# Mental Synchronization in Human Task Demonstration: Implications for Robot Teaching and Learning

Julia Oppenheim*
Johns Hopkins University

Jindan Huang*
Tufts University

Isabel Won
Johns Hopkins University

Chien-Ming Huang†
Johns Hopkins University


**Figure 1: We examined communicative acts used for mental synchronization in situated human task demonstration.**

## ABSTRACT

Communication is integral to knowledge transfer in human-human interaction. To inform effective knowledge transfer in human-robot interaction, we conducted an observational study to better understand how people use gaze and other backchannel signals to ground their mutual understanding of task-oriented instruction during learning interactions. Our results highlight qualitative and quantitative differences in how people exhibit and respond to gaze, depending on motivation and instructional context. The findings of this study inform future research that seeks to improve the efficacy and naturalness of robots as they communicate with people as both learners and instructors.

## CCS CONCEPTS

• **Applied computing** → **Psychology**.

## KEYWORDS

Task Demonstration; Grounding; Gaze; Backchanneling

**ACM Reference Format:**
Julia Oppenheim, Jindan Huang, Isabel Won, and Chien-Ming Huang. 2021. Mental Synchronization in Human Task Demonstration: Implications for Robot Teaching and Learning. In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction (HRI '21 Companion), March 8–11, 2021, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3434074.3447216

*Both authors contributed equally to this research.
†cmhuang@cs.jhu.edu


*HRI '21 Companion, March 8–11, 2021, Boulder, CO, USA*

ACM ISBN 978-1-4503-8290-8/21/03…$15.00
https://doi.org/10.1145/3434074.3447216

## 1 INTRODUCTION

Effective knowledge transfer between humans and machines is critical to the future of work in which humans and machines are envisioned to collaborate closely. A common way of human knowledge transfer is through physical demonstration. This is especially true for tasks that are spatially structured and require motor skills, such as assembly and repair of structural parts. During physical task demonstrations, people use a range of communicative acts, such as eye gaze, gestures, and head nods, to build and synchronize perceptual and procedural common ground [5]. The key insight driving the present exploration is that *people's natural communicative acts are integral to the understanding and learning of a task.* For example, in an episode of establishing joint attention, an instructor displays a referential gaze cue toward a task object of interest and then looks back at the learner; the look-back gaze behavior is meant to check whether the learner followed the attentional cue, thereby implicitly underscoring the relevance and importance of the object of reference in the demonstrated task.

In this paper, we report an observational study examining why and when human instructors use gaze to check on learners during physical task demonstrations, as well as how the learners respond with backchannel signals, to synchronize mutual understanding. This examination offers an empirical understanding of knowledge transfer in physical demonstrations. Its findings provide implications for enabling productive robot teaching and learning from human task demonstrations.

## 2 BACKGROUND

Grounding refers to the interactive process by which individuals build mutual understanding and maintain a common culture for their conversations or collaborative activities [5]. During the

grounding process, an individual monitors other individuals for signs of understanding and provides explicit or implicit feedback to mitigate misunderstanding [4]. Grounding can be achieved through a combination of behavioral channels during collaborative tasks. For example, a speaker may point or use representational gestures to indicate their intention [8], utilize mutual visual space to coordinate activities [12], and exhibit gaze cues to direct and establish joint attention [21].

Acting as a function of both perception and signaling [14], gaze can be used to monitor conversational partners for understanding, regulate conversation, and express conversational intention [1]. Exemplifying its versatility, gaze direction and duration varies depending on whether one highlights information, seeks approval, or signals willingness to communicate [15]. For example, although speakers tend to look at each other's faces during conversation [3], they tend to look at the workspace and task related targets during co-located collaborative physical tasks [7].

Also critical to grounding is backchanneling, when a listener directs a simple verbal or non-verbal response back to a speaker during conversation. It indicates that the listener follows and is paying attention to the speaker, and may influence the speaker's desire to continue talking. Common examples of backchannel signals are nods and short utterances such as "*uh huh, yeah*". Nodding in particular is meant to signal that the listener understands without interrupting the speaker [16, 20].

Prior research in HRI has investigated how to design communicative social cues, such as backchannel signals (e.g., [11, 13]), gaze (e.g., [2, 10]), and gestures (e.g., [9]), for robots to participate in productive joint activities with people. Different from these prior works, this study focuses on understanding how natural communicative cues are exhibited during situated learning interactions and how such understanding may inform the development of productive robot teachers and learners.

## 3 DATA COLLECTION AND CODING

We conducted a data collection study to understand the behavioral cues projected and processed during situated task demonstrations. In our study, one participant, the *instructor*, taught the other participant, the *learner*, one of the two assembly tasks: Pipe and Lego (Figure 2). The participants then switched roles in the other assembly task. Both tasks involved spatial awareness and repetitive assembly of structures. On average, *pipe teaching* interactions ($M = 7.03, SD = 1.25$) were comparable in minute duration to *Lego teaching* interactions ($M = 7.48, SD = 2.17$).

### 3.1 Procedure

Upon consenting to participate in this study, participants completed a brief background survey covering their age, area of study, previous teaching or tutoring experience, and spatial awareness skills. Additionally, they completed the Big Five Personality Test [6].

Participants then entered a phase of self-teaching at their respective stations, separated by a divider. They were randomly assigned to be either the *pipe instructor* or the *Lego instructor*. As tools for self instruction, the *pipe instructor* was supplied with pictures of the abstract pipe structure from three different angles. They also had access to a one-minute video explaining how to connect pipe

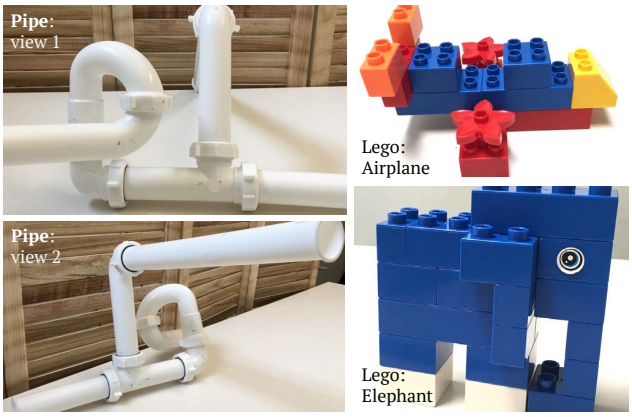


**Figure 2: The pipe and Lego structures used in our study.**

joints, as the proper protocol was not evident from the images. The *Lego instructor* received an image of both the elephant and airplane structures. Both images were from the official Lego instructions. Once the participants were acclimated to the tools available to them, they had 10–15 minutes to learn how to build their assigned task. During this time, both participants had unrestricted access to the learning materials available to them, as well as the pieces to the structures they were assigned to build. Participants entered this phase of self-learning under the known pretext that they would have to teach their assigned task to the opposite participant.

Upon completion of self-teaching, participants entered the knowledge exchange phase of the study. Both participants moved to the same table and sat directly across from each other. The *pipe instructor* then, using the pipe pieces, taught the *Lego instructor* how to build the pipe structure. The learner was instructed not to touch the pieces until the completion of first run-through of the structure. After pipe teaching concluded, the *Lego instructor* taught the other participant how to build the two Lego structures under the same conditions. Both knowledge exchange interactions were recorded from three views: a wide, side view of both of the participants, a frontal instructor view, and a frontal learner view (Figure 1).

Finally, both participants were allotted 10 minutes to demonstrate their newly acquired knowledge of the opposite participant's task. Both participants returned to their independent stations, and the *pipe instructor* built the two Lego structures while the *Lego instructor* built the pipe structure. During this phase, participants were not permitted to ask questions and worked completely from memory of the knowledge exchange. The whole study concluded after each participant completed this final task (or at the conclusion of 10 minutes). Each participant received a $10 Amazon gift card as compensation for their time. We then conducted a retrospective think aloud with one pair of participants, during which they watched their knowledge exchange and provided insight into their checking and backchannel behavior. These participants were compensated for this portion of the study, as well.

### 3.2 Participants

Twenty-two participants, all fluent English speakers aged 18–25, engaged in the hour-long, data collection study. These participants were divided into 11 dyads and had limited prior interaction with one another. One pair of participants piloted the study with three,

not two, Lego structures. Another pair did not follow the experimental protocol and jointly built the structures. As a result, nine out of the 11 pairs were included in our data coding and behavioral analysis. Of the participants included in our analysis, ten self-identified as female and eight as male. Twelve participants primarily studied or were studying engineering disciplines in college, and all but five participants had prior teaching or tutoring experience. On a scale of 1–5, 5 being excellent and 1 being poor, participants rated their spatial awareness skills at an average of 3.38 ($SD = 0.92$).

## 3.3 Behavioral Coding

Our coding focused on three areas: 1) why the instructor checked on (gazed toward) the learner, 2) how the learner responded to checking cues, and 3) the different phases of the instructional process.

*Motivations of Checking.* Through in-depth review of Lego and Pipe task videos, we identified three key event types that led to the instructor's checking behaviors:

- *Providing supplemental information.* The instructor provides information that supplements the demonstration in order to help the learner process the demonstration. This type was further coded with four categories that include 1) showing action, 2) labeling items, 3) making a reference, and 4) providing spatial information.
- *Highlighting task specifics.* The instructor draws the learner's attention to specific aspects of the demonstration. This event type contains three categories: 1) displaying part or all of the structure, 2) repeating teaching steps, and 3) providing strategic tips.
- *Conversing.* The instructor checks on the learner periodically as in common natural conversation.

*Teaching Phases.* We categorized the instructional process into key phases: delivery; review (overview, recap, re-demonstration, and strategy); feedback (confirmation, clarification, and correction); transition (preparation, regroup, and topic change); and other (engaging with the learner verbally or talking to oneself; "*Any question about this part?*", "*This is so confusing. I can't remember the exact structure but anyway we'll go with this.*").

Two people coded the data independently. One coder coded all nine videos of the Lego task, while the other coded all nine videos of the Pipe task. Both coders completed two videos of each task (n=4). Two of the videos were used to establish the coding scheme; the other two were used to verify coding reliability. Both coders achieved high agreements on motivations of checking (Lego: 92%; Pipe: 100%) and on teaching phases (Lego: 88%; Pipe 89%).

# 4 BEHAVIORAL ANALYSIS AND FINDINGS

## 4.1 Gaze Checking on Learners

We combined qualitative accounts from the retrospective think aloud with quantitative results across all pairs, to better understand how instructors use gaze to synchronize their mental model of the task with that of the learner.

*4.1.1 For what reasons do instructors check on learners?* Instructors check on learners for a variety of reasons. Table 1 illustrates how these motivations varied across teaching phases. Instructors

**Table 1: Distribution of gaze counts by checking motivations in different teaching phases.**

| | | *Teaching phases* | | | | | |
|---|---|---|---|---|---|---|---|
| | | Delivery | Review | Feedback | Transition | Other | Total |
| *Motivation of checking* | Supplement | 139 | 150 | 1 | 4 | 2 | **296** |
| | Highlight | 32 | 165 | 7 | 5 | 5 | **214** |
| | Converse | 11 | 20 | 8 | 41 | 87 | 167 |
| | Total | **182** | **335** | 16 | 50 | 94 | 677 |

checked on learners the most during the *delivery* and *review* teaching phases. Additionally, they checked most commonly to supplement and highlight aspects of their teaching. For example, during the Lego task, the instructor reported that she would find herself looking up at the student either *"if I did something myself that I thought would be confusing to him,"* or *"if I felt like I wasn't explaining something very well."* This provides evidence that instructors' checking behavior mainly served to ensure learners were following the integral steps of task. Instructors also checked on learners to look for non-verbal cues of understanding. For example, the instructor of the Pipe task explained that *"I looked over at her a couple times to kind of see if it looked like she was confused by anything, but she seemed like she had pretty intent focus."* This method of checking also serves to supplement instructor behavior during delivery and review phases of the task, as noted by one instructor who categorized pieces: *"if [I] labeled [them], as a certain shape, it would be easier for [the learner] to understand"*.

Moreover, another common reason for instructors to check on learners involved their own predictions of task complexity. For example, during the Lego task, the instructor reported that she looked up at the learner to ensure he was following along, based on her prediction of how complex the step was. She explained that "*I knew that if I was in his position, I'd need to see it [the task] another time.*" Instructors also often referred to their own understanding of the task as a guide for when they should check on the learners. The Lego task instructor explained, "*I remember I personally had a really hard time even figuring out from the picture how to build this ... so [I was] just making sure I was doing a decent job teaching him.*"

*4.1.2 How often do instructors check on learners?* On average, instructors checked on the learner 4.14 times per minute with a standard deviation of 3.96 (Pipe task: $M = 4.00, SD = 3.98$; Lego task: $M = 3.87, SD = 3.86$). The average interval between the instructors' gaze cues was 13.19 seconds with a standard deviation of 25.63 seconds (Pipe task: $M = 13.02s, SD = 27.81s$; Lego task: $M = 13.36s, SD = 23.06s$).

*4.1.3 Duration of gaze cues.* We explored whether gaze duration can be used to differentiate motivations of checking and teaching phases. To this end, we ran a two-way ANOVA, where checking motivation and teaching phase were set as two independent variables, and gaze duration was set as a dependent variable. Our analysis did not reveal any significant main effect of checking motivations ($F(2, 662) = 0.12, p = .891$) and teaching phases ($F(4, 662) = 2.22, p = .066$) on gaze duration, nor did we see significant interaction effect ($F(8, 662) = 0.67, p = .715$). Table 2 presents the data of gaze length, indicating that gaze duration was comparable under different reasons for checking and that teaching phases generally have more influence on gaze duration. The influence of teaching

**Table 2: Gaze duration breakdown by motivations of checking and teaching phases**

| *Motivation of checking* | Gaze duration (*ms*) | | *Teaching phases* | Gaze duration (*ms*) | |
|---|---|---|---|---|---|
| Supplement | M=813.96 | SD=507.12 | Delivery | M=**745.35** | SD=398.15 |
| Highlight | M=844.06 | SD=549.71 | Review | M=848.36 | SD=575.25 |
| Converse | M=925.56 | SD=715.41 | Feedback | M=**1309.25** | SD=611.54 |
| | | | Transition | M=839.32 | SD=858.46 |
| | | | Other | M=993.17 | SD=641.71 |

phase on gaze duration may stem from the discrete mapping between assumptions about the learner's mental model and each phase. For example, gaze during *delivery* was on average shortest in duration, whereas it was longest during *feedback*. As the default mode, *delivery* assumes that the learner builds a mental model equal to that of the instructor. However, *feedback* aims to reshape the learner's mental model of the task, and may therefore require more prominent grounding behaviors.

## 4.2 Behavioral Responses to Gaze Checking

In addition to understanding the instructor's checking behavior, we examined the learner's responses to checking during mental synchronization. We considered the learner's behavior as a response to the instructor's checking cues if the behavior happened between $T_{Gaze-begin}$ and $T_{Gaze-end}$ + 1s. We note that the learner may exhibit responses through more than one behavioral channel.

*4.2.1 How do learners respond to checking behavior?* Verbal and nodding backchannel responses to checking cues were most common amongst learners. Moreover, the frequency of these responses relative to gaze and lean was more pronounced when learners responded to supplemental or highlighting gazes versus conversational gazes (Figure 3). This data mirrors observations from the study, during which learners directed their gaze primarily at the task work space instead of the instructor. Additionally, we found that 40.4% of the learners' responses were exhibited in multiple channels while learning the Lego task, whereas only 26.4% were observed while learning the Pipe task.

*4.2.2 How long do learners take to respond to checking behavior?* We defined response time as the time between the beginning of a checking behavior and the beginning of its corresponding learner response. If there were multiple responses with respect to a checking behavior, we counted the first response. On average across the two tasks, the response time was 707.08 milliseconds; the Lego and Pipe tasks had similar response times (Lego task: $M = 709.34ms, SD = 536.28ms$; Pipe task: $M = 704.69ms, SD = 554.09ms$).

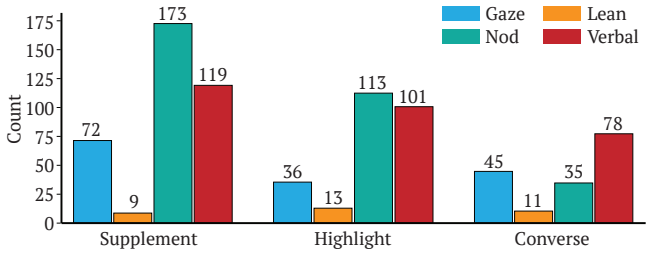


**Figure 3: Student responses to different gaze motivations.**

*4.2.3 How often do learners respond to checking behavior?* We counted the number of times the learner responded to a checking behavior and divided it by the number of checking behaviors. The overall response rate was 70.17% (Lego task: 75.62%; Pipe task: 64.72%). We speculate that because the Pipe task was structurally abstract and complex, it required more intense focus on the demonstration from the learner. This limited their processing bandwidth for responses. The need for intense focus may have also limited opportunity for multimodal responses to checking cues in the Pipe task (Sec. 4.2.1).

# 5 DISCUSSION

**Implications for Robot Teaching**. It is imperative that instructors check on learners periodically to ensure they follow the instruction. Our findings offer design parameters for enabling effective robot teaching. For example, robot instructors may check on human learners approximately four times per minute. However, it is important to note that this frequency must be adaptive to instructional context, rather than following a static, fixed timeline. Our study reveals that instructors checked on learners more frequently during teaching phases *delivery* and *review*, in which the instructors and learners sought to achieve mutual understanding. Moreover, robot instructors should also regularly check on human learners when *highlighting* and *supplementing* task instructions. For instance, robots should check whether learners attend to referenced aspects of the task through either verbal or gestural indications; this process parallels that of initiating joint attention [10].

**Implications for Robot Learning**. As learners, robots can leverage human instructors' gaze checking behaviors to infer which parts of the task demonstration are critical, interject questions, and request additional demonstrations. Furthermore, robot learners should provide appropriate backchannel signals to facilitate human teaching [13, 18]. Our findings suggest that nods and brief verbal responses are suitable for learning scenarios that involve intense focus on shared context. While backchanneling is instrumental, robot learners need not always respond to checking cues, which otherwise could be perceived as robotic and unnatural. In fact, a robot learner may deliberately show fewer backchannel signals to prompt a human teacher to slow down and repeat their instruction. Lastly, it is important to provide backchannel signals in a timely manner. Our results show that people tend to respond within one second. Responses that occur more than one second later may be regarded as a lack of response or as awkward [17].

**Limitations and Future Work**. Prior research has shown how people display social cues towards robots, even if they do not exhibit anthropomorphic features (e.g., [19]), and how human-inspired robot behavioral cues can facilitate human-robot interactions (e.g., [9]). However, whether people would exhibit behavioral cues and interaction patterns, as identified in this study, when teaching or learning a task from a robot requires further investigation. Future research should also examine how our findings may generalize to different manipulation tasks and real-world settings. Lastly, while we recognize that our exploration only involved a small sample size, we hope our findings can serve as an initial foundation for future work to develop computational models to enable robots to learn from and teach people effectively.

# REFERENCES


[1] Kendon A. 1967. Some functions of gaze-direction in social interaction. *Acta Psychol (Amst)* (1967), 26(1):22–63.

[2] Henny Admoni and Brian Scassellati. 2017. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction* 6, 1 (2017), 25–63.

[3] Michael Argyle and Mark Cook. 1976. Gaze and mutual gaze. (1976).

[4] Herbert H Clark. 1994. Managing problems in speaking. *Speech communication* 15, 3-4 (1994), 243–250.

[5] Herbert H Clark. 1996. *Using language*. Cambridge university press.

[6] John M Digman. 1990. Personality structure: Emergence of the five-factor model. *Annual review of psychology* 41, 1 (1990), 417–440.

[7] Susan R Fussell, Leslie D Setlock, and Elizabeth M Parker. 2003. Where do helpers look? Gaze targets during collaborative physical tasks. In *CHI'03 Extended Abstracts on Human Factors in Computing Systems*. 768–769.

[8] Susan R Fussell, Leslie D Setlock, Jie Yang, Jiazhi Ou, Elizabeth Mauer, and Adam DI Kramer. 2004. Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction* 19, 3 (2004), 273–309.

[9] Chien-Ming Huang and Bilge Mutlu. 2013. Modeling and Evaluating Narrative Gestures for Humanlike Robots.. In *Robotics: Science and Systems*. 57–64.

[10] Chien-Ming Huang and Andrea L Thomaz. 2011. Effects of responding to, initiating and ensuring joint attention in human-robot interaction. In *2011 Ro-Man*. IEEE, 65–71.

[11] Malte F Jung, Jin Joo Lee, Nick DePalma, Sigurdur O Adalgeirsson, Pamela J Hinds, and Cynthia Breazeal. 2013. Engaging robots: easing complex human-robot teamwork using backchanneling. In *Proceedings of the 2013 conference on Computer supported cooperative work*. 1555–1566.

[12] Jiazhi Ou, Lui Min Oh, Jie Yang, and Susan R Fussell. 2005. Effects of task properties, partner actions, and message content on eye gaze patterns in a collaborative task. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 231–240.

[13] Hae Won Park, Mirko Gelsomini, Jin Joo Lee, and Cynthia Breazeal. 2017. Telling stories to robots: The effect of backchanneling on a child's storytelling. In *2017 12th ACM/IEEE International Conference on Human-Robot Interaction (HRI*. IEEE, 100–108.

[14] Cañigueral R and Hamilton AFdC. 2019. The Role of Eye Gaze During Natural Social Interactions in Typical and Autistic People. *Frontiers in Psychology* (2019).

[15] Ponsleur Brett Holroyd Aaron Rich, Charles and Candace Sidner. 2010. Recognizing engagement in human-robot interaction. In *Proceedings of ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. 375–382.

[16] Allison Sauppé and Bilge Mutlu. 2014. How social cues shape task coordination and communication. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 97–108.

[17] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2008. How quickly should communication robots respond?. In *2008 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 153–160.

[18] Candace L Sidner, Christopher Lee, Cory Kidd, Neal Lesh, and Charles Rich. 2005. Explorations in engagement for humans and robots. *arXiv preprint cs/0507056* (2005).

[19] Maia Stiber and Chien-Ming Huang. 2020. Not All Errors Are Created Equal: Exploring Human Responses to Robot Errors with Varying Severity. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. 97–101.

[20] Tanya Stivers. 2008. Stance, alignment, and affiliation during storytelling: When nodding is a token of affiliation. *Research on language and social interaction* 41, 1 (2008), 31–57.

[21] Michael Tomasello et al. 1995. Joint attention as social cognition. *Joint attention: Its origins and role in development* 103130 (1995).