Good teachers with poor assumptions:

Teachers rationally select what information to share, but misrepresent learners' hypothesis spaces

Rosie Aboody[a], Joey Velez-Ginorio[b], Laurie Santos[a], & Julian Jara-Ettinger[a]

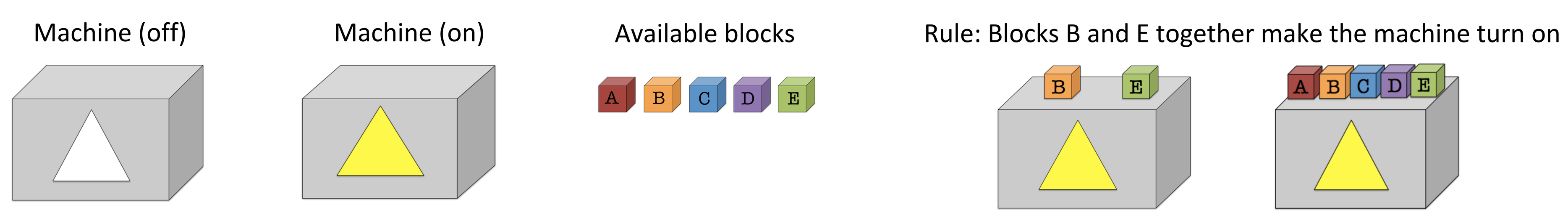
[a]Yale University, [b]Massachusetts Institute of Technology

Introduction

- When we teach, how do we decide what information to share? And, how effective are our teaching choices?
- Past research has found that adults teach **rationally** in short tasks (i.e., Shafto, Goodman & Griffiths, 2014), but **fail** to recognize and ameliorate learners' knowledge gaps in longer tasks (i.e., Chi, Siler & Jeong, 2004).
- We provide an account of teaching that unifies these findings.

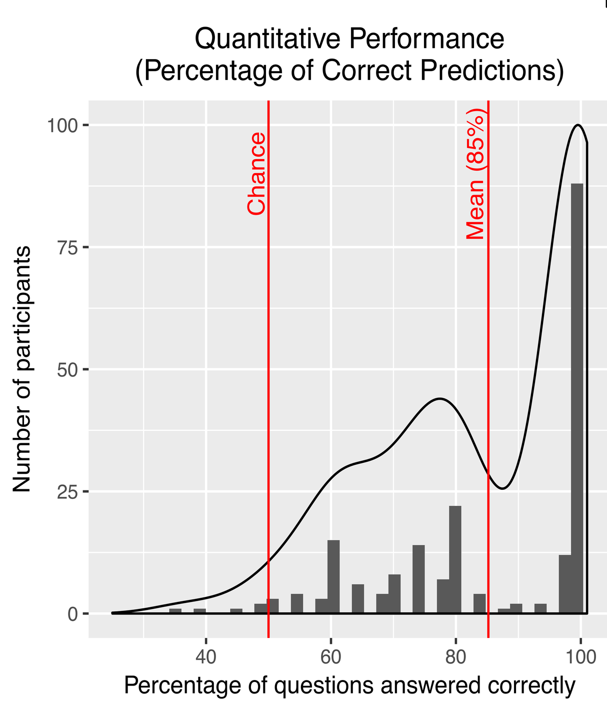
Experiment 1

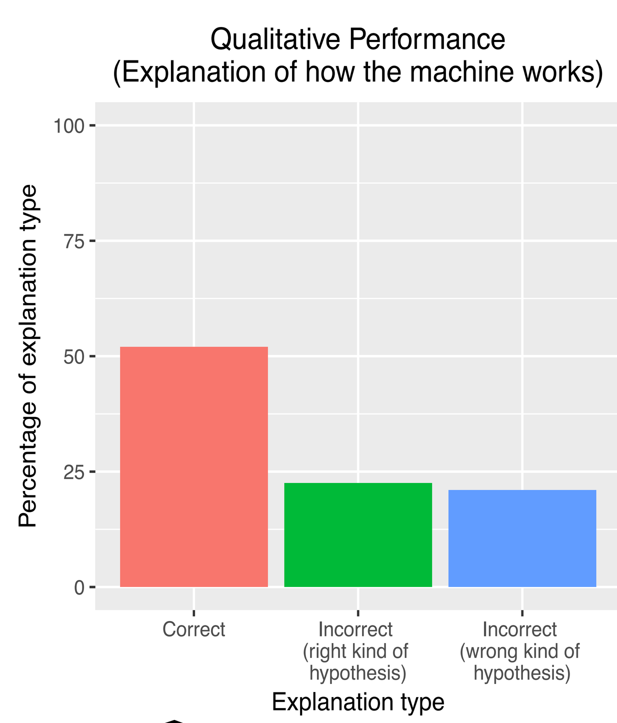
- *Participants*. Teachers, n = 20; Learners, n = 200; Amazon Mechanical Turk
- *Procedure*. Teachers learned how to activate a machine, and selected examples that would communicate this information to a learner. After seeing a teacher's examples, learners **explained** how the machine worked (qualitative), and answered **test questions** (quantitative).



- *Results.*

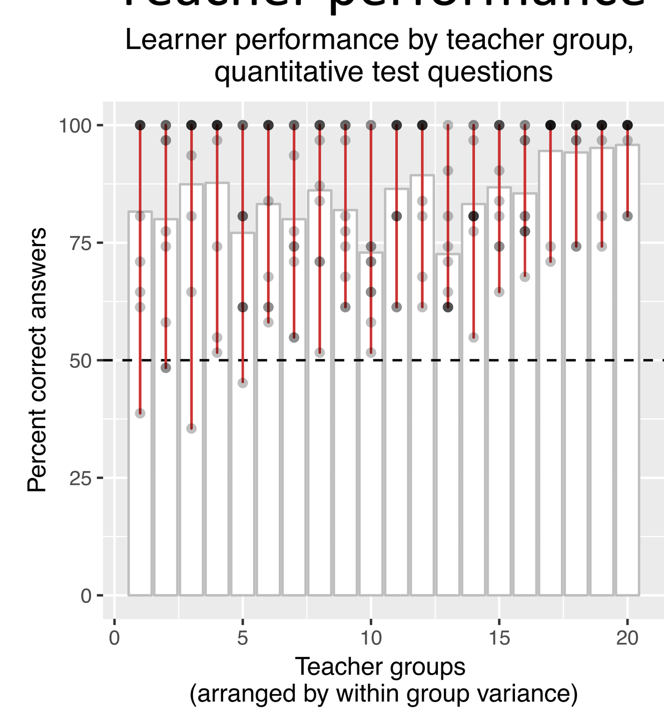
Learner performance





b = -.74, *t*(198) = -15.44, *p* < .01

Teacher performance

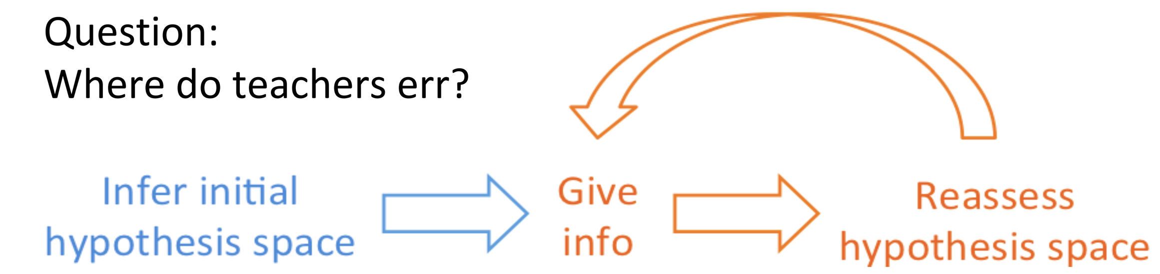


Most teachers were confident they had taught well. (Average teacher confidence: 6.6/7)

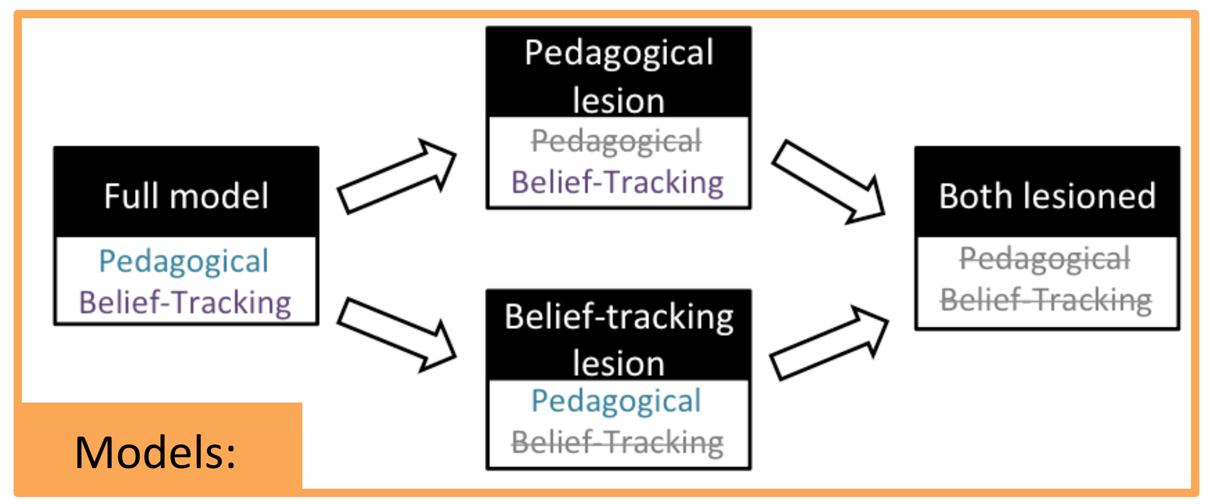
- *Discussion*. While 97% of learners scored above chance on the quantitative test questions, only 44% of learners scored perfectly, and only 52% of learners conveyed the correct activation rule in their qualitative explanation. **What kinds of teaching errors led these partial learner failures?**

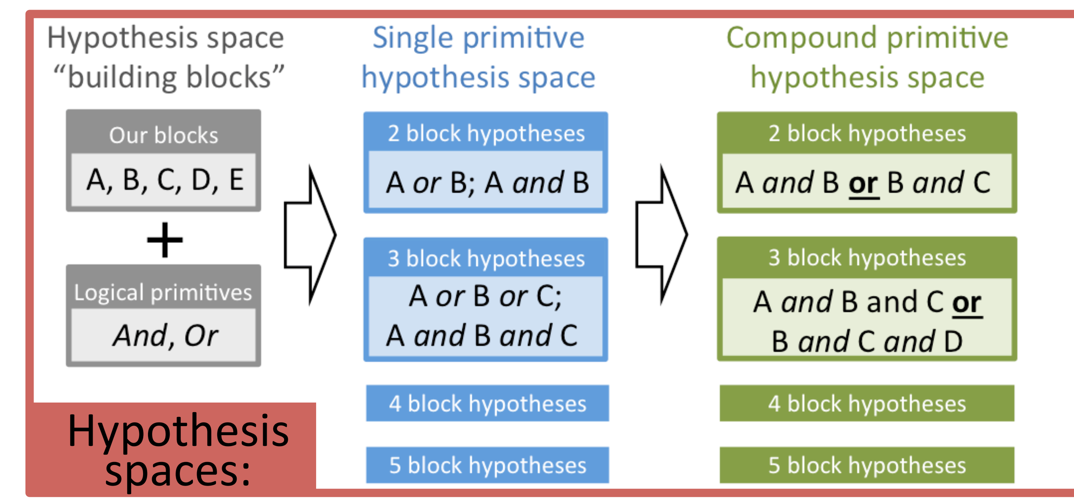
Model-Based Analysis

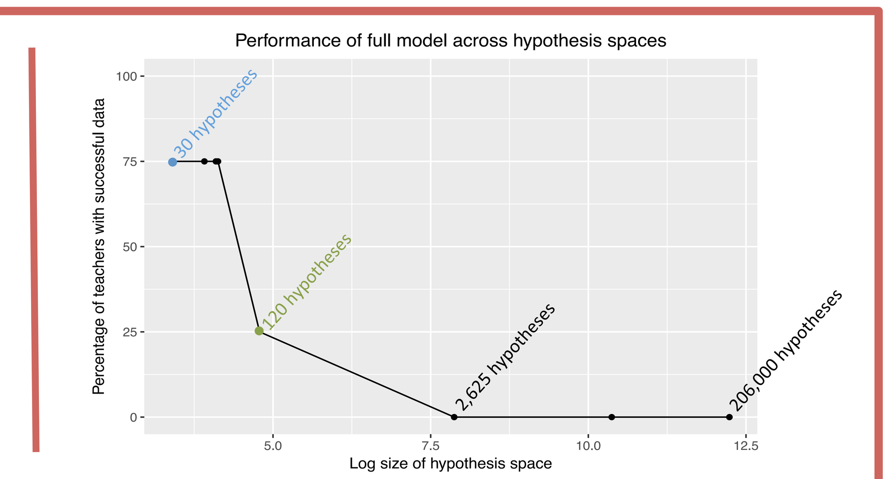
- *Questions*. Did teachers provide **bad data**, or did they produce **good data**, under a misconception about the types of hypotheses learners were considering?

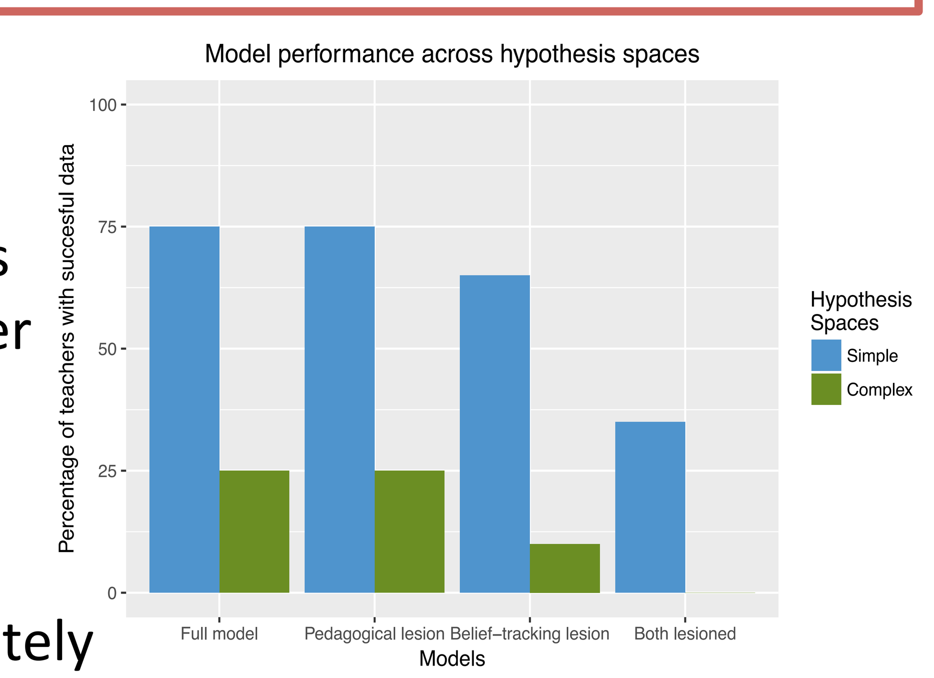


- *Models* & *Hypothesis spaces*. We built a computational model of a rational teacher: $p_{learner}(H|E) \propto p_{teacher}(E|H) * p_{learner}(H)$ (following Shafto et al., 2014), lesioning it to create three simpler models. If teachers produced bad data, **no** rational learner model should be able to learn from teachers' examples. But, if teachers produced good data under inaccurate assumptions about learners' initial knowledge states, our rational learner model *should* learn the machine's activation rule given the correct initial hypothesis space. Thus, we built 8 hypothesis spaces of varying size and complexity.









- *Results.* Multiple models learned the activation rule from the majority of teachers' examples. We find no evidence that teachers thought learners would interpret data pedagogically, although teachers do select data under these assumptions in other tasks (Shafto et. al., 2014). We find weak evidence that teachers tracked learners' beliefs over the course of the task. However, we find strong evidence that, in order to successfully teach, teachers *either* had to generate data under pedagogical assumptions, or had to track learners' beliefs throughout the task.
- *Discussion*. These results suggest that teachers selected **good data**, but failed to accurately assess learners' initial knowledge states (perhaps cursed by their own knowledge to overestimate that of learners).

General Discussion & Conclusion

- In sum, we find that teachers fail to accurately grasp the types of hypotheses learners initially consider, and thus cannot fully inform learners. Therefore, past divergent findings may have been caused by differences in the size of learners' potential hypothesis spaces, with teachers succeeding when the space of possible learner hypotheses is constrained, but failing as this space of possibilities grows.
- *Future Directions*. To provide additional support for this account, we will run a third experiment, in which we will limit learners' hypothesis spaces to the correct kinds of hypotheses. If our account is correct, and it is learners' initial hypothesis space that matters, learners should now succeed in this task.

Works Cited

Chi, Siler & Jeong (2004). *Cog and Instr, 22*(3), 363-387.

Shafto, Goodman & Griffiths (2014). *Cog Psych, 71*, 55-89.