We thank all reviewers for their constructive comments. All typographic errors pointed out will be corrected accordingly. Recall that $\mathcal{M} = \{P_i\}_{i=1}^{l}$ is the set of l candidate models, R is the unknown data generating distribution, D is the discrepancy measure (MMD or KSD), $J \in \arg\max_i D(P_i, R)$, and $\hat{J} \in \arg\max_i \hat{D}(P_i, R)$ (see L138 for details).

Rev 1, 3, 7: Why compare more than two models? Model comparison beyond two models is a more realistic scenario. Given the availability of possible solutions, e.g. zoo of GANs, it is unlikely that a practitioner will only consider two candidate models for a task. A popular approach is to rank candidate models by a fitness score (e.g., FID). These estimated scores are correlated since they are computed on the same set of observations. Simply ranking these scores without accounting for the randomness and correlation leads to uncontrolled false positive rate (FPR) e.g. Table 1 below, B (true best) is not selected $1 - 83\% = 17\%$ of the time and **@[Rev 7]** for Experiment 3 (CelebA), Model 4 (the "best") is not selected $1 - 63\% = 37\%$ of the time. By contrast, the two proposed tests (RelPSI and RelMulti) have a well controlled FPR and false detection rate (FDR), respectively, as noted by Rev 7. We will add to our manuscript.

Rev 3, 7: Multiple goodness-of-fit testing vs multiple model comparison. The two questions are fundamentally different. In multiple goodness-of-fit testing, the goal is to determine whether R (observed through samples) is in $\mathcal{M}$ i.e., find $P^* \in \mathcal{M}$ such that $D(P^*, R) = 0$. A PSI-based multiple goodness-of-fit test has been considered in [31] for several candidate GAN models. Since all models are wrong (Box. 1976), it has led to the trivial result of the rejection of all candidate models [31, section 5.3]. In multiple model comparison (our work), the goal is to find the model(s) which has the lowest (not necessarily zero) discrepancy to R i.e., find $P^* \in \arg\min_{P \in \mathcal{M}} D(P, R)$ with statistical significance. While the former may be addressed by reducing it to l individual goodness-of-fit tests (one for each candidate), the latter problem is more complicated since *finding P^* requires comparing l correlated estimates of D.*

All reviewers: **Why not use the previous relative tests?** The relative model comparison tests RelMMD, RelKSD (for $l = 2$) considered in [4, 17] require the practitioner to choose the ordering of models; that is, one has to decide a priori H_0: $D(P_1, R) \leq D(P_2, R)$ or H_0: $D(P_2, R) \leq D(P_1, R)$. It is not obvious how one would use these relative tests to find the best model(s) when $l > 2$. On the other hand, our proposed tests automatically determine the index $\hat{J}$ of the best model, and take into account the fact that the data used to find $\hat{J}$ are the same as the data used for testing each model against $P_{\hat{J}}$, creating the conditional null hypothesis (see L152, L169). This is the complication that did not exist in the previous relative tests, and is the crux of our proposal. **@[Rev 1] L179-180**, **@[Rev 7] L168**: The conditional H_0 reduces to the standard unconditional H_0 if the data used to find $\hat{J}$ are independent of the test data (i.e., conditioning on an independent event $\boldsymbol{Az}$). The independence can be achieved by data splitting, which is the basis of the proposed RelMulti.

Rev 2: L1, positive and negative. We follow the convention that when a test declares a significant result, it is positive. Thus model P_i is assigned positive when our test declares that it is worse than the best model (i.e., reject the null hypothesis). **L95, Mild conditions:** If $\mathbb{E}_p[k(x, x)] < \infty$, then the mean embedding μ_p exists [26]. In particular, if k is bounded (e.g., IMQ kernel, Gaussian kernel), μ_p always exists. **L231, TPR and random sample:** TPR is defined (in Appendix A) as the **population expectation** of the proportion of number of true positive models that are declared as positive, and is not random. **L234, definition of μ:** We define $\mu := D(P_1, R) - D(P_2, R)$. The discrepancy measure D can be MMD or KSD. **L271, H_0 is true:** In Experiment 1, $H_0 : D(P_1, R) \leq D(P_2, R)$ holds since $D(P_1, R) = D(P_2, R)$ (there is a typo on L267). **L288, sampling variability:** For all our experiments, we averaged the results over at least 100 trials (for Fig 1, it was 300 trials), with new samples redrawn in each trial.

Rev 3: Selected reference is not the best. It is true that the selection is noisy and we can pick a worse model than the actual best, i.e. $\hat{J} \neq J$ (assuming the best is unique). In this case, "$H_{0,i}^{\hat{J}} : D(P_{\hat{J}}, R) \geq D(P_i, R) \mid P_{\hat{J}}$ selected" will hold for a larger portion of the tests, and will only result in lower TPR. In particular, FPR is not affected. See Table 1. We emphasize that our **theoretical results** do not make an assumption that the reference is correctly selected. It is accounted for in TPR/FPR calculations and an incorrect rejection is made with probability no larger than α. **ME, FSSD, SCF:** We will provide a unified statement in the revised version. "**Complete**" refers to the complete U-statistic estimator and "**linear**" refers to the linear time estimator of [14, Section 6].

$P_{\hat{J}} =$	A	B	C
Sel	16%	83%	1%
CTPR	.115	.271	.009
CFPR	.010	0	0
A	0	.065	.017
B	.010	0	0
C	.229	.477	0

Table 1: Results of the toy experiment of Rev 3 using RelPSI-MMD. Results averaged over 5000 trials with sample size 100. Sel is the proportion of times a particular $\hat{J}$ is selected. CTPR (and CFPR) is the empirical TPR (and FPR) conditioned on the selected $\hat{J}$. The bottom half shows rejection rates of each model for different $P_{\hat{J}}$. $\alpha = 0.05$. We estimate FPR $= 0.001$ and TPR $= 0.2428$.

Rev 7: Gaussian kernel with KSD. [13] shows that if the KSD with a Gaussian kernel is used to measure the discrepancy between a collection of n points X_n from a non-convergent MCMC and a distribution p, then vanishing KSD does not imply that $X_n \sim p$. This is an issue only when X_n does not follow any distribution at all. It is irrelevant for goodness-of-fit/model comparison testing since X_n is assumed to follow a proper distribution. The KSD goodness-of-fit test will detect any discrepancy asymptotically (see Proposition 4.2 of [23]).