We thank the reviewers for their thoughtful feedback. We were pleased to see that several reviewers agree with our proposal to use distributional robustness to evaluate theories of generalization. Several reviewers also acknowledge the “significant computational effort” behind our empirical study and the “surprising findings” we uncover by going beyond average case.

R1 I cannot [...] see why [...] robust error is a better empirical quantity to look at. We argue that a theory that only works well on average does not explain generalization. As was presented in Sec. 1 and eloquently summarized by **R3**, a theory must “fit what we observe in nature". Because a theory is only as strong as its weakest link, we propose to use the framework of distributional robustness. In Sec. 6, we show that a focus on worst-case performance allows us to identify failure modes of existing generalization measures that would not have been revealed by an average-case analysis. Each of **R2**, **R3**, **R4** touch on the important difference between average and worst-case analysis here.

R3 Relationship to the Inductive Causation approach in Jiang et al. (2020) It is important to clarify key differences between the two approaches. Jiang et al. calculate the normalized conditional mutual information (Eq. (8) in [6]) and take the min over all conditioning sets of two hyperparameters (Eq. (9) in [6]). This is not equivalent to a min over all possible interventions: interventions correspond to setting the variables in the conditioning set to a specific value and their Eq. (6) and (7) take an *average over all values*. This is in sharp contrast with our method, which does measure robustness over all possible interventions. Furthermore, in the limit where we observe every possible intervention on the HPs, our method admits a causal explanation (see Sec 1.1). However, this is not necessarily true for the IC-based method of Jiang et al. The reason is that their conditioning sets are of size 2, which may leave open paths in the graph if more than 2 HPs act as confounders, resulting in non-causal mutual information. This issue cannot be avoided, since their measure collapses to 0 when all HPs are conditioned upon. Finally, our method also allows us to account for Monte Carlo noise in pairwise comparisons and avoids making graphical assumptions about variable relationships.

R1 R2 R3 R4 Choice of hyperparameter ranges and over/under-parametrization We agree that studying other datasets, more HPs, underparametrized networks, etc. is interesting. Yet, we feel the scope of our experiments provides sufficient evidence for the merits of our philosophical arguments and methodology in support of robust analysis.

R4 While the theoretical motivation behind using $\sup$ is clear, in practice it makes the evaluation criterion less robust to the coarseness of the underlying set of environments (e.g., 2 values for width → results different). Average-case analysis seems “more robust” in this way, but this is a deceptively benign property. If the boundaries of a robust analysis change, the conclusions may change because the *implied scope of the theory has changed*. In our analysis and supplement, we argue that one should dig into failures to localize them (see, e.g., Fig. 2 and App. D).

AC Spectral norm approximation method In short, thank you: we now use the same method as Jiang et al. (2020). In detail: Approximating the layer spectral norm $||\text{Conv}||_2$ with the reshaped filter spectral norm $||W'||_2$ (where W' is of shape $(c_{out} \times (c_{in} \times h \times w))$) was proposed by Yoshida and Miyato (2017; Sec. 3.2) and used for empirically stabilizing GAN training in Miyato et al. (2018). This method was also used in Neyshabur et al. (2017) and was favourably benchmarked in Sedghi et al. 2018. That said, estimates (such as by Tsuzuku et al. (2018; Cor. 1)) are too loose for comfort and so we now follow Jiang et al. (2020), who use the method of Sedghi et al. (2018), which exactly calculates the spectral norm of a convolutional layer. We have modified our code to use this method instead, and observe that the approximation was fairly accurate and our analysis does not change due to this modification. We will provide a comparison between the approximation and the exact method in the updated supplementary. Hurrah for peer review.

R3 It’s not clear to me how the single-network experiment relates to the distribution robustness. Generalization measures are expected to predict the exact numerical value of generalization error up to a constant rescaling factor and additive term. We apply robust regression (see Ben-Tal et al. (2009)) to learn constants that hold in all environments.

AC Give intuition for measures We agree that this is important and will update the paper accordingly.

R1 R2 R3 Extended discussion of findings and comparison to Jiang et al. We want to reiterate that the average-case results that we report are, up to the particular HP value choices, equivalent to the $\mathbf{\Psi}$ results reported in Jiang et al. [6]. We can expand on how our reproduction of their average-case study agrees with theirs in the additional page available for the camera-ready.

R2 Did the authors consider using the approach to formulate robust generalization measures? Absolutely, this is an interesting future work that we will mention in the camera-ready version.

R1 Reproducibility: public release of all the trained models, codes, and detailed setups We believe there is a misunderstanding: all experimental details—code, hyperparameter ranges, and data—were provided at submission (see App. C). We also plan to make a public release of the data and trained models upon acceptance.

R1 What exactly do you mean by the “test-set bound"? Apologies. We simply meant an estimate of risk using held-out data (e.g., a test set), using a simple average. We will address this and other clarity issues you raised.