We thank the reviewers for their very helpful feedback.

R1,R3,R4: Interpretability and Fig.2. **Answer:** Our text about interpretability was not clear enough and we will add the following: "In the canonical parametrisation, increasing the value of element a_{ij} in matrix $\mathbf{A}$ increases the calibrated probability of class i (and decreases the probabilities of all other classes), with the amount of change depending on the uncalibrated probability of class j. E.g., element $a_{2,8} = 0.63$ of Fig.2b increases class 2 probability whenever class 8 has high predicted probability, modifying decision boundaries and resulting in 26 less confusions of class 2 for 8 as seen in Fig.2c. Looking at the matrix $\mathbf{A}$ and vector $\mathbf{c}$, it is hard to know the effect of the calibration map without performing the computations. However, at $k+1$ *'interpretation points'* this is (approximately) possible. One of these is the centre of the probability simplex, which maps to $\mathbf{c}$. The other k points are vectors where one value is (almost) zero and the other values are equal, summing up to 1. Fig.2a shows the 3+1 interpretation points in an example for $k = 3$, where each arrow visualises the result of calibration (end of arrow) at a particular point (beginning of arrow). The result of calibration map at the interpretation points in the centres of sides (facets) is each determined by a single column of $\mathbf{A}$."

R4: "...Dirichlet calibration is very slightly different from matrix scaling..." **Answer:** While Dirichlet calibration can be viewed as a simple modification to matrix scaling (Dirichlet calibration is equivalent to the composition of log-transforming class probabilities and applying matrix scaling), we believe it is a novel modification as we are not aware of earlier works using log-transformed class probabilities in multi-class calibration (other than in the special case of binary calibration with beta calibration).

R4: "...why this is not a straightforward extension of beta calibration?" **Answer:** The derivation's starting point for Dirichlet calibration is indeed a straightforward generalization of the starting point of beta calibration [11]. Proof of our Theorem 1 is technically much more involved than the proofs in [11], but intuitively follows the same logic. The new challenge when moving from binary to multi-class was choosing the right regularization method (no regularization was applied in [11] for beta calibration).

R4: "I like the connection between Dirichlet calibration and matrix scaling. I'm just not sold on its significance. **Answer:** With this paper we have explored the family of calibration maps derived from Dirichlet distributions. The approach is however more general: e.g., it opens up the possibility of deriving further calibration maps from other members of the exponential family.

R4: "Explain why your method is significant even though it appears to have lower classwise-ECE than isotonic regression/vector scaling." **Answer:** Good calibration performance is easier to achieve with worse predictive performance (see also lines 108-113 of the main paper). Our methods improve predictive performance (e.g. log-loss) over the state-of-the-art, while retaining a comparable level of calibration. Based on critical difference diagrams, there are no significant differences in classwise-ECE between OvR_Isotonic and our Dirichlet_L2 (Supp.Fig.6f, non-neural) and between VecS and our MS-ODIR (Supp.Fig.11f, neural). Our methods have non-significantly worse classwise-ECE but better p-classwise-ECE (Supp.Figs.6h,11h).

R4: "Why is p-classwise-ECE chosen over classwise-ECE in the main paper?" **Answer:** These measures provide complementary information but due to space constraints we omitted classwise-ECE which does not show significant differences and we show it only in the supplementary.

R2: "This raises a general question of whether modeling pairwise class interactions is necessary or beneficial for multi-class calibration." **Answer:** MS-ODIR has lower log-loss than VecS in 13 out of 14 cases (Table 4) and this can only be due to non-zero off-diagonal values, indicating the importance of modelling pairwise class interactions. We apologize for the copy-paste error of average ranks in Table 4 (thanks to R2&R4 for noticing), the correct average ranks are $6.0, 3.5, 3.79, 2.93, 3.14, 1.64$ (as correctly shown in Supp.Tab.13).

R4: "...what are the effect sizes...?". **Answer:** In neural experiments, the effect size in predictive performance improvement of MS-ODIR over VecS measured by median relative reduction of loss (i.e. $(VecS - MS\_ODIR)/VecS$)) is 0.5% for error rate, 0.8% for log-loss and 0.7% for Brier score (calculated from Supp.Tabs.18,13,14). In non-neural experiments, the median relative loss reduction of Dir_L2 over OvR_Isotonic (i.e. $(Iso - Dir)/Iso$)) is 1.2% for error rate, 13.8% for log-loss and 1.0% for Brier score.

R3: "...I'm not sure I would want to switch to this new method." **Answer:** We suggest to switch to Dirichlet_L2 (non-neural) or MS-ODIR (neural), as they improve predictive performance and are on par with the state-of-the-art in calibration. These models are unique in modelling directly pairwise class interactions, and are likely to be particularly valuable when recalibrating after any dataset shift.

We thank **R4** for the suggestion to study which methods are better at calibrating which models (remains as future work) and for the suggested articles. We will cite Nixon et al. (6 weeks before us, independently, used SCE metric equivalent to our classwise-ECE), Kull et al. (earlier version of beta calibration [11]), Kuleshov and Liang (used one-vs-rest multi-class calibration), Hosmer and Lemeshov (studied model fitness in general, which partially inspects calibration).