We thank all the reviewers for their time and effort.

[Appropriateness for NeurIPS] As pointed out by Reviewer 1, reviewing practices are of interest to the NeurIPS community. Moreover, *our work is applicable beyond peer review as well, in applications such as admission or hiring*, where the evaluation assignments are not random and where one wishes to avoid excessive evaluations for the testing. We note that importantly, *past NeurIPS editions feature a number of works on statistical testing* (e.g., "Hypothesis Testing in Unsupervised Domain Adaptation with Applications in Alzheimer's Disease" (2016) and "Differentially Private Uniformly Most Powerful Tests for Binomial Data" (2018) to just name a few due to space constraints here), and we believe our work on statistical hypothesis testing aligns well with the scope of NeurIPS.

[Importance of the problem] It is extremely important to note that the claim "This review format is furthermore not widely adopted anyway" does not hold throughout most of academia. Even in computer science, most Theory conferences use single blind. There are massive ongoing debates on SB vs. DB in many fields of academia (including in fields such as databases which have switched to DB, but there is some push to move back to SB). Based on our experience in this debate, those supporting SB also argue for evidence of biases *in their specific community*. Indeed, an important roadblock in making the SB vs. DB arguments is the absence of rigorous evidence – and this is where our work significantly contributes by designing principled procedures to answer the pressing questions in these debates.

[Effect sizes] We thank Reviewer 1 for raising an important point of effect sizes. In fact, we should have mentioned that **the test statistic of our test represents the effect size**. This is in line with the seminal work by J. Cohen ("A power primer", 1992.) where the test statistic is suggested to estimate the effect size for the sign test. We are more than happy to add more in-depth discussion of the effect size in the revision.

[GLMs and other models] Our test makes significantly fewer assumptions than the tests based on GLMs and other parametric models. Specifically, our test *does not rely on strong modelling assumptions* (in contrast to GLMs) and *also holds when reviewer decisions can be completely subjective* (in contrast to models that assume a presence of "true" [latent] qualities). Indeed, the models suggested by the reviewers are at a risk of making spurious conclusions because of the restrictive natures of the suggested models which may not capture the highly complex decision-making process of the human reviewers.

[Assignment and matching] During the review period, we **have solved the open problem of designing the experimental procedure** that leads to a reliable testing. Specifically, we have designed the experimental procedure (allocation of reviewers to conditions, assignment and matching) that follows standard conference peer-review pipeline (i.e., allows to perform any assignment algorithm, including TPMS) and *does not inflate Type-I error rates of the testing procedures*. Our proposed procedure assigns reviewers to conditions and papers to reviewers jointly in a carefully selected manner, thereby avoiding issues pertinent to the setup of Tomkins et al. All the statements in the paper hold for this procedure, but do not require an idealized random assignment any more. If reviewers prefer, we can include it in the final version.

[Previous works] Conference peer review setup is not a fully randomized controlled trial (i.e., the reviewers are not assigned at random) and hence *past approaches fail due to idiosyncrasies of the process*. With respect to the specific work mentioned by Reviewer 1 (Bertrand & Mullainathan, 2004), their method assigns identities of authors to (fabricated) documents at random. In our setup, *random assignment of author identities to real (i.e., non-fabricated) submissions* is problematic due to various logistical and ethical issues including reviewers guessing actual authors thereby causing biases, not all authors/researchers agreeing to have their paper/name modified, and others — this opens a separate can of worms which should be rigorously addressed before using it in the peer review setting. We are definitely happy to add a discussion of this and other relevant papers (including those mentioned by Reviewer 3) in the revision.

[Permutation test] The standard way of performing the permutation test would fail to control for the Type-I error because of the additional confounding due to quality of submissions. Our test is a careful modification of the permutation test which provably controls for the Type-I error rate even in presence of such confoundings. In the final version, we are happy to detail the shortcomings of the standard permutation test in our setup.

[Real data experiment] Unfortunately, the data from the Tomkins et al. experiment is not available to us. Tomkins et al. mention in their work that releasing this data (even in an anonymized format) would make it possible to deanonymize reviewers. Through our developed toolkit, we are happy to assist any program chairs who are interested to conduct tests of biases in their respective research community.