We thank all three reviewers for their very thoughtful and detailed comments, with which we largely agree. Some requested additions (e.g., details about experiments and interventional datasets, an example of an empirical evaluation) were present in previous versions of the paper and were removed for space. We will add these sections to any revised version, either in the main body of the paper or the Supplemental Material. More details appear below.

Additional citations (R1) We thank Reviewer 1 for the highly relevant pointers to the statistics literature—these will be cited and discussed in any revised version. Within-study comparisons provide a great source of data that is both experimental and observational and should be discussed as a source of interventional data.

Originality (R1) Some of the points discussed in this paper have been touched on in previous work in statistics and quantitative social science (e.g., the citations mentioned by Reviewer 1). However, this work is not widely known in the machine learning community (the primary audience for NeurIPS), and this work often concerns a highly specific evaluation context (i.e., single-treatment/single-outcome, average treatment-effect, etc.). As seen by our survey, only a small fraction of the ML community performs evaluations that are both empirical and interventional, and there are very few standard data sets in the field that allow for this. We believe that further evidence, targeted at researchers within computer science and addressing the approaches and concerns of that community, is necessary to promote the wider adoption of more principled evaluation methods.

Evaluation guidance (R1) Reviewer 1 notes that the paper provides little guidance on how to perform a principled empirical evaluation. An example of an empirical evaluation was present in a previous version of the paper, but it was cut for space. It is clear that, for the type of evaluation we advocate to be better understood and adopted, more description and guidance is needed. We will include a detailed example in any revised version.

Interventional vs Structural measures on the correct structure (R2) Reviewer 2 is correct that, for a correctly learned structure, as long as the parameterization is correct, interventional and structural measures should agree. Interventional measures are valuable when the learned structure is only approximately correct, and empirical evidence indicates that this is almost always the case. We will add more discussion of this to any revised version to make this clearer. Some related experiments were moved for space to the Supplemental Material, showing how structural Hamming distance and structural intervention distance penalize over- and under-specification differently than total variation distance.

Untested influences (R2) The critique of untested influences refers to potential "unknown unknowns" in the data-generating system. In many real-world systems, even with strong domain knowledge, there often exist factors that are outside the researcher's knowledge. This is generally not possible in a synthetic system. While latent variables can be added, they are still defined and created by the researcher, limiting the realism of the data. We will clarify this in any revised version.

Number of samples in synthetic data experiments (R2) The number of samples was chosen to match the number of samples available from the software system. This is 2599 for networking, 473 for jdk, and 5000 for postgres. We acknowledge there may be issues with low sample sizes and will discuss this in any revised version.

Structural measures implicity assume a DAG (R2) Our intent with this statement was that some structural measures can only be used by algorithms that produce a DAG as output. There are many ways this assumption could be violated. Acyclicity is one such violation, which could hinder the use of certain structural measures in evaluation of an algorithm that outputs a cyclic graph. Another possible violation is an algorithm that does not output a graphical model, such as a recent work on learning probabilistic programs. We will clarify this intent in any future version.

Total variation distance calculation (R2) In Figure 2, TVD is calculated as the distance between the learned distribution and the true distribution, just as SHD is calculated as the distance between the learned structure and the true structure. Here, the true distribution is known because the data is synthetically generated, although such true distributions can also be obtained from interventional experiments as they are elsewhere in the paper.

Domain-specific simulations (R3) Domain-specific simulation systems are a very useful approach to consider, and we will add a discussion of them in any future version of the paper. We already mention one such system (the DREAM in-silico challenges), but we will add a far more general discussion of this approach. A sufficiently sophisticated simulation falls on a spectrum between purely synthetic and purely empirical data, and such simulations are valuable because they are often highly complex, are created by someone other than the researcher, and are created for a purpose other than evaluation.

Capabilities and limitations of current interventional datasets (R3) We agree that a more detailed discussion of the currently available interventional datasets would be highly beneficial. This sort of discussion was omitted for space but will be included in at least the Supplemental Material of any revised version.