We thank the reviewers for their feedback. Below we address specific comments.

All reviewers: We emphasize that our paper examines ways to *evaluate* distributions produced by *any learning algorithm*. While it can be used for designing specific learning algorithms in future, that is not our primary objective.

Reviewer #2: The reviewer asks for a clarification of the main message. We answer this question in 3 parts.

- **Importance of the properties we explored:** The importance of properness has been well established in the literature and is a major motivation behind the popularity of log loss minimization. As we mentioned, if the loss is not proper then even if the learner receives infinitely many samples the loss minimization recovers a wrong distribution. Sample properness is a natural extension of properness to finite samples. If a loss function is not sample-proper, then loss minimization on no reasonable amount of samples would lead to the right distribution. As we mentioned, sample properness is essential for distribution learning and has been studied in previous work for the specific case of log-loss. Concentration is also a natural property of losses, without it the estimated loss of a candidate distribution from samples is not a good proxy for its actual loss, and therefore, cannot be used for the purpose of distribution evaluation.
- **Prevalence of loss minimization for distribution learning:** A generative model is one that generates samples from some underlying distribution. Thus in our context, a generative model is exactly a candidate distribution $\mathbf{q}$ which is meant to estimate the target distribution $\mathbf{p}$. There are many methods to train and evaluate generative models, the most popular of which is the average log likelihood that the model assigns to a sample set. For example, in GANs (Goodfellow et al. 2014) the focus is on average likelihood maximization, *which is equivalent to log loss minimization.* Some of the other methods such as Jensen-Shannon divergence, contrastive divergence, etc. are also closely related to log loss minimization.
- **Clarification of main contributions:** The reviewer mentions that we study log-loss with respect to the above properties. While we do mention log-loss in this way, our main contribution is to show that many functions in addition to log-loss meet these properties when we consider calibrated distributions. On the other hand, no function (not even log-loss) possess these properties without calibration. As we showed in Figure 1, in some applications log loss is not the optimal loss function to be used. By putting forward alternative loss functions with desirable properties, our work paves the way for picking loss functions based on the domain's need.

Reviewer #3:

- **Concentration Results:** We disagree that our concentration results are not surprising given the folklore theorem. As we mention in the paper even the log loss (while being sample proper by the folklore theorem) *does not concentrate* without a calibration assumption. That is, our concentration result indeed has to leverage the structural properties of calibration in addition to the inverse concavity of the loss.
- **Definition of $\hat{\mathbf{p}}$:** As mentioned $\hat{\mathbf{p}}$ denotes the empirical distribution. As is standard, the probability of an element in $\hat{\mathbf{p}}$ is its relative frequency in the sample. Elements in the domain and not in the sample are assigned probability 0.
- **Efficient implementation:** We agree that efficient algorithms are an important direction for future work. We included some preliminary results in the supplemental (see Section 5) in this regard. Our main goal in this paper however is to lay out a formal foundation for evaluating distributions via desirable loss functions that future algorithms can rely on.

Reviewer #4: Thank you for your helpful and detailed comments, which we will implement in the final version.

- **Example**: We agree with this and have corrected the example. At a high level elements $3, \dots, N$ of $\mathbf{q}$ and $\mathbf{p}$ should have been switched, i.e., $q_{3:N} = 1/2(N-2)$. This fixes the calibration issue. As for sample-properness, note that the contribution of elements $x = 3 \dots, N$ is at most $\Theta(1/N)$ to the equation in line 218. With a constant probability $\hat{p}_1 \leq \frac{1}{4} - \frac{1}{\sqrt{m}}$ and $\hat{p}_2 \geq \frac{1}{4} + \frac{1}{\sqrt{m}}$. That is, $\ell(\mathbf{q}; \hat{\mathbf{p}}) - \ell(\mathbf{p}; \hat{\mathbf{p}}) \leq -1/m + \Theta(1/N) < 0$ for $m \in o(N)$. It is possible to strengthen this bound with a more careful analysis of the contribution of elements $3, \dots, N$, but the main message of this part remains the same regardless, that is, *linear-loss is not sample-proper.*
- **Is calibration a natural assumption:** Calibration has been used in a long line of work (including [11, 13]) as a natural requirement for probabilistic predictions. We consider a major contribution of our work to be in understanding how the classic notion of calibration relates to loss functions for evaluating/learning distributions, and in particular in showing that this restriction circumvents the impossibility result for satisfying the three natural criteria we considered.
- **On"for all distributions" results:** We agree that this is an interesting direction. One could strengthen Thms. 3 & 4 by taking a union bound over a finite net over all distributions. Thus, our results directly apply to this setting albeit with worsened sample complexity. We note that this worsening of sample complexity is needed, because the "for all distribution" version of our result would enable distribution learning in the worst-case, which is known to require $\Omega(N)$ samples. We note that there are methods that can perform ERM over a full class of distributions when the target distribution is not a worst-case distribution. Many such methods (e.g., GANs) perform an implicit optimization, and then are evaluated using various losses. In this case, the guarantees given in Theorem 3 & 4 are directly useful – they show that a small number of samples suffices to compare a finite set of outputs from various algorithms or parameter settings accurately.