Overhearers acquire perspective alignment through observational prediction errors

Yichi Zhang[1], Zhenguang G. Cai[1]

[1] Department of Linguistics and Modern Languages, The Chinese University of Hong Kong

Correspondence: yichizhang@link.cuhk.edu.hk

Background. Linguistic alignment, interlocutors' convergence on shared representations, has been studied mainly as dyadic coordination (Pickering & Garrod, 2004; Branigan et al., 2000). Yet conversations rarely involve only two parties (Clark, 1996; Holler & Kuhlen, 2026): overhearers and bystanders routinely witness others' interactions. Whether observers acquire alignment strategies from such observation, and by what mechanism, is unresolved. Two accounts make divergent predictions. A passive-priming account predicts item-specific effects driven by exposure frequency, insensitive to the relational context in which expressions are produced. An observational reinforcement-learning (RL) account (Burke et al., 2010; Zhou et al., 2024) predicts strategy-level updating driven by trial-by-trial discrepancies between predicted and observed demonstrator behavior, with effects that generalize across items and partners.

Method. Eighty-eight Mandarin-speaking adults completed a three-phase referential task adapted from Branigan et al. (2011). Mandarin admits an occupation-based form (*yi$_1$sheng$_1$*, "doctor") or an appearance-based form (*chuan$_1$ bai$_2$da$_4$gua$_4$ de ren$_2$*, "person in a white coat") for the same referent. In a *baseline session*, participants played a picture-matching-and-naming game (Fig. 1A) with a pre-recorded 5-year-old child whose references followed one perspective. In an *observational-learning session* (4 blocks × 12 trials), they observed an adult demonstrator play the same game with the child, predicted (made an explicit guess about) the demonstrator's reference, and then played the same game with the same child; in each trial. We manipulated, in a between-subject design, the demonstrator to have high or low perspective alignment with the child (aligning on 77.8% vs. 22.2% of trials respectively). In a *generalization session*, participants played picture-matching-and-naming game with a *new* child not present during the baseline or learning session. Trial-level behavior was modelled in two stages: Stage 1, a Rescorla-Wagner model fitted to participants' demonstrator-alignment predictions to verify learning and recover learning rates; Stage 2, Bayesian comparison of four models predicting participants' own alignment choices from cumulative observational prediction errors (PEs) versus an imitation alternative.

Results. (1) *Behavioral transmission and generalization.* A Prime × Group × Session interaction ($\beta = 5.30$, $p < .001$) showed that the high-alignment group increased its alignment from baseline to generalization (Prime × Session: $\beta = 2.18$, $p = .013$) while the low-alignment group decreased it ($\beta = -3.20$, $p < .001$). Both effects measured against a *new* child, never directly addressed during learning (Fig. 1B). (2) *Learning verified.* The RW model captured demonstrator-alignment predictions. (3) *Computational mechanism.* Model comparison favored a PE-based model over imitation (ΔWAIC = 24.94). A positive PE arises when the demonstrator aligns despite the participant predicting non-alignment (an upward surprise); a negative PE is the converse, the demonstrator fails to align when the participant expected it (a downward surprise). PE effects emerged in the second half of observation (Fig. 1C). In the high-alignment group, both surprise types updated behavior as predicted: more alignment after upward surprises, less after downward ones. In the low-alignment group, only downward surprises reached significance, consistent with upward surprises being rare when the demonstrator aligned on only 22% of trials. Together, this pattern, directional updates that track surprise, not the demonstrator's overall behavior, is the signature of error-driven learning over imitation. (4) *Alignment as social signal (exploratory).* In post-experiment ratings, observers rated high-alignment demonstrators as more perspective-taking than low-alignment demonstrators, indicating that participants tracked alignment as a perceptible attribute of the demonstrator.

Implications. ***First***, perspective alignment is *socially transmissible*: an observer's alignment is shaped by what they witness another speaker do, even without direct interaction with the addressee. ***Second***, this update signature indicates *strategy-level expectation updating* rather than item-level priming. ***Third***, transfer to a novel child, with whom no common ground was established, speaks to the partner-generality debate (Brennan & Clark, 1996; Barr & Keysar, 2002): under observational learning, what is acquired behaves as an *abstract communicative adjustment*, not a partner-specific convention. The findings extend dyadic accounts of alignment to multi-party settings, answer Holler and Kuhlen's (2026) call for multi-party psycholinguistic accounts and offer a computational account of how communicative norms propagate across speakers without direct interaction.

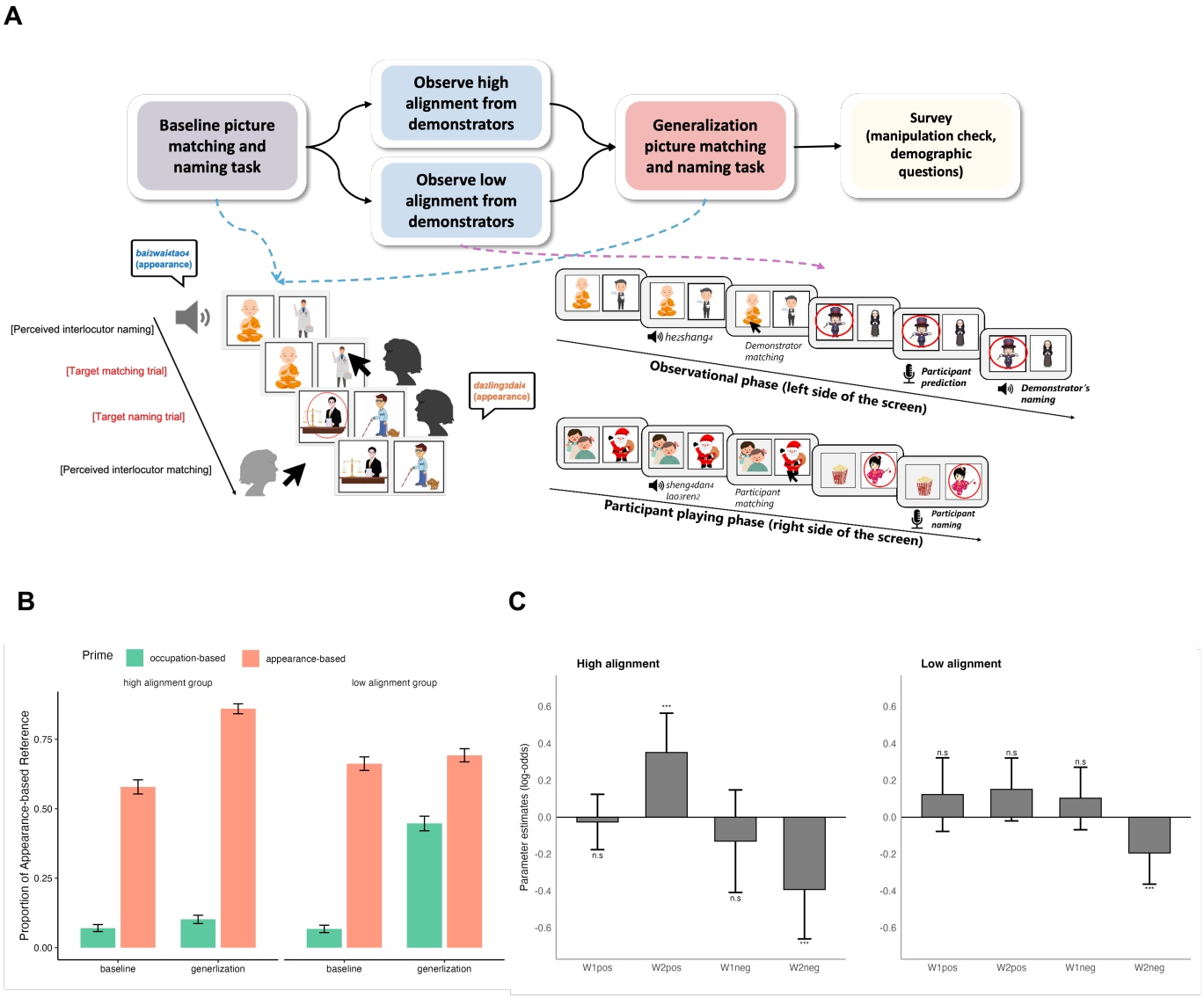

Figure 1. A. Picture matching and naming game in baseline and generalization session.Three-phase design (baseline → observational learning [high vs. low alignment] → generalization with novel child). B. Alignment effect cross sessions: high-alignment group .51 → .76; low-alignment group .59 → .24. C. Stage-2 weight estimates by group: significant W2_pos and W2_neg in the high-alignment group; significant W2_neg only in the low-alignment group.

References

Barr, D. J., & Keysar, B. (2002). Anchoring comprehension in linguistic precedents. *Journal of Memory and Language, 46*(2), 391-418.

Branigan, H. P., Pickering, M. J., Pearson, J., McLean, J. F., & Brown, A. (2011). The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition, 121*(1), 41-57.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic co-ordination in dialogue. *Cognition, 75*(2), B13-B25.

Brennan, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition, 22*(6), 1482.

Burke, C. J., Tobler, P. N., Baddeley, M., & Schultz, W. (2010). Neural mechanisms of observational learning. *PNAS, 107*(32), 14431–14436.

Clark, H. H. (1996). *Using language*. Cambridge University Press.

Holler, J., & Kuhlen, A. K. (2026). Psycholinguistic perspectives on face-to-face conversation. *Nature Reviews Psychology*, 1-16.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences, 27*(2), 169-190.

Zhou, Y., Han, S., Kang, P., Tobler, P. N., & Hein, G. (2024). The social transmission of empathy relies on observational reinforcement learning. *PNAS, 121*(9), e2313073121.