking condition (3.8)). The reduction to (3.12) was not used because $N_O = 3000$ is easily handled by modern personal computers.

True values of the items v_o were assumed to be normally distributed with a mean of 50 and standard deviation set to 15, but with v_o values truncated at 0 and 100. The assessor biases b_a were assumed to be normally distributed with a mean of 0 and a standard deviation of 15. Each assessment was considered to be done with high, medium or low confidence, and these were modelled using scaled uncertainties for the awarded scores, of $\sigma_{ao} = 5$, 10 or 15, respectively. The allocated scores follow equation (2.4), but truncated at 0 and 100.

With r assessors per item (which we took to be the same for each item in this instance), each simulation generated rN_O object scores s_{ao}. From these, we generated N_O value estimates $\hat{v}_o$ and N_A estimates of assessor biases $\hat{b}_a$ using the calibration processes. We then took the mean and maximum values of the errors in the estimates, $dv_o = |\hat{v}_o - v_o|$ and $db_a = |\hat{b}_a - b_a|$. Simple averaging also delivered a value estimate $\hat{v}_o$, as well as mean and maximal values of the errors dv_o. Finally, we determined the averages of the errors dv_o and db_a over 100 simulations. The results for these averaged mean and maximal errors in the scores are denoted by $\langle dv \rangle$ and $(dv)_{max}$, respectively, and those for the biases (for the calibrated approaches only) are denoted $\langle db \rangle$ and $(db)_{max}$.

Results for all three methods are presented in figures 2–4. The mean and maximal errors for the SA approach, the IBA method and the CWC approach are given in figures 2a–d and 3a–d. For demonstration purposes, we use three confidence levels rather than a continuous distribution. This allows us to clearly control differences in confidence levels in figures 2 and 3 and we do so by presenting four panels labelled (a)–(d). These represent different profiles, with the confidence for each assessment randomly allocated using probabilities for high, medium and low confidences in the ratios: (a) 1:1:1, (b) 1:1:2, (c) 1:2:1 and (d) 2:1:1. We observe that, for each method, the scores become more accurate (errors decrease) as the number of assessors per object r increases.

From figure 2a–d, with only two assessors per object, the SA method gives errors averaging about 10 points. Over $r = 6$ assessors per object are required to bring the mean error down to six points. Fisher's IBA, however, achieves this level of improvement with only two or three assessors. The CWC method delivers a further level of improvement of about one point. One also notes that, for the calibration approaches, relatively little is gained on average by employing more than four assessors per object. This result can be compared with [15] which found that five assessors per object was optimal in terms of accuracy over cost, for a procedure used by the Canadian Institutes of Health Research.

Figure 3 shows that IBA also leads to significant improvements in the maximal error values relative to those obtained through SA. With two assessors per object, maximal errors are reduced from about 45 to 30–35. The CWC approach does not appear to significantly improve upon this. However, with six assessors per object the maximal error value of about 25 delivered by the SA process is reduced to about 20 by IBA and to as low as 16 by CWC when half the assessments are done with a high degree of confidence in the scores.

Figure 4a gives the improvements achieved by the calibration methods as ratios of the mean errors coming from Fisher's IBA approach to the SA approach $\langle dv \rangle_{IBA}/\langle dv \rangle_{avg}$ and of the mean errors coming from the CWC approach to the SA approach $\langle dv \rangle_{CWC}/\langle dv \rangle_{avg}$. Smaller ratios mean greater
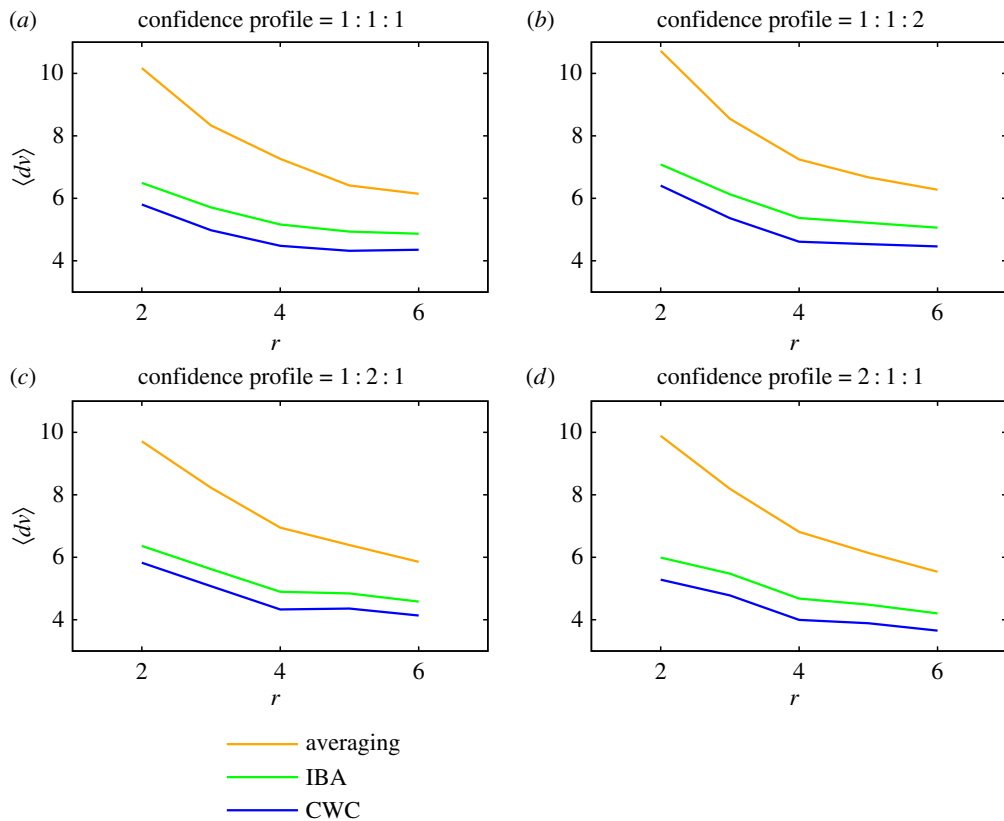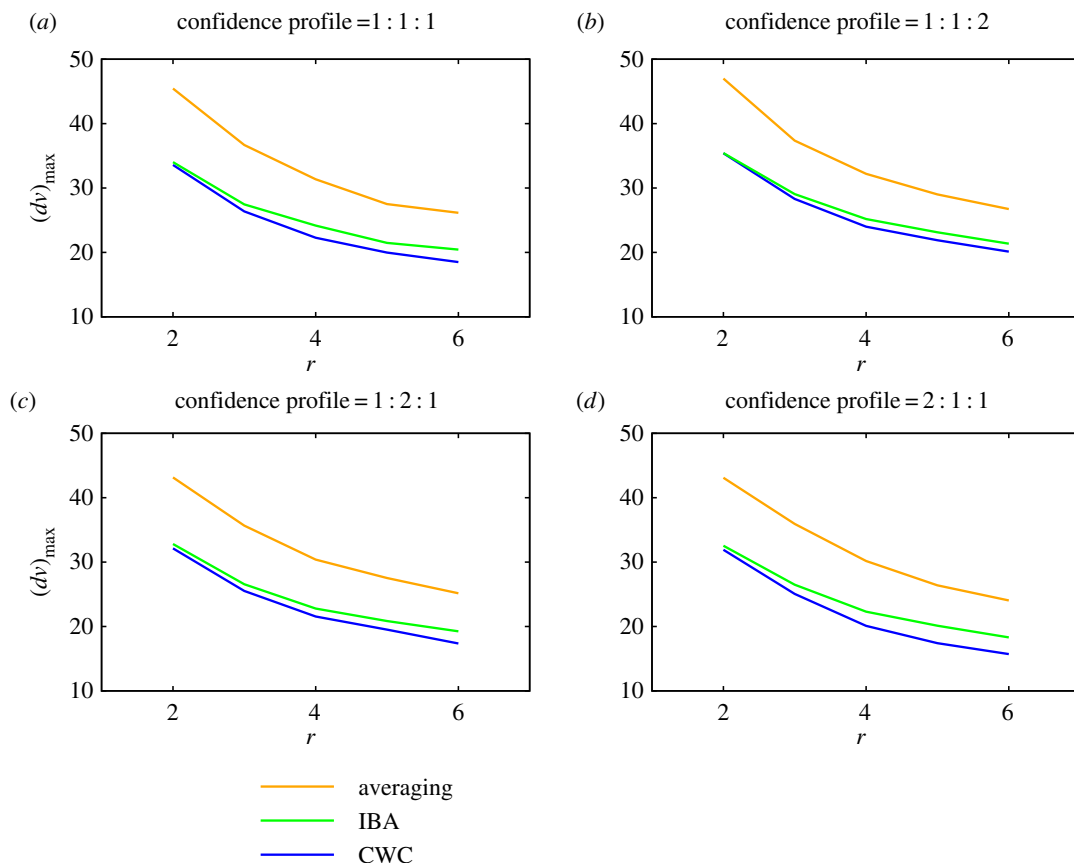
Figure 2. Mean errors plotted against the number r of assessors per object for the SA approach (upper curves, orange), the incomplete-block-analysis method (middle curves, green) and the calibration-with-confidence approach (lower curves, blue). The various panels represent different confidence profiles with probabilities for high, medium and low confidences in the ratios: (a) 1:1:1, (b) 1:1:2, (c) 1:2:1 and (d) 2:1:1.

accuracy on the part of the calibrated approaches. Figure 4b gives the analogous accuracy ratios for the maximal errors, namely $(dv)_{\text{max,IBA}}/(dv)_{\text{max,avg}}$ and $(dv)_{\text{max,CWC}}/(dv)_{\text{max,avg}}$. Figure 4a demonstrates that IBA delivers mean errors between about 60 and 80% of those coming from the SA approach, the better improvements being associated with lower assessor numbers. This is also the most desirable configuration for realistic assessments, as it represents employment of a minimal number of assessors per object. The CWC approach reduces errors by about a further 10 percentage points irrespective of the number of assessors.

4.2. Case study 2: grant proposals

To test CWC in a realistic setting, we adapted data from a university's internal competition for research funding, in which 43 proposals were evaluated by a panel of 11 assessors. Each proposal was graded by two assessors, who in addition each specified a confidence level in their grading in the form of high, medium or low. To respect confidentiality of the competition while making the data available, we not only anonymized the proposals and assessors but also made sufficient changes to the data (while preserving the statistical properties) so that attribution would not be possible. The actual panel used SA, but the assessors were also asked to provide confidences so that CWC could be applied for comparison. The panel awarded grants to the top 10 proposals. Our goals were firstly to see what differences would have been made by use of IBA or CWC, secondly to quantify the evidence for the three models from the data to determine which was most appropriate, and thirdly to compare the posterior uncertainties they provide.

To apply CWC, we translated the qualitative confidence levels of high, medium and low to values $c_{ao} = \lambda^2, 1, \lambda^{-2}$, respectively, with $\lambda = 1.75$. We chose $\lambda = 1.75$ as a reasonable guess at how the assessors used the confidence scale. One could include a computation to infer λ from the data, but our preference

Figure 3. Maximum errors plotted against the number r of assessors per object for the SA approach (upper curves, orange), the incomplete-block-analysis method (middle curves, green) and the calibration-with-confidence approach (lower curves, blue). The various panels represent different confidence profiles with probabilities for high, medium and low confidences in the ratios: (a) $1:1:1$, (b) $1:1:2$, (c) $1:2:1$ and (d) $2:1:1$.
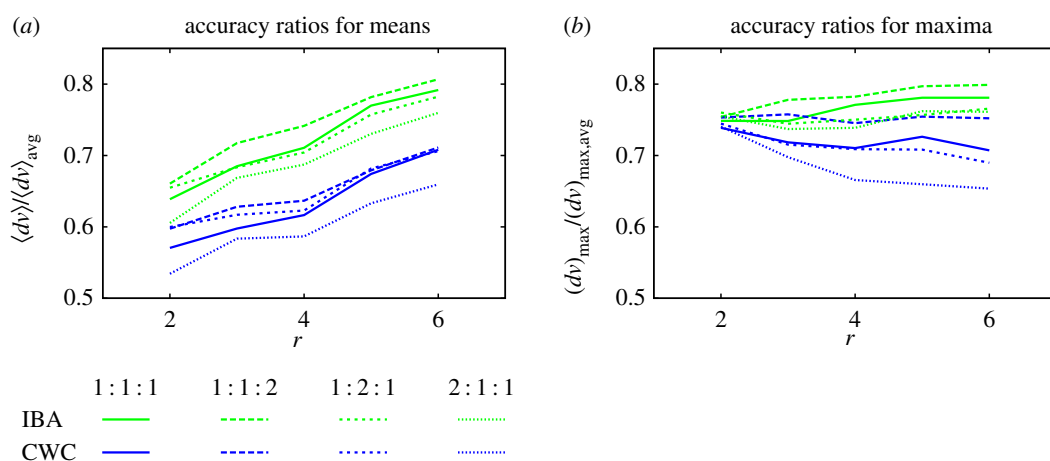
Figure 4. (a) The ratios $\langle dv \rangle_{IBA}/\langle dv \rangle_{avg}$ and $\langle dv \rangle_{CWC}/\langle dv \rangle_{avg}$ measure the mean improved accuracies of IBA (green curves) and CWC (blue), respectively, over SA. Smaller ratios indicate a greater degree of improvement over SA. (b) The analogous quantities for maximal errors are $(dv)_{max,IBA}/(dv)_{max,avg}$ and $(dv)_{max,CWC}/(dv)_{max,avg}$, respectively. The four line types correspond to relative probabilities of standard deviations of 5, 10 or 15, respectively, in the ratios $1:1:1$ (solid); $1:1:2$ (long-dashed); $1:2:1$ (short-dashed) and $2:1:1$ (dotted lines).

is for panel chairs to ask assessors to provide uncertainties rather than qualitative confidence levels, as indicated in §2, so we did not implement the inference of λ.

Figure 5a–c shows the resulting values inferred by the three methods, projected into the planes of (SA; IBA), (IBA; CWC) and (CWC; SA), respectively. Figure 5d is a Bland–Altman or Tukey mean-difference
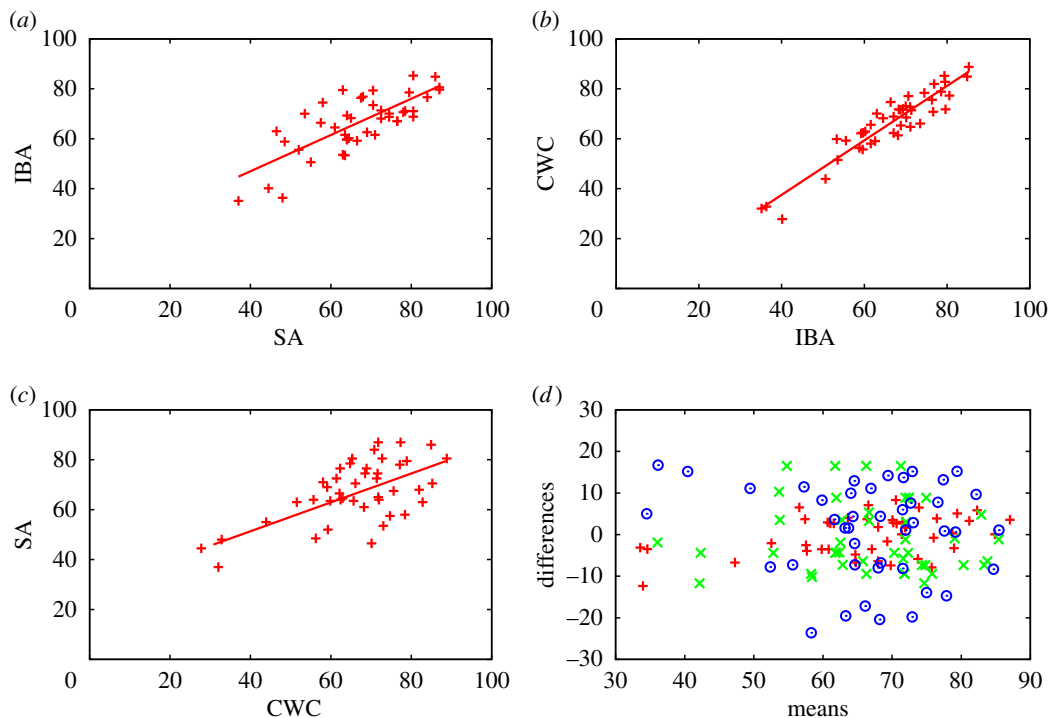
Figure 5. Correlations between the results coming from the three methods applied to Case Study 2. The three panels give the correlations between the outputs of (*a*) IBA and SA; (*b*) CWC and IBA and (*c*) SA and CWC. The coefficients of determination are given, respectively, by $R^2 = 0.5701; 0.8807$ and 0.3772. Panel (*d*) is a Bland–Altman or Tukey mean-difference plot of differences between results from pairs of approaches against their averages. The symbols '+' (red) compare CWC to IBA ($\mathcal{V}_{CWC} - \mathcal{V}_{IBA}$ versus $(\mathcal{V}_{IBA} + \mathcal{V}_{CWC})/2$); '×' (green) compare IBA to SA ($\mathcal{V}_{IBA} - \mathcal{V}_{avg}$ versus $(\mathcal{V}_{avg} + \mathcal{V}_{IBA})/2$); '○' (blue) compare SA to CWC ($\mathcal{V}_{avg} - \mathcal{V}_{CWC}$ versus $(\mathcal{V}_{CWC} + \mathcal{V}_{avg})/2$).

Table 2. The 43 grant proposals are identified as OA, OB, OC, . . . OZ, OA′, OB′, . . . OP′, OQ′. Here they are ranked according to their $\mathcal{V}_{avg}$, $\mathcal{V}_{IBA}$ and $\mathcal{V}_{CWC}$ values, representing the outcomes of SA, the IBA and CWC approaches. Proposals identified by CWC as belonging to the top 10 but missed by IBA are highlighted in boldface. Proposals identified by IBA or CWC as belonging to the top ten but missed by SA are highlighted in italics. Proposals which are not in the CWC top 10 are underlined.

rank	SA $\mathcal{V}_{avg}$	IBA $\mathcal{V}_{IBA}$	CWC $\mathcal{V}_{CWC}$
1	OH (87.0)	OA (85.3)	OA (88.8)
2	OP (87.0)	OC (84.9)	*OB* (85.2)
3	OC (86.0)	OH (80.6)	OC (84.9)
4	OS (84.0)	*OP* (79.7)	*OD* (82.8)
5	OA (80.5)	*OD* (79.5)	*OE* (82.0)
6	OM (80.5)	*OB* (79.4)	OF (78.9)
7	OZ (80.5)	OF (78.6)	***OG*** (78.4)
8	OF (79.5)	*OE* (76.9)	OH (77.3)
9	OA′ (78.5)	OS (76.7)	**OI** (77.1)
10	OI (78.0)	*OJ* (76.4)	*OJ* (75.6)

plot [16]. The correlations are not strong, though as we would expect, the correlation of IBA with CWC is stronger than those of either with SA. In particular, we note that the set of proposals rated in the top 10 varies substantially with the method used (table 2). The reason for the differences is that IBA and CWC attribute a significant range of biases to the assessors (table 3).

In the absence of 'true' values for the proposals, how can one decide which is the best method to use, and hence which outcome is preferred?

Table 3. Assessor statistics: assessors are labelled AA, . . . AK according to increasing CWC-biases (5th column). Here, we give the mean scores they awarded, standard deviations and IBA-biases too. The mean score awarded over all assessments was 66.9.

assessor	mean	s.d.	bias (IBA)	bias (CWC)
AK	84.2	16.6	14.6	17.7
AJ	61.0	19.2	8.7	12.6
AI	64.6	10.0	0.0	9.7
AH	76.6	9.1	10.0	9.1
AG	71.9	6.9	8.8	8.8
AF	65.9	5.6	5.7	2.0
AE	72.3	15.5	2.8	1.1
AD	61.0	21.9	−5.0	−3.6
AC	62.3	9.6	−12.4	−15.6
AB	58.3	6.4	−12.8	−16.6
AA	49.1	12.1	−20.7	−25.2

Table 4. Residuals (scaled by mean confidence in the case of CWC).

method	SA	IBA	CWC
residual	8602	4388	3156

Table 5. Bayesian log-evidences.

method	SA	IBA	CWC
log-evidence	−385	−389	−387

A first answer is to compare the 'residuals' that the methods leave after the least-squares fit. In the case of SA, this means the value of equation (3.1) obtained by taking the v_o to be the averaged scores and $b_a = 0$. For IBA, the residual is the value of (3.1) at the least-squares fit, taking all the $c_{ao} = 1$. For CWC, we take the value of (3.1) at the least-squares fit, divided by the average confidence over all assessments. The residuals are presented in table 4. From this point of view, we see clear improvement progressively from SA to IBA to CWC, providing an apparently compelling argument for the use of CWC.

As IBA and CWC have more free parameters (the biases) than SA, however, one should penalize them appropriately to make a correct comparison. Also although normalizing the residual for CWC by the average confidence sounds sensible, it is not clear it is the right way to compare CWC with IBA.

A principled answer is provided by Bayesian model comparison. In this procedure, the evidence provided by the data in favour of each model is quantified, and the best model is the one with the highest evidence. The procedure to quantify the evidence for the three models is described in appendix E. It depends on assumptions about the prior probability distribution for the parameters of the models, but we took 'ball' priors on the true values and on the biases (constrained by the degeneracy-breaking condition) and a truncated Jeffreys' prior on the variance of the noise. In the notation of appendix E, the parameters for the prior probability distributions were $\sigma_O = 22.5$, $\sigma_A = 15$, $w_{max} = 900$, $w_{min} = 1$. As the evidences come out to be small numbers (around 10^{-168}), we took their (natural) logarithms. The resulting log-evidences are shown in table 5. Simple averaging wins, but these values are so close together that we cannot make a strong conclusion about which method is most justified by the data. Furthermore, adjusting the prior probability distributions and the confidence weights changes which method has the highest evidence. We suspect that differences between the evidences for the models would become apparent if each proposal had been evaluated by more than two assessors.

A third approach is to evaluate the posterior uncertainty in the values assigned to the objects for the three methods, as detailed in appendix D, using (D 1) for IBA and CWC, and (D 2) for SA. The results are given in table 6. On this basis, the most precise results are given by CWC. None of them are very precise, however. A posterior uncertainty of eight means that we should consider values for the objects to have a

Table 6. Confidence-weighted root mean square uncertainties for the values (and biases in the cases of IBA and CWC). For SA, the weighting is according to the number n_o of assessors for object o.

method	SA	IBA	CWC
uncertainty	14.1	8.4	8.0

one-third chance of differing by more than eight from the outputted values. This means that for IBA and CWC, only the top three proposals of table 2 are reasonably assured of being in the top 10.

As the object of the competition was only to choose the best 10 proposals to fund, rather than assign values to each proposal, it might have been more appropriate to design just a classifier system (with a tunable parameter to make the right number in the 'fund' class) but our goal was to use it as a test of CWC.

The fact that three different methods with roughly equal evidence lead to drastically different allocation of the grants, and with large posterior uncertainties, highlights that better design of the panel assessment was required. Large variability of outcome even when just using SA but with different assessment graphs was already noted by Graves *et al*. [17]. A moral of our analysis is that to achieve a reliable outcome, the assessment procedure needs substantial advance design. We continue a discussion of design in appendices C and F, but substantial treatment is deferred to a future paper.

4.3. Third context: assessment of students

We also tested the method on undergraduate examination results for a degree with a flexible options system [18] and on the assessment of a multi-lecturer postgraduate module.

In the former case, as surrogates for the confidences in the marks we took the number of Credit Accumulation and Transfer Scheme (CATS) points for the module, which indicate the amount of time a student is expected to devote to the module (for readers used to the European credit transfer and accumulation system, 2 CATS points are equivalent to 1 ECTS point). The amount of assessment for a module is proportional to the CATS points. If it can be regarded as consisting of independent assessments of subcomponents, e.g. one per CATS point, with roughly equal variances, then the variance of the total score would be proportional to the number of CATS points. As the score is then normalized by the CATS points, the variance becomes inversely proportional to the CATS points, making confidence directly proportional to CATS points. The outcome of our analysis indicated significant differences in standards for the assessment of different modules, but as most modules counted for 15 or 18 CATS, this was not a strong test of the merits of including confidences in the analysis, so we do not report on it here.

For the postgraduate module, there were four lecturers plus module coordinator, who each assessed oral and written reports for some but not all of the students, according to availability and expertise (except the coordinator assessed them all). Each assessor provided a score and an uncertainty for each assessment. The results were combined using our method and the resulting value for each student was reported as the final mark. The lecturers agreed that the outcome was fair.

5. Discussion

We have presented and tested a method to calibrate assessors in a panel, taking account of differences in confidence that they express in their assessments. From a test on simulated data we found that calibration with confidence (CWC) generated closer estimates of the true values than Additive incomplete block analysis (IBA) or simple averaging (SA). A test on some real data, however, provided little evidence to distinguish between the methods, though they produced wildly different rankings, suggesting that the assessment procedure for that context needed more robust design. Nevertheless, CWC came ahead on posterior precision. We note that the default of assuming all assessment confidences to be equal results in IBA, which already represents a useful improvement over SA.

One of the principal conclusions from our analysis is that to achieve reliable outcomes from the methods we tested, requires good design of the assessment graph (showing which objects are evaluated by which assessors and with what confidences).

All three methods we compared are based on least-squares fitting. They may therefore be considered overly sensitive to outliers. An alternative approach which is less sensitive to outliers is based on medians rather than means. For example, Tukey's Median Polish [19] is a median-based version of Fisher's IBA. It would be good to develop a version of it that takes confidences into account too.

Some other drawbacks of our CWC method are:

— It requires assessors to give reliable uncertainties; if assessors differ in their confidence estimates the method gives higher weight to those who give higher confidences. In particular, one needs to guard against an assessor giving unwarrantedly high confidence for a particular assessment. There is a case for calibrating confidences too.
— Bias effects may be more subtle than just an additive effect; for example, an assessor may be more generous (or perhaps tougher) on topics in which they have high confidence, or they may use a shorter or longer part of the scale than other assessors.
— Some organizations insist on round-number scores; this goes against the spirit of our approach and is awkward for assessors who may rightly wish to rate an object as between two of the allowed grades. The requirement is perhaps based on the laudable idea of not wishing to imply higher accuracy than is warranted, yet in our opinion this is better dealt with by reporting an uncertainty for each result on a continuous scale.
— Some organizations may insist that scores cannot go beyond certain limits, which is awkward for an assessor if after evaluating several objects highly they find there are some they wish to rate even higher.

There are a number of refinements which one could introduce to the core method, addressing some of these drawbacks. These include how to deal with different types of bias, different scales for confidence, different ways to remove the degeneracy in the equations, how to deal with the endpoints on a marking scale, and how to choose the assessment graph. Some suggestions are made in the appendices, along with mathematical treatment of the robustness of the method and of computation of the Bayesian evidence for the models.

An advantage of our type of calibration is that it does not produce the artificial discontinuities across field boundaries that tend to arise if the domain is partitioned into fields and evaluation in each field carried out separately. In the UK Research Assessment Exercise 2008, for example, there is evidence that different panels had different standards [20]. Although RAE2008 stated that cross-panel comparisons are not justified, some universities have used such comparisons to help decide on how much to resource different departments. Our approach would take advantage of cross-panel referrals (which was part of RAE2008 for work in the overlaps between panels) to infer relative standards and hence to normalize the outcomes.

We suggest that a method such as this, which takes into account declared confidences in each assessment, is well suited to a multitude of situations in which a number of objects is assessed by a panel. We acknowledge, however, that this approach requires an investment in training assessors to estimate their uncertainties and in constructing a sufficiently strongly connected assessment graph. Different panels will deal with the trade-off between investment of effort and accuracy of results in different ways.

Appendix A. Scale for confidences

We motivated the model by proposing that the noise terms be of the form $\sigma_{ao}\eta_{ao}$ with the η_{ao} independent zero-mean random variables with unit variance, so that the σ_{ao} are standard deviations. Nevertheless, multiplying all the confidences by the same number does not change the results of the least-squares

13

fit, nor our quantifications of robustness (appendices C and D). Thus, the η_{ao} can be taken to have any variance w, as long as it is the same for all assessments. It is only ratios of confidences that have significance.

The fitting procedure can be extended to infer a best-fit value for w. Even if the assessors provide confidences based on assuming $w = 1$, the best fit for w is not 1 in general. Assuming independent Gaussian errors, the maximum-likelihood value for w comes out to be

$$\bar{w} = \frac{R}{N},$$

where

$$R = \sum_{ao} c_{ao}(s_{ao} - \bar{v}_o - \bar{b}_a)^2 \tag{A 1}$$

is the residual from the least-squares fit $(\bar{v}, \bar{b})$ for (v, b) and N is the total number of assessments. The posterior distribution for w, given a prior distribution, is obtained in appendix D.

Appendix B. Degeneracy-breaking conditions

We can remove the degeneracy in equations (3.2) and (3.3) in different manners from equation (3.8) used here. Indeed, use of (3.8) can lead to an average shift from the scores to the true values. This does not matter if only a ranking is required, but if the actual values are important (e.g. for degree classification), then a better choice of degeneracy-breaking condition is needed.

A preferable confidence-weighted degeneracy-breaking condition is

$$\sum_a C_a' b_a = 0, \tag{B 1}$$

which from (3.3) automatically implies $\sum_o C_o v_o = \sum_{ao} c_{ao} s_{ao}$, thus avoiding the possibility of such systematic shifts.

From a theoretical perspective, however, the best choice of degeneracy-breaking condition is to choose a reference value v_{ref} (think of a notional desired mean) and require

$$\sum_{ao} c_{ao}(v_o - b_a) = C v_{\text{ref}}, \tag{B 2}$$

where

$$C = \sum_{ao} c_{ao}. \tag{B 3}$$

Using the notation in (3.6) and (3.7), this can equivalently be written as

$$\sum_o C_o v_o - \sum_a C_a' b_a = C v_{\text{ref}}. \tag{B 4}$$

To reduce the possible average shift from confidence-weighted average scores to true values, the reference value v_{ref} should be chosen near the confidence-weighted average score

$$\bar{s} = \sum_{ao} \frac{c_{ao} s_{ao}}{C}. \tag{B 5}$$

Choosing v_{ref} exactly equal to $\bar{s}$ gives (B 1), which makes the confidence-weighted average bias come out to 0 and the confidence-weighted average value come out to $\bar{s}$. We will show in appendix C, however, that the results are a factor $\sqrt{2}$ more robust to changes in the scores if v_{ref} is chosen to be fixed rather than dependent on the scores.

For any affine choice of degeneracy-breaking condition on the biases, $\sum_a \beta_a b_a = \gamma$, the reduced system (3.12) can be solved either by replacing one of the equations by the degeneracy-breaking condition as in §3, or by appending an additional unknown s, adding $\beta_a s$ to the left-hand side of each equation (3.12), and appending the degeneracy-breaking equation as an additional equation. The latter option has the advantage of preserving the symmetry of the matrix representing the system of equations and hence twice as efficient algorithms to solve them (symmetric indefinite factorization). The additional unknown s comes out to be 0 because of the relation $\alpha(V, B) = 0$ mentioned after (3.10).

Appendix C. Robustness to changes in the scores

Here, we present our approach to the quantification of the robustness of our method to small changes in the scores, using norms that take into account the confidences.

For $s = (s_{ao})_{(a,o) \in E}$, define the operator K by

$$Ks = \begin{bmatrix} V \\ B \end{bmatrix}, \tag{C1}$$

as a shorthand for the definitions in equations (3.4) and (3.5), so that equations (3.2) and (3.3) can be written as

$$L \begin{bmatrix} v \\ b \end{bmatrix} = Ks. \tag{C2}$$

Thus, if a change Δs is made to the scores, we obtain changes Δv, Δb of magnitude bounded by

$$\left\| \begin{matrix} \Delta v \\ \Delta b \end{matrix} \right\| \leq \|L^{-1}K\| \|\Delta s\|, \tag{C3}$$

where L^{-1} is defined by restricting the domain of L to (3.8) and its range to $\alpha(V, B) = 0$, and appropriate norms are chosen. In this appendix, we propose that appropriate choices of norms are

$$\|\Delta s\|_{\text{scores}} = \sqrt{\sum_{ao} c_{ao} \, \Delta s_{ao}^2}, \tag{C4}$$

$$\|(\Delta v, \Delta b)\|_{\text{results}} = \sqrt{\sum_{ao} c_{ao}(\Delta v_o^2 + \Delta b_a^2)} = \sqrt{\sum_o C_o \, \Delta v_o^2 + \sum_a C'_a \, \Delta b_a^2}, \tag{C5}$$

and the associated operator norm from scores to results for $\|L^{-1}K\|$. With the confidence-weighted degeneracy-breaking condition $\sum C'_a b_a = 0$, (B1) instead of (3.8), we obtain

$$\|L^{-1}K\| \leq \frac{\sqrt{2}}{\sqrt{\mu_2}}, \tag{C6}$$

where μ_2 is the second smallest eigenvalue of a certain matrix M formed from the confidences (see (C10)). In particular, this gives

$$|\delta v_o| \leq \frac{1}{\sqrt{C_o}} \frac{\sqrt{2}}{\sqrt{\mu_2}} \sqrt{\sum_{ao} c_{ao} \, \delta s_{ao}^2}. \tag{C7}$$

The factor of $\sqrt{2}$ can be removed if one switches to an ideal degeneracy-breaking condition as in (B2) of appendix B.

As a consequence, to maximize the robustness of the results, the task for the designer of E is to make none of the C_o much smaller than the others and to make μ_2 significantly larger than 0. The former is evident (no object should receive significantly less assessment or less expert assessment than the others). The latter is the mathematical expression of how well connected is the graph Γ (equivalently Γ_A). To design the graph Γ requires a guess of the confidence levels that assessors are likely to give to their assessments (based on knowing their areas of expertise and their thoroughness or otherwise) and a compromise between assigning an object to only the most expert assessors for that object and the need to achieve a chain of comparisons between any pair of assessors.

We now go into detail, derive the above bounds and describe some computational short cuts.

One can measure the size of a change Δs_{ao} to a score s_{ao} by comparing it to the declared uncertainty σ_{ao}. Thus, we take the size of Δs_{ao} to be $\sqrt{c_{ao}} \, |\Delta s_{ao}|$. We propose to measure the size of an array Δs of changes Δs_{ao} to the scores by the square root of the sum of squares of the sizes of the changes to each score, as in (C4). Supremum or sum-norms could also be considered but we will stick to this choice here.

It is also reasonable to measure the size of a change Δv_o to a true value v_o by comparing it to the uncertainty implied by the sum of confidences in the scores for object o. Thus, the size of Δv_o is defined to be $\sqrt{C_o} \, |\Delta v_o|$, where C_o is the total confidence in the assessment of object o. Similarly, we measure the size of a change Δb_a in bias b_a by $\sqrt{C'_a} \, |\Delta b_a|$ where C'_a is the total confidence expressed by a given assessor. Finally, we measure the size of a change $(\Delta v, \Delta b)$ to the vector of values and biases by the square root of sum of squares of the individual sizes, as in (C5).

The size of the operator $L^{-1}K$ is measured by the operator norm from scores to results, i.e.

$$\|L^{-1}K\| = \sup_{\Delta s \neq 0} \frac{\|L^{-1}K\Delta s\|_{results}}{\|\Delta s\|_{scores}}. \tag{C 8}$$

The operator $L^{-1}K$ is equivalent to orthogonal projection with respect to the norm (C 4) from the scores to the subspace Σ of the form $s_{ao} = v_o + b_a$ with a degeneracy-breaking condition to eliminate the ambiguity in direction of the vector $v_o = 1, b_a = -1$.

The tightest bounds in (C 3) are obtained by choosing the degeneracy-breaking condition to correspond to a plane perpendicular to this vector with respect to the inner product corresponding to equation (C 5). Thus, we choose degeneracy-breaking condition (B 2).

Theorem. *For a connected graph Γ and with the degeneracy-breaking condition (B 2), the size of the change $(\Delta v, \Delta b)$ resulting from a given array of changes Δs in scores is bounded by*

$$\|(\Delta v, \Delta b)\|_{results} \leq \frac{1}{\sqrt{\mu_2}} \|\Delta s\|_{scores}, \tag{C 9}$$

where μ_2 is the second smallest eigenvalue of the matrix

$$M = \begin{bmatrix} I_{N_O} & D \\ D^T & I_{N_A} \end{bmatrix}, \tag{C 10}$$

$$D^T_{ao} = \frac{c_{ao}}{\sqrt{C_o C'_a}}, \tag{C 11}$$

N_A, N_O *are the numbers of assessors and objects, respectively, and for $k \in \mathbb{N}$, I_k is the identity matrix of rank k.*

Proof. Firstly, the orthogonal projection in metric (C 4) from s to the subspace Σ never increases length. Secondly, if $\Delta s_{ao} = \Delta v_o + \Delta b_a$ with $\sum_{ao} c_{ao}(\Delta v_o - \Delta b_a) = 0$ then

$$\|\Delta s\|^2_{scores} = \sum_{ao} c_{ao}(\Delta v_o + \Delta b_a)^2 = g^T M g, \tag{C 12}$$

where g is the vector with components

$$g_o = \tilde{v}_o := \sqrt{C_o}\, \Delta v_o, \tag{C 13}$$

$$g_a = \tilde{b}_a := \sqrt{C'_a}\, \Delta b_a. \tag{C 14}$$

Then, because we restricted to the subspace orthogonal to the null vector in results-norm, and M is non-negative and symmetric

$$g^T M g \geq \mu_2 \sum_i g_i^2 = \mu_2 \|(\Delta v, \Delta b)\|^2_{results},$$

where index i ranges over all objects and assessors. Positivity of μ_2 holds as soon as the graph Γ is connected, because M is a transformation of the weighted graph-Laplacian to scaled variables [21], so dividing by μ_2 and taking the square root yields the result. ∎

The computation of the eigenvalue μ_2 of M can be reduced from dimension $N_A + N_O$ to dimension N_A by

Proposition. *If $N_A \geq 2$, the second smallest eigenvalue μ_2 of M is related to the second largest eigenvalue λ_2 of $D^T D$ by*

$$\mu_2 = 1 - \sqrt{\lambda_2}. \tag{C 15}$$

If $N_A = 1$ and $N_O \geq 2$, then $\mu_2 = 1$. If both are 1, then $\mu_2 = 2$.

Proof. The equations for an eigenvalue–eigenvector pair $\mu, (\tilde{v}, \tilde{b})$ of M are

$$\tilde{v} + D\tilde{b} = \mu\tilde{v} \tag{C 16}$$

$$D^T\tilde{v} + \tilde{b} = \mu\tilde{b}. \tag{C 17}$$

Applying D^T to the first equation, multiplying the second by $(1 - \mu)$, and then substituting for $(1 - \mu)D^T\tilde{v}$ in the second yields

$$D^T D\tilde{b} = (1 - \mu)^2\tilde{b}. \tag{C 18}$$

Thus, either $\tilde{b} = 0$ or $(1 - \mu)^2$ is an eigenvalue λ of $D^T D$. In the first case, equation (C 16) implies $\mu = 1$, so if $\mu \neq 1$ then $(1 - \mu)^2$ is an eigenvalue of $D^T D$.

Conversely, if $(\lambda, \tilde{b})$ is an eigenvalue–eigenvector pair for $D^{\mathrm{T}}D$ with $\lambda \neq 0$ then $\lambda > 0$ because $D^{\mathrm{T}}D$ is non-negative, so put $\tilde{v} = \pm D\tilde{b}/\sqrt{\lambda}$ to see that $(\tilde{v}, \tilde{b})$ is an eigenvector of M with eigenvalue $\mu = 1 \pm \sqrt{\lambda}$. If $\lambda = 0$ and $D\tilde{b} = 0$ then $\mu = 1$ is an eigenvalue of M with eigenvector $(\tilde{v}, \tilde{b})$ for any $\tilde{v}$ with $D^{\mathrm{T}}\tilde{v} = 0$, e.g. $\tilde{v} = 0$.

Thus, there is a two-to-one correspondence between eigenvalues μ of M not equal to 1 and positive eigenvalues λ of $D^{\mathrm{T}}D$ (counting multiplicity): $\mu = 1 \pm \sqrt{\lambda}$. Any remaining eigenvalues are 1 for M and 0 for $D^{\mathrm{T}}D$. The degeneracy gives an eigenvector $\tilde{v}_o = \sqrt{C_o}, \tilde{b}_a = -\sqrt{C_a'}$ of M with eigenvalue 0 and it corresponds to an eigenvalue 1 of $D^{\mathrm{T}}D$. All other eigenvalues of M are non-negative because M is. All other eigenvalues of $D^{\mathrm{T}}D$ are less than or equal to 1 by the Cauchy–Schwarz inequality. So if the second largest eigenvalue λ_2 of $D^{\mathrm{T}}D$ (counting multiplicity) is positive, then the second smallest eigenvalue μ_2 of M (counting multiplicity) is $1 - \sqrt{\lambda_2}$. If $\lambda_2 = 0$, then $\mu_2 = 1$ because existence of λ_2 implies $N_{\mathrm{A}} \geq 2$ so M has dimension at least three and we have only two simple eigenvalues $\mu = 0$ and 2 from the simple eigenvalue 1 of $D^{\mathrm{T}}D$, so M must have another one but any other value than 1 would give a positive λ_2; so the same formula holds. If there is no second eigenvalue of $D^{\mathrm{T}}D$ (because $N_{\mathrm{A}} = 1$), then if $N_{\mathrm{O}} \geq 2$ the second largest eigenvalue of M must be 1 by the same argument. If both N_{A} and N_{O} are 1, then the second largest eigenvalue of M is the other one associated with the eigenvalue 1 of $D^{\mathrm{T}}D$, namely 2. ∎

Note that

$$(D^{\mathrm{T}}D)_{aa'} = \frac{C_{aa'}}{\sqrt{C_a' C_{a'}'}}$$

is a similarity transformation of (3.13). As examples of second eigenvalues, putting unit confidences on the graphs in the left column of figure 1 we calculate $\lambda_2 = \frac{1}{3}, \frac{2}{3}, 1$ for cases (a), (b), (c) in the right column, giving $\mu_2 = 1 - \sqrt{\frac{1}{3}}, 1 - \sqrt{\frac{2}{3}}, 0$, respectively.

Finally, a user may prefer to use the degeneracy-breaking condition (B 1) rather than (B 2), perhaps out of uncertainty about what value of v_{ref} to use. Or a user may be happy to use (B 2) with v_{ref} equal to the confidence-weighted average score, but wants v_{ref} to follow this average score if changes are made to the scores. That comes out equivalent to using (B 1). So we extend our discussion of robustness to treat this case. We find it makes the bounds increase by a factor of only $\sqrt{2}$.

Proposition. *For Γ connected and using degeneracy-breaking condition (B 1), the size of $(\Delta v, \Delta b)$ resulting from changes Δs to the scores is at most $(\sqrt{2}/\sqrt{\mu_2}) \| \Delta s \|_{\mathrm{scores}}$.*

Proof. If the degeneracy-breaking condition (B 2) gives a change $(\Delta v, \Delta b)$ for a change Δs to the scores, then switching to degeneracy-breaking condition (B 1) just adds an amount k of the null vector $\mathbf{n} = (\mathbf{1}, -\mathbf{1})$ to achieve $\sum_a C_a' (\Delta b_a - k) = 0$, i.e.

$$k = \frac{\sum_a C_a' \Delta b_a}{C}. \tag{C 19}$$

In the results metric, the null vector has length $\sqrt{\sum_o C_o + \sum_a C_a'} = \sqrt{2C}$. Thus, the correction has length $|k|\sqrt{2C} = \sqrt{2/C} \, |\sum C_a' \Delta b_a|$. Using the condition (B 2), we can write $\sum_a C_a' \Delta b_a = \frac{1}{2}(\sum_a C_a' \Delta b_a + \sum_o C_o \Delta v_o)$, which one can recognize as one half of the inner product of $(\mathbf{1}, \mathbf{1})$ with $(\Delta v, \Delta b)$ in results-norm, so it is bounded by $\sqrt{C/2} \, \|(\Delta v, \Delta b)\|$. Thus, the length of the correction vector is at most that of $(\Delta v, \Delta b)$. The correction is perpendicular to $(\Delta v, \Delta b)$, thus the vector sum has length at most $\sqrt{2} \, \|(\Delta v, \Delta b)\|$. ∎

One may also ask about robustness with respect to changes in the confidences c_{ao}. If an assessor declares extra high confidence for an evaluation, for example, that can significantly skew the resulting v and b. The analysis is more subtle, however, because of how the c_{ao} appear in the equations and we do not treat it here.

Appendix D. Posterior probability distribution

Another point of view on robustness is the Bayesian one. From a prior probability on (v, b) and a model for the η_{ao}, one can infer a posterior probability for (v, b), whose inverse width tells one how robust is the inference.

In the case of flat prior on (v, b), prescribed w, Gaussian noise, and an affine degeneracy-breaking condition, the posterior is Gaussian with mean at the value solving equations (3.2) and (3.3) and the degeneracy-breaking condition, and with covariance matrix related to L^{-1}. Specifically, the posterior

probability density for (v, b) is proportional to

$$\exp -\frac{S}{2w},$$

constrained to the degeneracy-breaking hyperplane, where

$$S = \sum_{ao} c_{ao}(s_{ao} - v_o - b_a)^2.$$

Using (C 12) and (A 1), this can be written as

$$\exp -\frac{1}{2w}(g^T M g + R),$$

with $(\Delta v, \Delta b)$ being the deviations of (v, b) from the least-squares fit. Thus, the covariance matrix in these scaled variables is wM^{-1}, where for degeneracy-breaking condition $\gamma^T g = K$, M^{-1} is interpreted as the limit as $t \to \infty$ of $(M + t\gamma\gamma^T)^{-1}$. Using the degeneracy-breaking condition (B 2) or equivalently (B 4) for which γ is in the null direction of M and diagonalizing the matrix, we obtain widths $\sqrt{w/\mu_j}$ for the posterior on g in the eigendirections of M, where μ_j are the positive eigenvalues of M. Thus, the robustness of the inference is again determined by μ_2, but scaled by $\sqrt{w}$.

A slightly more sophisticated approach is to consider w to be unknown also. Given a prior density ρ for w (which could be peaked around 1 if the assessors are assigning confidences via uncertainties, but following Jeffreys would be better chosen to be $1/w$ if there is no information about the scale for the confidences), the posterior density for (w, v, b) is proportional to

$$\rho(w)w^{-N/2} \exp -\frac{S}{2w},$$

where again N is the number of assessments. The maximum of the posterior probability density is determined by the least-squares fit for (v, b) (which is independent of w) and the following equation for w:

$$\frac{\rho'(w)}{\rho(w)} - \frac{N}{2w} + \frac{S}{2w^2} = 0.$$

For N large, the peak of the posterior has w near the previously determined maximum-likelihood value $\bar{w} = R/N$. For example, taking Jeffreys' prior, the peak is at $w = R/(N - 2)$. Integrating over w (with Jeffreys' prior) one finds the marginal posterior for (v, b) to be proportional to

$$(g^T M g + R)^{-N/2}.$$

Incorporating an affine degeneracy-breaking condition, this is a $(N_O + N_A - 1)$-variate Student distribution with $\nu = N - N_O - N_A + 1$ degrees of freedom. Its covariance matrix is $w^* M^{-1}$ with

$$w^* = \frac{R}{\nu - 2}$$

and M^{-1} interpreted by imposing the chosen degeneracy-breaking condition as above.

So for the degeneracy-breaking condition (B 2), the robustness of the inference is given by widths $\sqrt{w^*/\mu_j}$ for $j \geq 2$, in the eigendirections of M on g. In particular, the confidence-weighted root mean square uncertainty σ for the components of the vector (v, b) is

$$\sigma = \sqrt{\frac{w^*}{2C} \sum_{j \geq 2} \frac{1}{\mu_j}} = \sqrt{\frac{R \operatorname{Tr} M^{-1}}{2(\nu - 2)C}}, \tag{D 1}$$

where Tr denotes the trace and, again, M^{-1} is interpreted by restricting to the degeneracy-breaking plane. Marginal posteriors for each v_o and b_a can be extracted, but it must be understood that in general they are significantly correlated. One way to do this in the case of degeneracy-breaking condition (B 2) is to find the orthogonal matrix O to diagonalize M as $O^T D O$ with $D = \operatorname{diag}(\mu_j)$, and then the posterior variance of g_i is $w^* \sum_{j>1} O_{ji}^2/\mu_j$, but there may be ways to evaluate it without diagonalizing M.

For the case of SA, the root mean square posterior uncertainty in the values, weighted by the numbers n_o of assessors for object o, is

$$\sigma = \sqrt{\frac{R}{N - N_O}}, \tag{D 2}$$

where R is defined in (E 3) of appendix E. This can be derived in an analogous fashion to (D 1) via a Student distribution again, but with $M = I_{N_O}$.

Appendix E. Model comparison

Here we describe the method used in case study 2 to compare the three models.

Bayesian model comparison is based on computing how much evidence there is for each proposed model (e.g. ch. 28 of [22]). The evidence for a model M given data D is $P(D \mid M)$. Given strength of belief $P(M)$ in model M prior to the data (relative to other models), one can multiply it by the evidence to obtain the posterior strength of belief in model M. It is convenient to replace multiplication by addition, thus we define the log-evidence

$$\mathrm{LE}(M \mid D) = \log P(D \mid M).$$

If the model M has free parameters μ, then

$$P(D \mid M) = \int P(D \mid M, \mu) P_M(\mu) \, \mathrm{d}\mu,$$

where $P_M(\mu)$ is a prior probability density on μ.

Let there be N_O objects, N_A assessors, let s_{ao} be the score returned by assessor a for object o, c_{ao} the confidence in this score in the case of calibration with confidence, s be the collection of scores and N be their number.

First, we compute the evidence for SA. Then we treat calibrate with confidence (CWC) and lastly IBA because it is a special case of CWC.

E.1. Simple averaging

For SA, the model is that

$$s_{ao} = v_o + \varepsilon_{ao}$$

for some unknown vector v of 'true' values v_o, with ε_{ao} iid normal $N(0, w)$ for some unknown variance w. Then the probability density for the scores s is

$$P(s \mid v, w) = \prod \frac{e^{-(s_{ao} - v_o)^2/2w}}{\sqrt{2\pi w}} = (2\pi w)^{-N/2} e^{-(1/2w) \sum (s_{ao} - v_o)^2},$$

with the product and sum being over the assessments that were carried out.

To work out the evidence for SA the model must include a prior probability density for v and w. The simplest proposal would be $\Delta^{-N_O} L^{-1} w^{-1}$ on $v_o \in [v_{\min}, v_{\max}]$, $w \in [w_{\min}, w_{\max}]$, where $\Delta = v_{\max} - v_{\min}$ and $L = \log(w_{\max}/w_{\min})$. This is the product of a 'box' prior on v and Jeffreys' prior on w (truncated to an interval and normalized). For comparison with the other models, however, it is easier to replace the box prior on v by a 'ball' prior, giving

$$P_{SA}(v, w) = \frac{1}{Z_O L w}$$

on

$$\sum_o (v_o - v_{\mathrm{ref}})^2 \leq N_O \sigma_O^2, \tag{E 1}$$

for some anticipated average score v_{ref} and upper estimate of the width σ_O of the distribution of values v_o. The normalization is

$$Z_O = \frac{(\pi N_O \sigma_O^2)^{N_O/2}}{\Gamma(N_O/2 + 1)},$$

where Γ is the Gamma function. For $w_{\min}$ is it reasonable to choose u^2 where u is the smallest change any assessor could contemplate. For $w_{\max}$ it is reasonable to choose σ_O^2.

For each object o,

$$\sum_a (s_{ao} - v_o)^2 = n_o (v_o - \bar{s}_o)^2 + R_o,$$

where n_o is the number of assessors for object o, $\bar{s}_o$ is the mean of their scores, and the residual

$$R_o = \sum_a (s_{ao} - \bar{s}_o)^2. \tag{E 2}$$

Thus

$$P(s \mid v, w) P_{SA}(v, w) = \frac{1}{Z_O L w} (2\pi w)^{-N/2} e^{-(1/2w) \sum_o n_o (v_o - \bar{s}_o)^2} e^{-R/2w},$$

where

$$R = \sum_o R_o. \tag{E 3}$$

To integrate this over v and w, we assume the bulk of the probability distribution lies in the product of the ball and the interval, and so approximate by extending the range of integration to $\mathbb{R}^{N_O} \times (0, \infty)$. Integrating the exponential over v_o produces a factor

$$\sqrt{\frac{2\pi w}{n_o}}.$$

Thus, integrating over all components of v yields

$$\frac{1}{Z_O L w}(2\pi w)^{-(N-N_O)/2} e^{-R/2w} \prod n_o^{-1/2}.$$

Integrating this over w, we obtain the evidence

$$P(\text{SA} \mid s) = \frac{1}{Z_O L}(\pi R)^{-(N-N_O)/2} \Gamma\left(\frac{N-N_O}{2}\right) \prod n_o^{-1/2}$$

and the log-evidence

$$\text{LE}(\text{SA} \mid s) = \log \Gamma\left(\frac{N-N_O}{2}\right) - \frac{N-N_O}{2} \log \pi R - \frac{1}{2}\sum_o \log n_o - \log Z_O - \log L.$$

E.2. Calibration with confidence

For calibrate with confidence (CWC), the model is

$$s_{ao} = v_o + b_a + \sigma_{ao}\eta_{ao},$$

for some unknown vectors v of true values v_o, and b of assessor biases b_a, with η_{ao} iid normal $N(0, w)$ for some unknown variance w. The uncertainties σ_{ao} correspond to confidences c_{ao} by $\sigma_{ao} = 1/\sqrt{c_{ao}}$, which are considered as given (one could propose a generative model for them too, but that would require further analysis). Then the probability density for s is

$$P(s \mid v, b, w) = \prod \frac{e^{-c_{ao}(s_{ao}-v_o-b_a)^2/2w}}{\sqrt{2\pi w/c_{ao}}} = (2\pi w)^{-N/2} e^{-(1/2w)\sum c_{ao}(s_{ao}-v_o-b_a)^2} \prod c_{ao}^{1/2}.$$

For prior probability density over the parameters v, b, w, we want to build in a degeneracy-breaking condition. We used $\sum_a b_a = 0$ in our calculation, thus we take prior 'density'

$$P_{\text{CWC}}(v, b, w) = \frac{1}{Z_O Z_A L w} \delta\left(\sum_a b_a\right)$$

on the product of the balls (E 1) and $\sum_a b_a^2 \le N_A \sigma_A^2$ and interval $[w_{\min}, w_{\max}]$, where δ is the delta function. Here, σ_A is an estimated upper bound for the standard deviation of the biases, and the normalization is

$$Z_A = \frac{1}{\sqrt{N_A}} \frac{(\pi N_A \sigma_A^2)^{(N_A-1)/2}}{\Gamma((N_A+1)/2)}.$$

Note that the interpretation of w is not the same as for SA, so one might choose a different prior for it. For example, if the σ_{ao} are fairly accurate values for the uncertainties in the scores then the prior for w should be peaked around $w = 1$, but if they are on an undetermined scale a truncated Jeffreys prior is sensible. The only thing is that one might want to choose a different interval for it, but for application to IBA where the $c_{ao} = 1$ or to CWC if the c_{ao} are on a scale centred around 1, such as we have used to translate the quantitative high/medium/low confidence ratings, the same interval should be reasonable. Similarly, one might want to use a different value for σ_O if one believes that the spread in values is more due to variation in assessor bias than true value, but in our case we think it reasonable to use the same σ_O.

Thus,

$$P(s \mid v, b, w)P_{\mathrm{CWC}}(v, b, w)$$

$$= \frac{1}{Z_O Z_A L w}(2\pi w)^{-N/2}e^{-(1/2w)\sum c_{ao}(v_o + b_a - s_{ao})^2}\delta\left(\sum_a b_a\right)\prod\sqrt{c_{ao}}.$$

Again we assume the bulk of this lies in the product of balls and interval, so we approximate its integral by extending the domains to infinity. Now

$$\sum c_{ao}(s_{ao} - v_o - b_a)^2 = h^{\mathrm{T}}Ah + R,$$

where h is the vector with $N_O + N_A$ components, $h_o = (v_o - \bar{v}_o)$, $h_a = (b_a - \bar{b}_a)$, $(\bar{v}, \bar{b})$ is any least-squares fit to this model (without loss of generality satisfying the degeneracy-breaking condition), the residual R is now $\sum c_{ao}(s_{ao} - \bar{v}_o - \bar{b}_a)^2$ (as in (A 1)) and A is the matrix with block form

$$\begin{bmatrix} \mathrm{diag}(C_o) & c^{\mathrm{T}} \\ c & \mathrm{diag}(C_a') \end{bmatrix}.$$

Choose one assessor, say n, and integrate over b_n. This yields

$$\frac{1}{Z_O Z_A L w}(2\pi w)^{-N/2}e^{-(1/2w)(\tilde{h}^{\mathrm{T}}\tilde{A}\tilde{h}+R)}\prod\sqrt{c_{ao}},$$

with $\tilde{h}$ being the remaining components of h and

$$\tilde{A} = \begin{bmatrix} \mathrm{diag}(C_o) & \tilde{c}^{\mathrm{T}} \\ \tilde{c} & \mathrm{diag}(C_a') + C_n' E \end{bmatrix},$$

where $\tilde{c}_{ao} = c_{ao} - c_{no}$ and $E_{aa'} = 1$, restricted to $a, a' \neq n$, which takes into account that $b_n = -\sum_{a\neq n}b_a$.

Thus, the integral over $\tilde{h}$ is

$$\frac{1}{Z_O Z_A L w}(2\pi w)^{-\nu/2}e^{-R/2w}\frac{\prod\sqrt{c_{ao}}}{\sqrt{\det\tilde{A}}},$$

where $\nu = N - N_O - N_A + 1$.

Finally, we integrate over w to obtain

$$P(\mathrm{CWC}\mid s) = \frac{1}{Z_O Z_A L}(\pi R)^{-\nu/2}\Gamma\left(\frac{\nu}{2}\right)\frac{\prod\sqrt{c_{ao}}}{\sqrt{\det\tilde{A}}},$$

and the log-evidence is

$$\mathrm{LE}(\mathrm{CWC}\mid s) = \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\log\pi R + \frac{1}{2}\sum\log c_{ao}$$

$$- \frac{1}{2}\log\det\tilde{A} - \log Z_O - \log Z_A - \log L.$$

E.3. Incomplete block analysis

The model for IBA is the same as for CWC but taking the confidences $c_{ao} = 1$ for all the assessments. Thus, the log-evidence for IBA given the scores s is

$$\mathrm{LE}(\mathrm{IBA}\mid s) = \log\Gamma\left(\frac{\nu}{2}\right) - \frac{\nu}{2}\log\pi R - \frac{1}{2}\log\det\tilde{A} - \log Z_O - \log Z_A - \log L,$$

with the appropriate changes to R and $\tilde{A}$.

Appendix F. Potential refinements to the method

One could develop refinements to the basic model (2.4). For example, assessors might have not only an additive bias but also different scales, so, for example,

$$s_{ao} = \lambda_a v_o + b_a + \sigma_{ao}\eta_{ao}. \tag{F 1}$$

Fitting λ, v, b is more complicated, however, than just v, b.

An assessor may have a bias correlated with their confidence [23] or with some other feature like familiarity [24]. Assessors may like to give round-number scores or the organizers of the panel may insist

on them. Assessors may have different scales for confidence, so their confidences may need calibrating as well as their scores.

Another problem is that often assessors are asked to assign scores in a fixed range $[A, B]$, e.g. 1–10. Then any model for bias really ought to be nonlinear to respect the endpoints. One way to treat this is to apply a nonlinear transformation to map a slightly larger interval (a, b) onto $\mathbb{R}$, e.g.

$$x \mapsto s = \frac{x - (a+b)/2}{(b-x)(x-a)} \tag{F 2}$$

or

$$x \mapsto s = \log \frac{b - x}{x - a}, \tag{F 3}$$

apply our method to the transformed scores, scaling the confidences by the inverse square of the derivative of the transformation, and then apply the inverse transformation to the 'true' values. On the other hand, it may be inadvisable to specify a fixed range because it requires an assessor to have knowledge of the range of the objects before starting scoring. Thus, one could propose asking assessors to use any real numbers and then use equation (F 1) to extract true values v. A simpler strategy that might work nearly as well is to allow assessors to use any positive numbers but then to take logarithms and fit equation (2.4) to the log-scores. The assessor biases would then be like logarithms of exchange rates. The confidences would need translating appropriately too.

One issue with our method is that the effect of an assessor who assesses only one object is only to determine their own bias, apart from an overall shift along the null vector $(v, b) = (1, -1)$ for the rest. To rectify this, one could incorporate a prior probability distribution for the biases (indeed, this was done by Platt & Burges [8] in the form of a regularizer).

An interesting future project is to design the graph Γ optimally, given advance guesses of confidences and constraints (like conflicts of interest) or costs for the number of assessments per assessor. 'Optimality' would mean to achieve maximum precision or robustness of the resulting values. For instance, in each case of figure 1, each assessor has the same amount of work and each object receives the same amount of attention, but (a) achieves full connectivity with a resulting value for μ_2 of $1 - \sqrt{\frac{1}{3}} \approx 0.42$, whereas (b) achieves moderate connectivity and a smaller value of $\mu_2 = 1 - \sqrt{\frac{2}{3}} \approx 0.18$ and (a) is not even connected and has $\mu_2 = 0$.

References

1. Meadows M. 2006 *Can we predict who will be a reliable marker?* Manchester, UK: AQA Centre for Education Research and Policy.

2. Paul SR. 1981 Bayesian methods for calibration of examiners. *Br. J. Math. Stat. Psychol.* **34**, 213–223. (doi:10.1111/j.2044-8317.1981.tb00630.x)

3. Næs T, Brockhoff P, Tomic O. 2010 *Statistics for sensory and consumer science*. Chicester, UK: Wiley.

4. Fisher RA. 1940 An examination of the different possible solutions of a problem in incomplete blocks. *Ann. Eugenics* **10**, 52–75. (doi:10.1111/j.1469-1809.1940.tb02237.x)

5. Giesbrecht FG. 1986 Analysis of data from incomplete block designs. *Biometrics* **42**, 437–448. (doi:10.2307/2531064)

6. Flach PA, Spiegler S, Golenia B, Price S, Guiver J, Harbrich R, Graepel T, Zaki MJ. 2009 Novel tools to streamline the conference review process; experiences from SIGKDD'09. See http://research.microsoft.com/pubs/122784/ReviewerCalibration.pdf.

7. Guiver J. 2010 Calibrating reviews of conference submissions. See http://blogs.msdn.com/b/infernet_team_blog/archive2011/09/30/calibrating-reviews-of-conference-submissions.aspx.

8. Platt J, Burges C. 2012 Regularised least squares to remove reviewer bias. See

http://research.microsoft.com/en-us/um/people/cburges/papers/ReviewerBias.pdf.

9. Ge H, Welling M, Ghahramani Z. 2013 A Bayesian model for calibrating conference review scores. See http://mlg.eng.cam.ac.uk/hong/nipsrevcal.pdf.

10. Thorngate W, Dawes RM, Foddy M. 2008 *Judging merit*. New York, NY: Psychology Press.

11. Hubbard DW. 2014 *How to measure anything*, 3rd edn. New York, NY: Wiley.

12. Sattler DN, McKnight PE, Naney L, Mathis R. 2015 Grant peer review: improving inter-rater reliability with training. *PLoS ONE* **10**, e0130450. (doi:10.1371/journal.pone.0130450)

13. Meyer CD. 2001 *Matrix analysis and applied linear algebra*. Philadelphia, PA: SIAM.

14. Golub GH, van Loan CF. 1996 *Matrix computations*. Baltimore, MD: Johns Hopkins University Press.

15. Snell RR. 2015 Menage a Quoi? Optimal number of peer reviewers. *PLoS ONE* **10**, e0120838. (doi:10.1371/journal.pone.0120838)

16. Bland JM, Altman DG. 1986 Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* **327**, 307–310. (doi:10.1016/S0140-6736(86)90837-8)

17. Graves N, Barnett AG, Clarke P. 2011 Funding grant proposals for scientific research: retrospective

analysis of scores by members of grant review panel. *Br. Med. J.* **343**, d4797. (doi:10.1136/bmj.d4797)

18. Parker S. 2014 A test of a method for calibration of assessors. Final year undergraduate project report, University of Warwick.

19. Tukey JW. 1977 *Exploratory data analysis*. Reading, MA: Addison Wesley.

20. Kenna R, Berche B. 2011 Normalisation of research evaluation results across academic disciplines. *Res. Eval.* **20**, 107–116. (doi:10.3152/095820211X1294137 1876625)

21. Chung FRK. 1996 *Spectral graph theory*. Providence, RI: American Mathematical Society.

22. MacKay DJC. 2003 *Information theory, inference and learning algorithms*. Cambridge, UK: Cambridge University Press.

23. Song T, Wolfe EW, Hahn L, Less-Petersen M, Sanders R, Vickers D. 2014 Relationship between Rater Background and Rater Performance. See https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/schools/022_Song_RaterBackground_04_21_2014.pdf.

24. Fuchs D, Fuchs LS. 1986 Test procedure bias: a meta-analysis of examiner familiarity effects. *Rev. Educ. Res.* **56**, 243–262. (doi:10.3102/003465430 56002243)