We thank all the reviewers for a thorough reading and their helpful comments. Our responses are below. We refer to sections/pages etc in the full version of our submission.

Instances that are not perfectly identifiable (Reviewer 1). As stated in §6, our results can be extended to the case where not all pairs of hypotheses can be distinguished. There is, however, some loss in the performance guarantee, which now also depends on the maximum degree of the *similarity graph* G (defined in §6, first paragraph). Graph G contains an edge for every pair of hypotheses that are not identifiable from each other. Let $d = 1+$ max-degree(G). Note that G is empty (and $d = 1$) for perfectly-identifiable instances (assumed in §1-5).

Example: Consider hypotheses $\{1, 2, \cdots m\}$ and m tests where the i^{th} test (for any $i = 1, \cdots m$) has (a) outcome $-$ for hypotheses $\{1, \cdots i - 1\}$, (b) outcome $*$ for hypothesis i and (c) outcome $+$ for hypotheses $\{i + 1, \cdots m\}$. The similarity graph G here is a line with edges $(i, i + 1)$ for all $i = 1, \cdots m - 1$. So $d = 3$ for this instance.

Our current description in §6 gives a policy that stops when the compatible hypotheses H is a subset of any star in G, which we call the *neighborhood stopping criterion*. (The paragraph on "Non-binary outcomes" was unfortunately misplaced.) The description in pages 15-16 outlines how to obtain a non-adaptive $O(d \cdot \log m)$-approximation and an adaptive $O(d + \min(h, r) + \log m)$-approximation for neighborhood stopping. Note that this matches the results stated in Theorem 3.2 and Corollary 4.11.1 where $d = 1$. In fact, our adaptive algorithm's guarantee is stronger: the cost of our algorithm is at most $O(\min(h, r) + \log m) \cdot OPT + d$.

The stopping criterion suggested by Reviewer 1 requires the compatible hypotheses H to be a clique in G (so there is no further test to distinguish between them). We call this the *clique stopping criterion*; note that this is a stricter requirement than neighborhood stopping. Our adaptive algorithm can be easily extended to this criterion. Note that $|H| \leq d$ at the end of our policy for neighborhood stopping. We then continue performing tests that distinguish within H until H is completely indistinguishable (i.e., a clique in G). The number of additional tests is at most d (each test reduces $|H|$ by at least one), which does not affect our worst-case guarantees.

We also tested our algorithms on the WISER dataset (without preprocessing) using both the neighborhood and clique stopping criteria and the results are reported below (for uniform distribution). The resulting similarity graph has $d = 54$ and the number of hypotheses $m = 414$. The preprocessed instance (reported in the submission and reproduced in the first column below) has a smaller set of hypotheses, chosen so that they are perfectly identifiable (we used a greedy rule that iteratively drops the highest-degree hypothesis in G). While we agree that preprocessing can change the objective in an unpredictable way, we think that it still preserves some structure of the original dataset.

Algorithm	Wiser (preprocessed)	Wiser (neighborhood stopping)	Wiser (clique stopping)
# Hypotheses (m)	255	414	414
ODTN$_r$	8.357	11.163	11.817
ODTN$_h$	9.707	11.908	12.506
Non-Adap	11.568	16.995	21.281
Low-Adap	9.152	16.983	20.559

In summary, we present extensions of our results to output a set of scenarios along with a witness that no further distinguishing between any pair in this set is possible (since we return a clique in the similarity graph). In doing this, we achieve a generalized performance ratio of $O(d + \min(h, r) + \log m)$ in the adaptive setting, and promising results on preliminary experiments with the WISER dataset.

Arbitrary probabilities in the noise model (Reviewer 4). As stated in §6 (last paragraph), our results continue to hold in the setting where each noisy outcome has a different (arbitrary) probability to be $+/-$. Theorem 3.2 and Corollary 4.11.1 are unchanged. The approximation ratio in Theorem 5.1 increases by a factor of $\frac{1}{\delta}$ where $\delta > 0$ is the minimum probability of any noisy outcome (assumed to be $\frac{1}{2}$ in §1-5). We decided to focus on the simpler (but representative) case of uniform ± 1 noise in §1-5 only to reduce notational clutter.

Other concerns (Reviewer 4). We acknowledge that our bounds in the proofs of Lemma 3.4 and Proposition 2 were sloppy but they can be fixed by only adding a further constant factor in the guarantees, that are absorbed in the big-oh. Apologies for the typo in the proof of Lemma 3.5. The values of r and h as reported in Section 7 are not mixed up (e.g. in Table 7). Despite the higher value of r in this data set, the performance of the $ODTN_r$ algorithms is superior potentially as a result of the influence of the other dominating logarithmic factor.

Improved presentation of §4 (Reviewer 2). We agree that the presentation in §4 can include more details and better explanations. If the paper is accepted, we will use the extra content page for this.