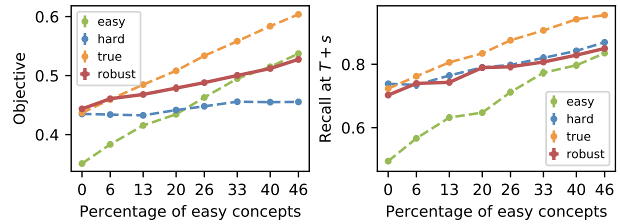
We thank the reviewers for their valuable suggestions. Please find our answers (**A**) for each reviewer (**R**) below.

R1, R3: *Parameter sensitivity study (ln 264–270)*



A: We had conducted a sensitivity study on simulated learners before choosing the HLR parameters for our user study. These detailed results were omitted from the original submission. We report them below and we will include these results in the revision. In this experiment, we consider two groups of concepts: "easy/common" concepts with $\theta = (10, 5, 0)$, and "hard/rare"concepts with $\theta = (3, 1.5, 0)$. Other configurations are kept the same as our user study, with $T = 40$, $n = 15$ and $s = 10$. We vary the number of "easy" concepts from $\{0, 1, \ldots, 8\}$ (i.e., up to 50% of the concepts being easy), and consider four types of teachers: (i) "easy": $\theta = (10, 5, 0)$ for all concepts; (ii) "hard": $\theta = (3, 1.5, 0)$ for all concepts; (iii) "true": using the true parameters for each concept; (iv) "robust": $\theta = (6, 2, 0)$ for all concepts. We plot the performances of these different teachers measured by the two metrics considered in simulations (i.e., the objective value and future recall). As shown in the figures, the "robust" teacher performs well on both metrics, and hence is used for our user study on the German dataset.

R1, R3: *Time scale of real-life learning settings*

A: After the publication of this work, it is quite conceivable to apply these ideas in real-life language learning scenarios. We consider collaborating with existing language learning platforms as a natural step for future work.

R1: *Fit the parameters of the HLR model, and compare them with the current parameters of choice*

A: Thanks for the suggestion. Indeed, one can infer θ for each concept from historical data. We have collected 800 user entries from the random teacher on the German dataset (and 3200 entries on Biodiversity), and it is possible to take existing user study histories and fit an HLR model to get an estimate of θ. We plan to include the results in the revision.

R2: *Relevant pre-existing work: optimal teaching with exemplars; references on HLR memory model*

A: Thanks for pointing us to these references. We will certainly include them in the revision. However, there might be some misunderstanding about the differences between the exemplar-based setting (Patil et al. 2011, Nosofsky et al. 2018) and the setting of our work. Patil et al. (2011) and Nosofsky et al. (2018) investigated the problem of choosing the optimal exemplars (based on the Generalized Context Model) for teaching a *classification* task; whereas for our case, the exemplars for each class are already given (in other words, we have only one "exemplar" per class), and we aim at optimally teaching the learner to *memorize* the (label of) exemplars. It is unclear how one can adapt the algorithm to our setting, as they are addressing two orthogonal problems. We will explain these points in the updated paper.

R2: *Alternative baselines: (1) Optimal "forgetless" leaner; (2) different levels of forgetting*

A: Under our problem setting (as explained in our previous response), if the learner is "forgetless", then after teaching each concept the recall probability becomes 1, leading to a trivial teaching scenario (by showing each concept once). To see the effect of different levels of forgetting, please refer to our sensitivity study results in response to **R1**.

R2: *Significance tests*

A: We performed χ^2 tests (with contingency tables where rows are algorithms and columns are observed outcomes) and obtained very similar statistics with Table 1 (see below). We will include these statistics in the updated paper.

	German				Biodiversity				Biodiversity (common)				Biodiversity (rare)			
	GR	LR	RR	RD	GR	LR	RR	RD	GR	LR	RR	RD	GR	LR	RR	RD
p-value (χ^2 tests)	–	0.0652	0.0197	0.0151	–	0.0017	<0.0001	<0.0001	–	0.3111	0.8478	0.0047	–	0.0001	<0.0001	<0.0001

R3: *Upper-bound on $F(\pi^g)$*

A: This is a very interesting suggestion. It will be interesting to establish an upper bound for the greedy or optimal algorithm under particular model configurations, e.g., to provide a necessary condition for achieving a certain target utility under the HLR model (similar to Thm 5). We will further explore this question as future work.

R3: *Teaching interface: Special consideration for teaching "two consecutive time steps"*

A: In our teaching interface, there was no gap between two teaching iterations. The learner could copy the answer from the previous iteration to the next question if the same concept was shown. Therefore, we treat this case specially to mitigate such effect. An alternative way is to introduce a small break between two teaching iterations.

R3: *Minor Comments*

A: Thanks for the detailed suggestions. We will fix all the issues with careful proofreading, especially in the Appendix. A few specific answers: (i) *Proof of Thm 1 (Page 14, line 449)*: Yes, we will clarify that $g_i(\cdot) \leq 1$ is due to μ being submodular, and (ii) *Proof of Thm 2 (Page 16)*: Inequality (17) does not require an expectation of the second summand as $(\sigma^g_{1:t}, y^g_{1:t})$ is the *observed* history. For the inequality between lines 481-482: this is a good point – we will revise this inequality by imposing the condition only on the first part of the expectation. This does not affect the rest of the proof.