We are very happy to hear that all reviewers found the paper interesting and well written. We thank the reviewers for all the constructive feedback on our paper and for all the suggestions for future work. We will naturally take all comments into account when revising the paper. Below we provide point-by-point responses to selected comments.

All reviewers commented on that it would have been interesting to see results on real-world datasets.

We intentionally did not include any evaluations on benchmark datasets such as CIFAR-10 or ImageNet, since we wanted to focus on an experimental confirmation of the derived theoretical properties of the kernel-based estimators and the illustration of its advantages over the commonly used expected calibration error (ECE). In contrast to Guo et al., we did not want to make any claims about the calibration of different neural network architectures, nor the re-calibration of uncalibrated models.

As pointed out by Reviewer #3, the calibration measures that we consider only depend on the predictions (and true labels), not on how these predictions are computed. We therefore believe that directly specifying the predictions in a "controlled way" results in a cleaner and more informative numerical evaluation.

That being said, we recognize that his approach might have resulted in an unnecessary disconnect between the results of the paper and a practical use case. We have therefore conducted additional evaluations with different neural networks such as DenseNet, ResNet, GoogleNet, Inception, and VGGNet trained on the CIFAR-10 dataset, using the same binning and kernel choices as in our submission. As we have argued in our submission, the raw calibration estimates are not interpretable and can be misleading, and hence we have only considered the proposed approximations of the probability of falsely rejecting a calibrated model. The distribution-free bounds are typically between 0.99 and 1, indicating again their weakness, whereas the bounds based on the asymptotic distribution of linear unbiased SKCE yield values between 0.09 (for ResNet-34) and 0.91 (for GoogleNet). On the other hand, the approximations obtained by consistency resampling, both with uniform and data-dependent binning, are almost always 0 (apart for GoogleNet with uniform binning). It seems the quadratic bound with (only) 100 bootstrap samples for the asymptotic distribution yields also 0 for all models. We will add these additional experimental results to the revised supplementary material.

Reviewer #1

"it could be useful to study the impact of the chosen kernel and its hyperparameters"

Indeed, the impact of the kernel and its hyperparameters on the estimators and, in particular, on the bounds of the type I error is an important research question. Apart from the choice of the kernel bandwidth, we had refrained from discussing it in our initial submission. In our opinion, this question deserves a more exhaustive study than, what we felt, would have been possible in this work. We will state the need and importance for future research in this area more clearly.

Reviewer #2

"Consider reporting computational time for the proposed estimators and experimental setup details"

We agree that the computational time, although dependent on our Julia implementation and the hardware used, might provide some insights to the interested reader in addition to the algorithmic complexity. However, in our opinion, a fair comparison of the suggested estimators and p-value approximations should take into account the error of the methods as well, similar to work precision diagrams for numerical differential equation solvers. In our case, one could quantify the bias and variance of the estimators and the p-value approximations. A simple comparison of the methods (with the same kernel and binning choices as in our submission) for fixed numbers of classes and varying number of samples has shown the expected scaling of the computation time with increasing number of samples and revealed that even for 1000 samples and 1000 classes the biased SKCE and the quadratic unbiased SKCE can be evaluated in around 0.1 sec. As a comparison, for this setting, the evaluation of the linear unbiased SKCE takes around 10^{-4} sec, of the ECE with bins of uniform width around 10^{-3} sec, and of the ECE with data-dependent bins around 0.1 sec.

Reviewer #3

"I found l.213-216 a bit too mysterious . . . Is it not possible for other estimators to estimate (3). Does this method allow (3) to become tractable, or are you simply observing that one should keep in mind that test statistics come from (3). "

We simply mean that the statistical tests are designed to test if (3) does not hold. Other estimators are indeed possible, but establishing deviation inequalities such as our Lemmas 2 and 3 has to be done on a case-by-case basis, and this could very well prove to be difficult in many cases.

References

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR, 08 2017.