



# Are web-based Amazon Mechanical Turk studies on speech perception reliable?



Crystal Lee\*, Lauren Oey\*, Emily Simon\*, Xin Xie, T. Florian Jaeger  
University of Rochester (\*authors contributed equally)  
{clee69, loey, esimon3}@u.rochester.edu, {xxie13, fjaeger}@ur.rochester.edu

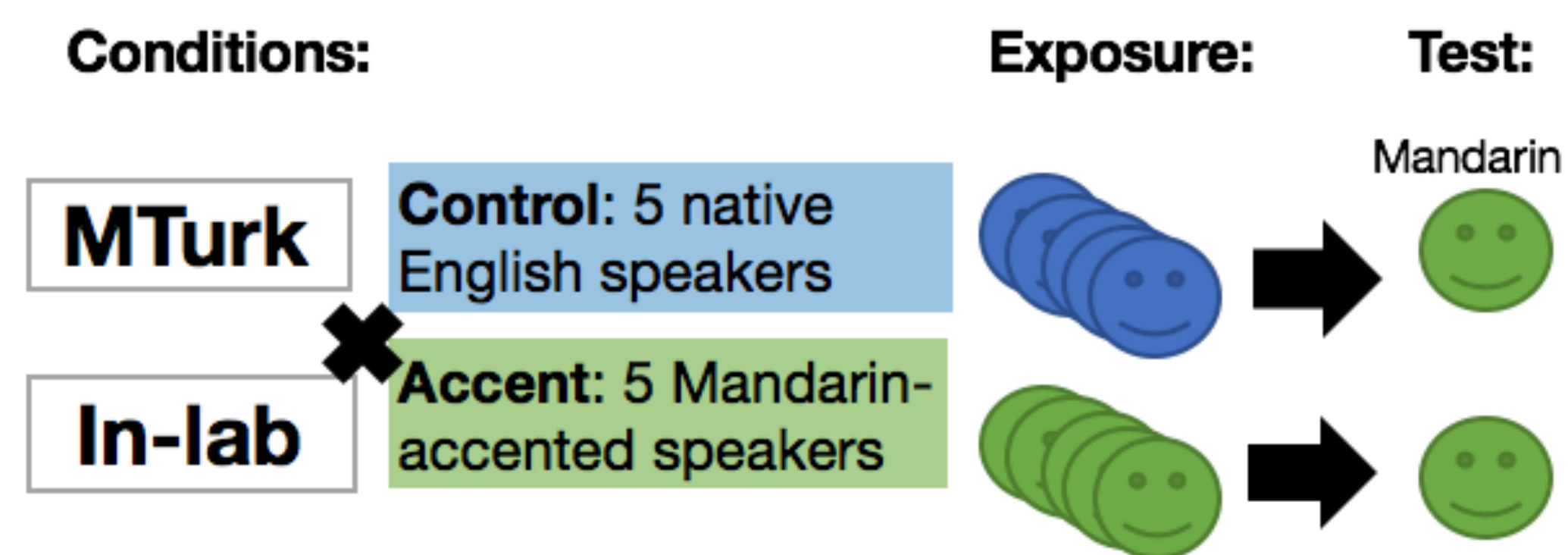
## Introduction

We examine the validity of web-based Amazon Mechanical Turk (MTurk) for **speech perception experiments**. Over MTurk, sound equipment varies between participants, which might affect the reliability of such studies. We replicated an influential study on accent perception both over the web and in the lab<sup>2</sup>.

## Previous work

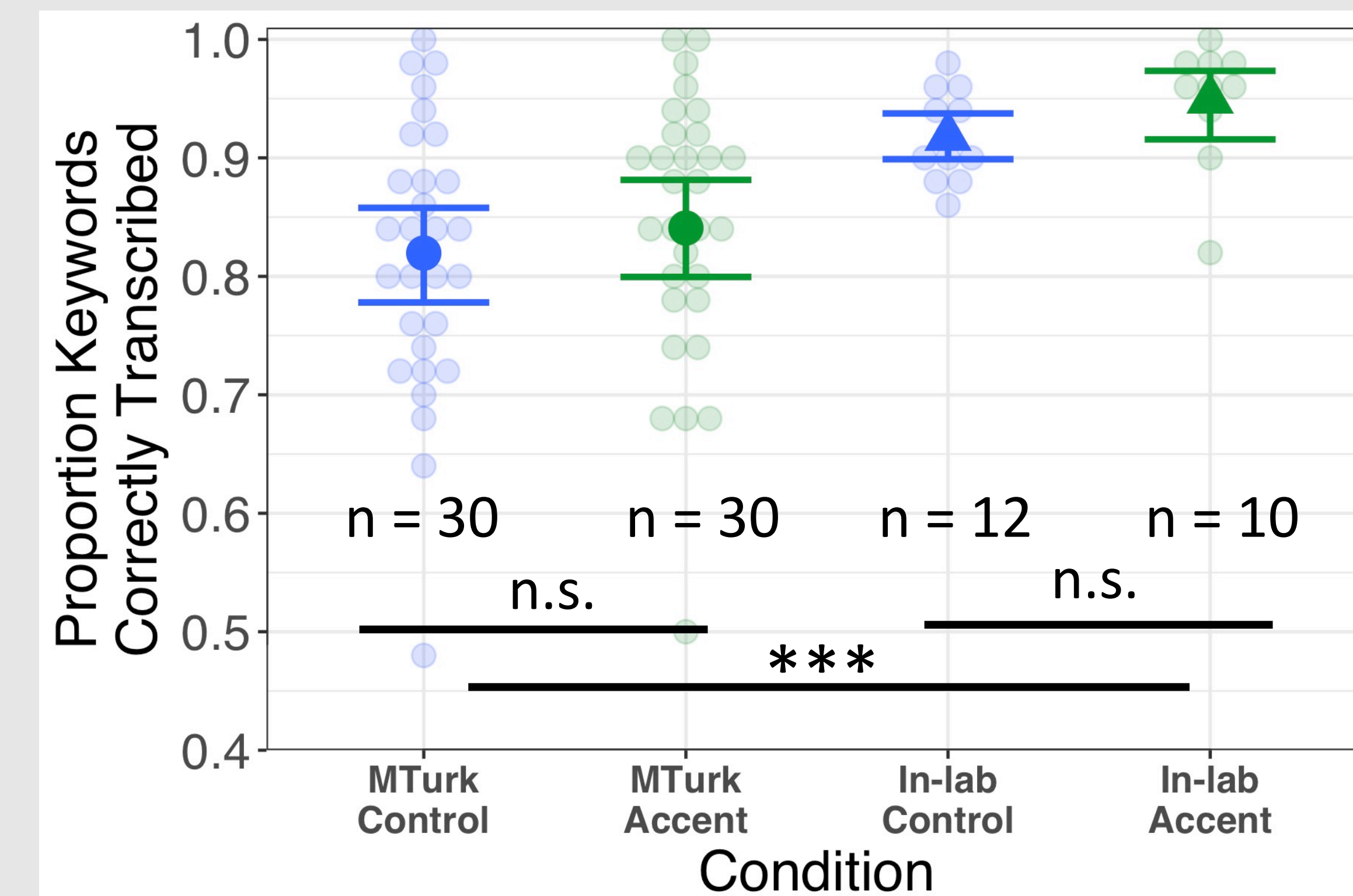
Limited studies on speech perception on MTurk<sup>1,4,6</sup> have yielded conflicting evidence on validity of the use of MTurk<sup>3,4,5</sup>.

## Our experimental paradigm



- **Exposure:** transcribe simple sentences of either English (**Control**) or Mandarin-accented (**Accent**) speakers
- **Test:** both **Control** and **Accent** conditions transcribe a novel Mandarin-accented speaker

## Results: accuracy during test



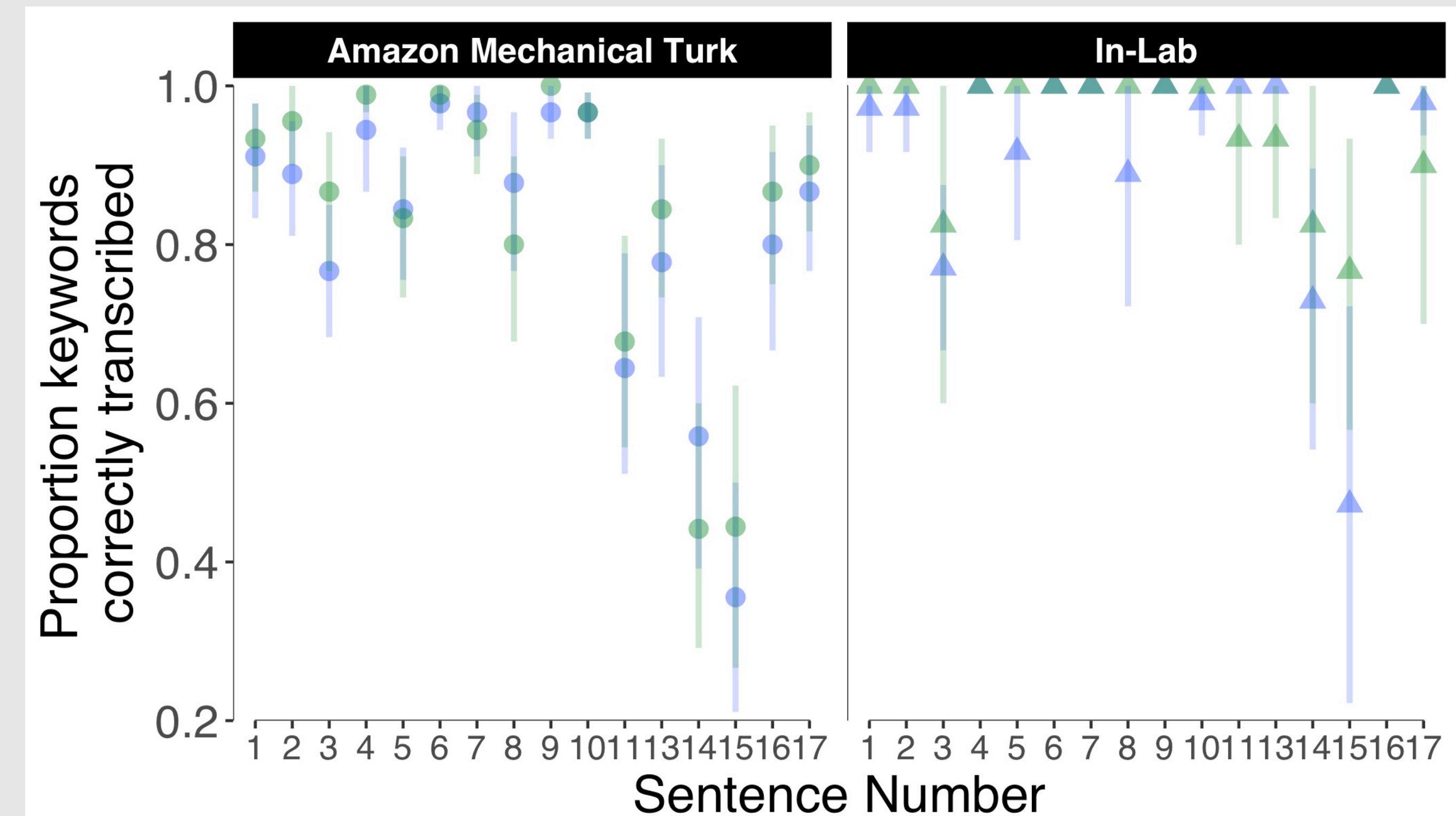
- In-lab participants (undergrads) significantly outperform MTurk workers
- No difference in performance between **Control** and **Accent**

- **By-item performance is similar across testing conditions**

## Post-hoc analyses

Other predictors we explored did not explain difference in performance between In-lab and MTurk

- Time of day
- Gender
- Age



## Conclusion

- **No evidence for a strong generalization effect**
- Both testing conditions showed a **trend in a similar direction**
- Better performance for In-lab, **but overall similar effects:**
  - **Similar trend for (null) effect of Exposure Condition for both In-lab and MTurk**
  - **Very similar by-item effects**

References: <sup>1</sup>Bainton et al., 2016. <sup>2</sup>Bradlow & Bent, 2008. <sup>3</sup>Crump et al., 2013. <sup>4</sup>Fort et al., 2011. <sup>5</sup>Sprouse, 2011. <sup>6</sup>Weatherholtz et al., 2016.

Acknowledgements: NICHD R01 HD075797 to TFJ. All views expressed are those of the authors and do not necessarily reflect those of the sponsors. We are grateful to Ann Bradlow for providing access to the Wildcat corpus, discussions through the undergraduate research course BCS 206/207, and support of the Dean of Undergraduate Studies.