We thank the reviewers for their positive and constructive feedback. We will address all minor suggestions in the final version of the paper. We now reply to the main points raised by the individual reviewers.

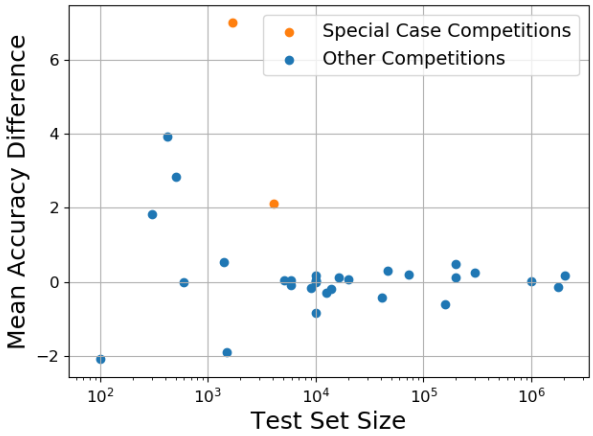
Reviewer 1

Point 1: We agree with the overall conclusion of the reviewer, with the exception that the non-uniform p-values in Section 3.3 are only limited evidence for adaptive overfitting due to the strong null hypothesis. Nevertheless, some of the plots in Section 3.2 also indicate small differences between public and private accuracy scores that are indicative of mild overfitting. We will update our discussion in Section 3.5 to more clearly reflect this.

Point 2: We will update our figures with alpha values for the scatter plots. Figure 1 does show confidence intervals, but they are small due to the scale of the axes and the test set sizes. We will clarify this in the figure captions.

Point 3: Figure 4 of our submitted paper already points towards one covariate that is correlated with adaptive overfitting: in-class vs. other competitions (mainly featured competitions). This indicates that quality control performed by Kaggle for the featured competitions may avoid certain causes of public vs. private accuracy differences.

Moreover, we will include the figure on the right in the final version of our paper. The figure shows the relationship between overfitting and test set size for accuracy competitions. It also points towards 10,000 examples as a possible recommendation for minimum test set size. While a reliable recommendation for applied machine learning will require a broader investigation that goes beyond the scope of our current paper, we believe that our candidate hypothesis can still inform future work on adaptive overfitting.



Point 4: While we agree that ranking submissions is an important aspect of the Kaggle platform, we believe that score differences are more fundamental from the perspective of adaptive overfitting. If many models with essentially the same performance are submitted to a competition, small variations can lead to large rank changes. However, these rank changes are not problematic if the performance of each method stays the same up to statistical fluctuations. Hence we view ranking mainly as an artifact of the competition format and not inherently important for safely deploying machine learning in real applications. Having said that, we have compared our results to those of the "Meta Kaggle: Competition Shake-up" notebook on Kaggle and found similar conclusions when considering changes in ranking. We will comment on this in the final version of our paper.

Reviewer 2

Regarding the relationship between competition type or test set size and overfitting, please see Point 3 in our response to Reviewer 1.

We certainly plan to investigate further competition platforms and machine learning tasks (including regression). In fact, we have already contacted 20 individual ML competitions hosted outside Kaggle in order to gather data.

Regarding the ICML'19 workshop submission mentioned by Reviewer 2: please note that the corresponding workshop submission deadline was after the NeurIPS submission deadline. Due to the NeurIPS guidelines for anonymous submissions, we cannot comment on this point beyond stating that all applicable rules are being followed to the best of our knowledge.

Reviewer 3

We agree with the reviewer that the non-i.i.d. case (i.e., distribution shifts) is crucial for deploying machine learning and plan to investigate this point in future work. But as stated by the reviewer, the purpose of the current paper is the i.i.d. case since it allows us to clearly separate adaptive overfitting from distribution shift and show that adaptive overfitting is currently not substantively affecting machine learning practice.

Regarding lines 275–278: with "not enough data", we mean that the MetaKaggle dataset only provides the aggregate score for each submission and not the loss on each example in the test set. This makes it impossible to use techniques such as bootstrapping to compute error bounds since we cannot re-sample the test set. In contrast, the accuracy metric is simple enough so that the aggregate score is a sufficient statistic to compute tight (even exact) confidence intervals (Clopper-Pearson confidence intervals).