We thank all the reviewers for their thorough reviews. All reviews expressed that there were “valuable contributions" in the paper, and R3 and R4 said the work was of “high practical significance" and “original, useful and clearly presented". The reviewers also had many constructive suggestions and questions that we will address below.

R1, scaling vs ours: R1’s first major question was if, due to Theorem 4.1, one should “use the scaling method for calibration" and use our method “as a surrogate in order to check if the (scaling method’s) error is smaller than some threshold". We cannot do this because the calibration error of the scaling method can be much higher than for our method. Theorem 4.1 says that our method is *at least* “almost as well-calibrated as the best possible recalibrator" in “G after a certain number of samples", but it could be much better calibrated, as in Example 3.2. This is because the binned version of a function has lower calibration error than the original function (Proposition 3.3, used in line 206 of Theorem 4.1 proof sketch). This is a fundamental issue with scaling methods—binning only lower bounds their calibration error.

R1, additional experiment: As suggested, we ran synthetic experiments to compare our calibrator with the underlying scaling method. The ground truth $P(Y = 1|z)$ is from the Platt scaling family G but with noise. Varying n, we compute 90% confidence intervals from 1000 trials. With 10 bins, $n = 3000$ the ℓ_2^2 calibration error is 5.2 ± 1.1 times lower for our method than the scaling method—our method does even better for larger n. And unlike scaling methods, our method has measurable calibration error—if we are not calibrated we can get more data or use a different scaling family.

R1, proof of Theorem 4.1: $R1$ mentioned that to use Lemma D.1, we would in fact need $1/\epsilon^4$ points to achieve ℓ_2^2-CE $\leq 2\epsilon^2$. We made a mistake (thanks for catching it), but it can be easily repaired as follows. In lemma D.1, we can actually get a convergence rate of $1/n$ instead of $1/\sqrt{n}$ for the MSE, using standard asymptotic results of M-estimators (under regularity conditions). Technical details: The asymptotic result gives a *parameter* convergence rate of $1/\sqrt{n}$ which leads to a $1/n$ convergence rate in the MSE *loss*. We have updated Lemma D.1, which fixes line 575. We have fixed the theorem statement (see below) and applications of the lemmas to clarify they are probabilistic. We also implemented synthetic experiments to sanity check these bounds (see ‘R2, validating bounds’).

R1, other concerns: We agree with all of R1’s detailed comments and will fix them (for example we have toned down line 214 to say “we showed that *current techniques* cannot accurately measure the calibration error of scaling methods").

R2 had 2 main concerns. 1. **Well-balanced binning**: R2 was concerned that our framework requires well-balanced binning to hold, which may not hold on real data. We believe this is a misunderstanding—in step 2 of our algorithm we *choose* bins so that an equal number of calibration points land in each bin. We then *prove* (instead of require) that the well-balanced property holds in the population (Lemma 4.3). 2. **Binary vs Multi-class**: Our theory generalizes to the multiclass setting. Top-label calibration is a binary calibration problem (lines 103 - 104) where $Z \in [0, 1]$ is the model’s confidence for the top class, and Y is 1 if the model’s prediction was correct, and 0 otherwise. Marginal calibration requires each class to be independently calibrated, which transforms into K binary calibration problems where K is the number of classes. **Notational issues and typos**: We thank the reviewer for identifying them, and will fix these.

R2, validating bounds: As R2 suggested, we added synthetic experiments to validate the bound in Theorem 4.1. Our theory predicts that $n \lesssim 1/\epsilon^2 + B$ for our method but for histogram binning $n \lesssim B/\epsilon^2$. In the first experiment, we fix B and vary n—we see that $1/\epsilon^2$ is approximately linear in n for both calibrators. In the second experiment, we fix n and vary B—as predicted by the theory, for our variance-reduced calibrator $1/\epsilon^2$ is nearly constant, but for histogram binning $1/\epsilon^2$ scales close to $1/B$. When we increase from 5 to 20 bins, our method’s ℓ_2^2-CE decreased by $2\% \pm 7\%$ but for histogram binning it increased by 3.71 ± 0.15 *times*—we will include details and plots in the paper.

R3 had a number of useful suggestions and questions. They mentioned that the use of big-O in Theorem 4.1 was confusing—we have rephrased the theorem as shown below. We agree that the DEMOGEN dataset is a good resource to tap into for a more extensive analysis of calibration—we will mention this as potential future work and cite the dataset/paper. Regarding line 178: histogram binning bins the Y values, but not the outputs of a *recalibrator function*. We will address R3’s other suggestions (e.g. connection with scoring rules) in the next revision.

Theorem 4.1: Assume regularity conditions on $\mathcal{G}$ (Lipschitz, injective, and all conditions in Theorem 5.23 in Asymptotic Statistics, Vaart, A.) Given $\delta \in (0, 1)$, there is a constant c such that *for all* B, ϵ, with $n \geq c\Big(B \log B + \frac{\log B}{\epsilon^2}\Big)$ samples, the variance-reduced algorithm finds $\hat{g}_{\mathcal{B}}$ with ℓ_2^2-CE$(\hat{g}_{\mathcal{B}}) \leq \min_{g \in \mathcal{G}} \ell_2^2$-CE$(g) + \epsilon^2$, with probability $\geq 1 - \delta$.

R4 had a good suggestion—checking if the debiased estimator was less sensitive to the number of bins when used for scaling methods. We repeated the experiment in Section 3.1, but observed similar results, which we will add to the paper. Regarding motivation, we believe practitioners use calibration in many ways although [5] (Gneiting and Raftery, Science 2005) propose maximizing sharpness subject to a calibration error budget. We believe practitioners implicitly do this—calibration/reliability is an important diagnostic metric and when it is unsatisfactory the forecast and its granularity are changed. R4 mentioned that our method is less efficient than scaling methods—note that our method typically has lower (better) calibration error than the scaling method, as binning decreases the calibration error, e.g. see “R1, additional experiment”. We thank R4 for the detailed comments and will fix the issues identified.