Auton Robot (2015) 39:313–329
DOI 10.1007/s10514-015-9461-0

Effective task training strategies for human and robot instructors

Allison Sauppé[1] · **Bilge Mutlu**[1]

Received: 18 November 2014 / Accepted: 3 July 2015 / Published online: 23 July 2015
© Springer Science+Business Media New York 2015

Abstract From teaching in labs to training for assembly, a role that robots are expected to play is to instruct their users in completing physical tasks. While instruction requires a range of capabilities, such as use of verbal and nonverbal language, a fundamental requirement for an instructional robot is to provide its students with instructions in a way that maximizes their task performance. In this paper, we present an autonomous instructional robot and investigate how different instructional strategies affect user performance and experience. Our analysis of human instructor–trainee interactions identified two key instructional strategies: (1) *grouping* instructions together and (2) *summarizing* the outcome of subsequent instructions. We implemented these strategies into a humanlike robot that autonomously instructed its users in a pipe-assembly task. To achieve autonomous instruction, we also developed a repair mechanism that enabled the robot to correct mistakes and misunderstandings. An evaluation of the instructional strategies in a human–robot interaction study showed that employing the grouping strategy resulted in faster task completion and increased rapport with the robot, although it also increased the number of task breakdowns. A comparison of our results with the human instructor–trainee interactions revealed many similarities, areas where our model for robot instructors could be improved, and the nuanced ways in which human instructors use training strategies such as summarization. Our findings offer strong implications for the design of instructional robots and directions of future research.

Keywords Repair · Interactive robot systems · Human–robot interaction · Instructional systems · Instructional strategies · Task training · Human instruction · Autonomous robot systems

1 Introduction

As robots enter instructional roles such as teaching in classrooms, training for assembly on a shop floor, and teaching medical students surgical procedures, they will need to effectively present task instructions, providing clarifications and corrections when needed, to improve task outcomes and user experience. Robots' success in instruction will depend on their effectiveness first in their use of language, including linguistic and nonverbal cues (Andrist et al. 2013; Boucher et al. 2012; Huang and Mutlu 2012; Staudte and Crocker 2009), and second in their presentation of task information, including what information they disclose at a given moment, how they present task information, and how they correct misunderstandings. This paper focuses on the latter problem of effectively presenting task information and explores how robots might adopt the strategies that human instructors use to present task information and what strategies might be most effective.

Human instructors carefully plan instructions to maximize their students' ability to integrate the material, such as first choosing a subgoal to address in a task and plan future instructions to address the chosen subgoal to help contextualize the instructions (Blaylock et al. 2003; Grosz and Kraus 1996). To aid participants in completing the step, instruc-

This is one of several papers published in *Autonomous Robots* comprising the "Special Issue on Robotics Science and Systems".

✉ Allison Sauppé
asauppe@cs.wisc.edu

Bilge Mutlu
bilge@cs.wisc.edu

[1] Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI, USA

tions are iteratively refined until they are atomic. Instructors might also engage the student in the instruction, encouraging "learning by doing" to enable the student to achieve a deeper understanding of the instructions by performing them (Alfieri et al. 2011). These discourse strategies might inform how a robot should order instructions and engage participants. While some strategies are known, formalizations suitable to implementation on a robot do not yet exist.

In addition to an effective method of delivery, task-based instruction requires instructors to monitor student understanding and progress and to provide feedback and corrections. As the instructor and student progress in the task, they may encounter *breakdowns*—misunderstandings or miscommunication concerning the task goals—that can impede task progress. Instructors need to *repair* these breakdowns by resolving such differences in understanding. Failure to repair breakdowns might lead to compounded breakdowns later in the interaction, further hindering progress. This repair is often context-specific in that it requires knowledge of prior actions and current expectations in order to succeed. Additionally, humans use a variety of techniques to repair breakdowns (Hirst et al. 1994) and adapt their use of these techniques to the context of the interaction (Reigeluth et al. 1980).

In this paper, we build a better understanding of these instructional and repair strategies by collecting and analyzing data from human instructor–trainee pairs on task instruction. We then implement models of these strategies on an autonomous robot system that guides users through a pipe-assembly task, mimicking real-world assembly tasks in which robots are expected to participate (Fig. 1). This system enables the robot to use each of the teaching strategies employed by human instructors to provide students with task instructions and to autonomously handle repair when breakdowns arise. Using this system, we conducted an exploratory human–robot interaction study to assess the tradeoffs between different instructional strategies in measures such as the number of repairs conducted, task completion time, and user experience with the robot. We compare our results to those found from the human instructor–trainee pairs, highlighting similarities and areas for improvement in developing robot instructors. In summary, our work makes the following contributions:

1. A better understanding of human–human instruction.
2. Models for planning instructions and repairing breakdowns and their implementation in a robot system.
3. The validation of our models and their implementation in an instructional scenario and an understanding of the effectiveness of different instructional strategies.
4. A "gold standard" comparison of our models and their implementation against the human instructor–trainee data in order to identify components of the model which could be improved.
5. The demonstration of an integrated process for designing effective robot behaviors that involves modeling human behaviors, implementing the resulting model in robots, and evaluating implemented behaviors in a user study.

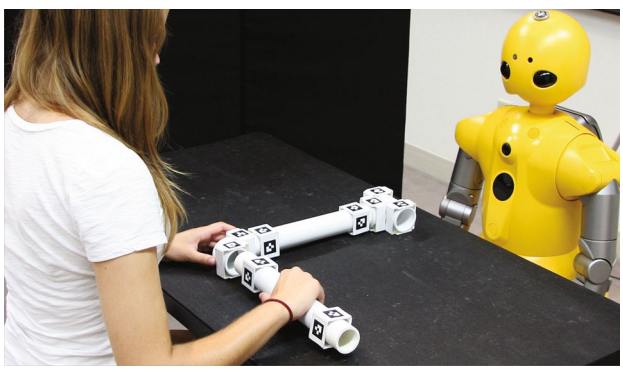


Fig. 1 The robot autonomously guiding a participant in assembling pipes

2 Background

In order to enable robots to successfully fulfill instructional roles, it is necessary to understand what instructional strategies would be best for robots to follow. We draw inspiration from how humans give task instruction to model and implement teaching strategies that maximize task outcomes and student experience in human–robot instruction. This section reviews prior work on strategies that humans use in presenting task information and on the development of instructional robots.

2.1 Instruction in human–human interaction

Effectively communicating a series of instructions is a complex task that has been studied at a number of levels, including how human instructors develop and communicate instructions for their students. Prior work has suggested that instructors follow a discourse planning process based on iterative refinement, where the instructor first picks a subgoal to complete and then further decomposes the subgoal into atomic actions (Blaylock et al. 2003; Grosz and Kraus 1996). Instructions are then ordered based on logical segmentations of steps to help students contextualize the task (Grosz and Sidner 1986). These models provide important insights into how instructors break task goals into a set of instructions.

Successfully directing a student in a task also relies on feedback from the student. Despite the best efforts of instructors, there will inevitably be instances of *breakdowns*—misunderstandings or miscommunication concerning task goals—that can either impede ongoing progress or lead to

breakdowns in the future (Zahn 1984). To correct breakdowns, humans engage in *repair*, a process that allows participants to correct misunderstandings and helps ensure that all participants have a similar understanding of the relayed information (Hirst et al. 1994; Zahn 1984). The process of engaging in repair is often context-sensitive (Seedhouse 1999). For example, when a topic is being discussed in a classroom, the instructor frequently initiates repair to clarify students' statements. However, when the classroom is engaged in a task, students are more likely to offer repair to or seek repair from their peers. Additionally, the likelihood of initiating repair can also be dependent on context. While earlier studies suggested that people have a preference for noticing and correcting mistakes on their own (Schegloff et al. 1977), more recent work indicates that a preference for others to engage in repair exists when the conversational partner is better equipped to handle the repair. Such a situation often arises in groups with disparate levels of knowledge, such as parent–child or instructor–student groups (Norrick 1991; Tomasello et al. 1990).

2.2 Instruction in human–robot interaction

Prior research in robotics has explored how robots might function in instructional settings, such as daycare facilities and classrooms (Kanda et al. 2007; Tanaka and Kimura 2009; Tanaka and Movellan 2006), and aid in task instruction, such as offering assistance in a hand washing task (Hoey et al. 2005) and giving directions in a cooking task (Torrey et al. 2007). In addition to enabling robots to accurately convey instructions, researchers have also explored how a robot's use of social cues, such as gaze and gestures, can aid students in performing a task (Huang and Mutlu 2012). To effectively achieve task goals in the range of instructional settings robots are expected to participate in, robots might adapt their instructions to accommodate the user's task expertise and recall of task steps. For instance, Torrey et al. (2006) explored how adapting the comprehensiveness of the robot's instructions to its user's expertise might affect task outcomes and user experience. They found that more comprehensive instructions resulted in fewer mistakes among novices, while experts rated the robot as more effective, more authoritative, and less patronizing when it provided brief descriptions. Foster et al. (2009) studied the effects of the order in which the robot provided task goals along with instructions on student recall of task steps, showing that providing task goals prior to issuing task steps resulted in fewer requests for repetition by the student later in the task.

Just as repair is necessary in human instruction, robots must also be capable of identifying breakdowns and offering repair for effective human–robot instruction. Prior work has explored a variety of techniques to alleviate the need for repair, such as taking into account the speaker's perspective (Trafton et al. 2005) or mitigating the negative impact of breakdowns through framing (Lee et al. 2010). While these studies point to instructional and repair strategies as key elements of the design of instructional robots, enabling robots to use strategies that maximize task outcomes and student experience requires a better understanding and models of effective task instruction. The following section details our work on developing such models.

3 Modeling

To better understand human teaching strategies, we collected video data of human–human interactions during an instructional pipe-assembly task that resembled assembly tasks in which robots might guide humans, such as furniture assembly. Below, we discuss our data collection process, analysis, and the instruction models we constructed from the data.

3.1 Data collection

We collected video data from eight instructor–trainee dyads during a pipe-assembly task. In each of these interactions, one participant (the instructor) first learned how to connect a set of pipes into a particular formation from a pre-recorded video. Instructors were given as much time as necessary to re-watch the video and were provided use of the pipes during training. Upon learning the instructions, the instructor trained the second participant (the trainee) on how to correctly assemble the pipes without the aid of the video (Fig. 2).

Eight males and eight females aged 18 to 44 ($M = 23.75$, $SD = 8.56$) were recruited from the local community. Each interaction was recorded by a video camera equipped with a wide-angle lens to capture the participants and the task space. The instructional portion of the task, excluding the time the first participant spent learning how to construct the pipes, ranged from 3:57 to 6:44 min ($M = 5{:}11$, $SD = 2{:}19$).

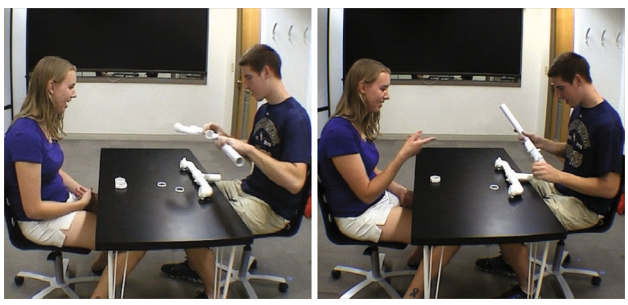


Fig. 2 The *instructor* (participant on the *left*) directing the *student* (participant on the *right*) in assembling a predetermined pipe configuration

3.2 Analysis

The analysis of our data helped us to better understand different strategies instructors use to deliver instructions and confirmed examples for our understanding of repair gained from the literature. In our data, we observed instructors organizing their instructions along two major factors: how many instructions they gave at once, and whether or not they gave a high-level summary of what the next few instructions would accomplish. We coded our videos with these two factors, and examined the effects these factors had on two outcomes: time spent per step in the task and the number of breakdowns encountered. To ensure reliability of the coding, a second coder analyzed the videos, with the inter-rater reliability showing substantial agreement between the primary and secondary coders (79 % agreement, Cohen's $\kappa = .74$) (Landis and Koch 1977). The coding resulted in 104 data points, including 54 data points with no grouping or summarization from eight instructors (71 % agreement), 32 data points with grouping and no summarization from another eight instructors (72 % agreement), nine data points with no grouping and summarization from six instructors (78 % agreement), and nine data points with grouping and summarization from five instructors (89 % agreement).

Further analysis of the data involved a two-way analysis of variance (ANOVA), including grouping, summarization, and the interaction between them as fixed-effect factors. For main and interaction effects, we used α levels of .050 and .10 for significant and marginal effects, respectively. We conducted four contrast tests to understand the effects of each factor in the absence or presence of the other factor. These contrast tests were evaluated using a Bonferroni-adjusted α level of .0125 (.05/4) for significance. Results from this analysis showed that grouping significantly reduced time spent per step, $F(1, 4.18) = 17.61$, $p = .013$, $\eta^2 = .808$, while significantly increasing the number of breakdowns, $F(1, 6.09) = 7.96$, $p = .03$, $\eta^2 = .567$. Summarization had only a marginal effect on time spent per step, $F(1, 22.66) = 3.89$, $p = .061$, $\eta^2 = .147$, and no effect on the number of breakdowns, $F(1, 5.67) = .707$, $p = .434$, $\eta^2 = .111$. Contrast tests across conditions showed that, when the instructor did not provide a summary, grouping instructions significantly reduced the time spent per step, $F(1, 8.86) = 21.97$, $p = .001$, $\eta^2 = .713$, but had no effect on breakdowns, $F(1, 7.87) = 4.25$, $p = .074$, $\eta^2 = .351$.

In addition to insights gained into how instructors organized their instructions, our analysis showed that instructors always initiated repair verbally when a breakdown occurred, regardless of whether they became aware of the breakdown verbally, such as a question by the trainee, or visually, such as noticing that the task space was not configured correctly. We found that 65 % of these repairs were *trainee-initiated*, while 35 % of repairs were *instructor-initiated*.

Trainee-initiated repair—also called *requests*—always involved verbal statements that clarified or confirmed instructor expectations when the trainee either did not understand or misunderstood an instruction. These statements ranged from brief queries (e.g., "What?") to more detailed requests, such as "Where should the pipe go?" Consistent with prior work that associated confusion with not understanding and clarification with misunderstanding (Gonsior et al. 2010; Hirst et al. 1994; Koulouri and Lauria 2009), we classified requests into the categories *confusion*, *confirmation*, and *clarification*.

Where trainee-initiated repair was directed towards better understanding expectations, instructor-initiated repair clarified or corrected the trainee's perceptions of the task. Instructors initiated repair under one of two circumstances: *mistake detection* and *hesitancy*. When instructors noticed the trainee performing an action that the instructor knew not to be consistent with the goals of that instruction, such as picking up the wrong piece, they verbally corrected the trainee. When instructors noticed that the trainee was hesitating to take action, which was indicated by an average delay of 9.84 s in following an instruction, they asked if the trainee needed help.

3.3 Model

Our analysis informed the development of a model with two components: *instructional strategies* and *repair*.

3.3.1 Instructional strategies

As noted in our analysis, instructor strategies for organizing instructions involved two factors: grouping and summarization. In *grouping*, instructors vary the number of instructions given from $1 \ldots i$ before the student completes the instructions. Instructors may provide one instruction at a time and allow the student to carry it out before providing the next instruction or offer grouped instructions by conveying i instructions, given that $i > 1$, prior to the student fulfilling the instructions. When instructors provide *instruction summarization*, they preface their instructions with a high-level summary of the goal of the subsequent k instructions. For example, when the next four steps will result in a set of pipes forming a U-shape, the instructor may say "Now, we'll be taking a few pipes and connecting them into a U-shape" prior to giving the first step. While we categorized instructional strategies into the grouping and summarization factors, our analysis demonstrated that all four possible combinations of these factors were exhibited, as illustrated in Table 1. Algorithm 1 outlines how the robot integrated summarization and grouping in its instructions.

Table 1 Examples of how the two strategies that we identified in our modeling study, *instruction grouping* and *instruction summarization*, can be jointly used in instruction

Instruction summarization	Instruction grouping	
	Not grouped	Grouped
Not summarized	Instructor: Now take this [*points toward pipe*] and just attach it like that [*makes connecting motion*] <student acts>. Then take this one [*points toward joint*] and put it here. <student acts>	Instructor: You will now connect these two and then connect them to this piece [*points toward piece*] so they will be pointing straight up. <student acts>
Summarized	Instructor: So you are going to use these two to connect them in and form a U-shape. So take one of these [*points toward pipe*] <student acts>, and then one of those [*points toward washer*] <student acts>, and you will want the skinny side facing out. <student acts>	Instructor: OK and you want to start with one arm. So the arms are going to screw onto the smooth side, so they will go onto the top of the t-piece. So you are going to want to take a washer first, and you will want to put the fat side towards the curve of the washer and then put the washer on top of that, and then put the t-piece there. <student acts>

3.3.2 Repair

Regardless of the instructional strategy utilized, we observed instructors engage in three forms of repair: *requests*, *hesitancy*, and *mistake detection*. Below, we describe these behaviors and present model components for determining whether repair is needed and, if so, how it might be performed.

Requests: All trainee requests, including questions and statements, were considered as requests for repair. To enable the model to determine the appropriate response, we classified requests into semantic categories using semantic-language modeling. For example, "Which piece do I need?" and "What piece should I get?" were recognized as the same question.

Algorithm 1 Pseudocode for how the robot integrated grouping and summarization into its instruction-giving.

```
current ← x
bool summarize?
bool grouping?
if summarize? then
    summarize(current, current + k)
end if
for y ← 0; y < i do
    instruction(current)
   if !grouping? then
    action(current)
   end if
    y ← y + 1
    current ← current + 1
end for
if grouping? then
   for z ← 0; z < i do
    action(z)
    z ← z + 1
   end for
end if
```

Hesitancy: Depending on the task, indicators such as time elapsed since the last interaction or time elapsed since the workspace was last changed can signal hesitancy in performing instructions. For the pipe-assembly task, we chose to use the time elapsed since the workspace was last changed as a conservative predictor of hesitancy-based breakdowns, as using time elapsed since the last interaction could result in incorrectly inferring hesitancy while the trainee is still working. Based on our observations of how long human instructors waited before offering repair, we considered 10 s of no change to the workspace to indicate a hesitancy-based breakdown.

Mistake detection: While requests and hesitancy-based breakdowns are triggered by the student's action or inaction, mistake detection requires checking the student's work. In our proposed model, we chose a simulation-theoretic approach to direct the robot's behavior in relation to the participant. This approach posits that humans represent the mental states of others by adopting their partner's perspective to better understand the partner's beliefs and goals (Gallese and Goldman 1998; Gray et al. 2005). This approach has been used in designing robot behaviors and control architectures to allow robots to consider their human partner's perspective (Bicho et al. 2011; Nicolescu and Mataric 2003). In the context of an instructional task, the instructor has a mental model of an action that they wish to convey to the trainee. Following instruction, the instructor can assess gaps in the trainee's understanding or performance by comparing the trainee's actions to their mental model of the intended action and noting the differences that occur.

The majority of the breakdowns we observed were highly dependent on the ongoing context and thus could be unique to assembly tasks. For example, if a participant selected the wrong piece for the structure, an instructor may reiterate or rephrase the description for the correct piece. This response to the breakdown may only be suited to tasks that can be broken down into individual components. Only requests for

repetition were independent of task context. To ensure that our model is applicable to other task types, we included an abstraction that allows designers to implement the algorithm according to their own task needs (e.g., breakdowns that are unique to their task).

Following the simulation-theoretic approach, we defined a set of instruction goals $P = \{p_1, \ldots, p_n\}$ for the robot regarding the result of the participant's action or inaction given the current instruction. Depending on the task, P may vary at each step of the instruction, as some instruction goals may no longer be applicable, while others may become applicable. As the participant engages in the task, the robot will evaluate whether the current state of the workspace is identical to the set of instruction goals P^*. If any of the individual task goals p_k do not match p_k^*, then there is a need for repair.

How repair is carried out depends on which task goal p_k has been violated. As we observed in our analysis of the human–human interactions, the instructor repaired only the part of the instruction that was currently incorrect. Additionally, there is an inherent ordering to the set P that is informed by the participant's perception of the task. The participant's ordering of P is informed by *elaboration theory*, which states that people order their instructions based on what they perceive as being the most important and then reveal lower levels of detail as necessary (Reigeluth et al. 1980). By imposing an ordering of decreasing importance on the set P based on these principles for a given task, we can ensure that each p_k takes precedence over any p_{k+n} for $n > 0$. If multiple p_k are violated, then the task goal with the lowest k is addressed first. An example of this ordering can be seen if a participant has picked up the wrong piece and attached it in the wrong location. The instructor first repairs the type of piece needed and then the location of that piece.

Although we discuss the model for detecting mistakes in terms of task steps and goals, this model can also be extended to understanding and repairing verbal mistakes. For example, if the participant mishears a question and responds in a way that is inconsistent with the answers expected, then repair is needed. The appropriate answers of the intended question can be formalized as p_k, and any answer that does not fulfill p_k can be considered as a cause for repair.

4 System

To create an autonomous system that implements our models, we contextualized our task in the same scenario used for modeling human–human interactions. Using our findings from the previous stage, we designed our system to—upon user request—process both verbal and visual information in order to check the participant's workspace and to detect and repair breakdowns.

4.1 Hardware

We implemented our model on a Wakamaru humanoid robot (Fig. 1). Our model uses information provided by both video and audio using a Microsoft Kinect stereo camera and microphone-array sensor. Video is captured at 12 frames/second. The camera and microphone were suspended three feet above the participant's workspace, as shown in Fig. 5. This camera setup provided a visible range of the workspace of 43 in. by 24 in. A second stereo camera was placed behind the robot to track the participant's body and face.

4.2 Architecture

The architecture for our model involved four modules: *vision*, *listening*, *dialogue*, and *control*. The vision and listening modules capture and process their respective input channels. The control module uses input from these modules to decide the need for repair and relays the status of the workspace to the dialogue module if feedback from the robot is needed.

The pipe-assembly task used in our implementation involved multiple copies of five types of pieces: three types of pipes (short, medium, and long) and two types of joints (elbow and t-joints). In order for the workspace camera to identify these pieces, we used eight unique augmented reality (AR) tags: two tags for the elbow joints, three tags for the t-joints, and one tag each for the short, medium, and long pipes. The orientation of each tag was used to identify object type, location, and rotation. The location and orientation of tags on pipes and joints were consistent across each type of object, and tag locations on each object were known to the system.

In our model, we defined a set P that describes which possible expectations can be violated by the participant. Consistent with elaboration theory, our study of human instructor–trainee interactions revealed the following ordering of task expectations:

- *Timely action* (p_0): The participant acted in a timely fashion.
- *Correct piece* (p_1): The participant used the correct piece.
- *Correct placement* (p_2): The participant placed the piece in the correct location relative to the current workspace.
- *Correct rotation* (p_3): The participant rotated the piece correctly relative to the current workspace.

The first expectation, p_0, ensures that the participant does not hesitate for too long—which might indicate confusion—when adding the next piece. Based on our previous analysis, we considered a 10 s delay in changing the workspace after the last instruction to indicate hesitancy. The remaining expectations, p_1, p_2, and p_3, ensure that the participant

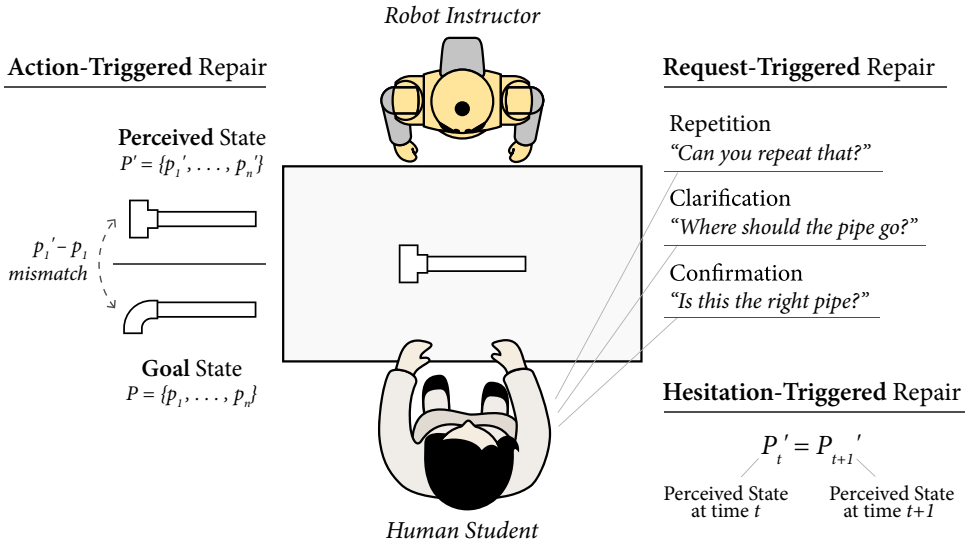


Fig. 3 Examples of the three types of repair. In *action-triggered repair*, the student's configuration of pieces does not match what the robot knows to be the correct configuration. *Request-triggered repair* is initiated when the student directs a question or statement to the robot that requires the robot to respond appropriately. In *hesitation-triggered repair*, the workspace remains unchanged for more than 10 s, prompting the robot to offer assistance

chooses the correct piece to add, adds the piece in the correct location, and rotates the piece correctly. Figure 3 illustrates these expectations.

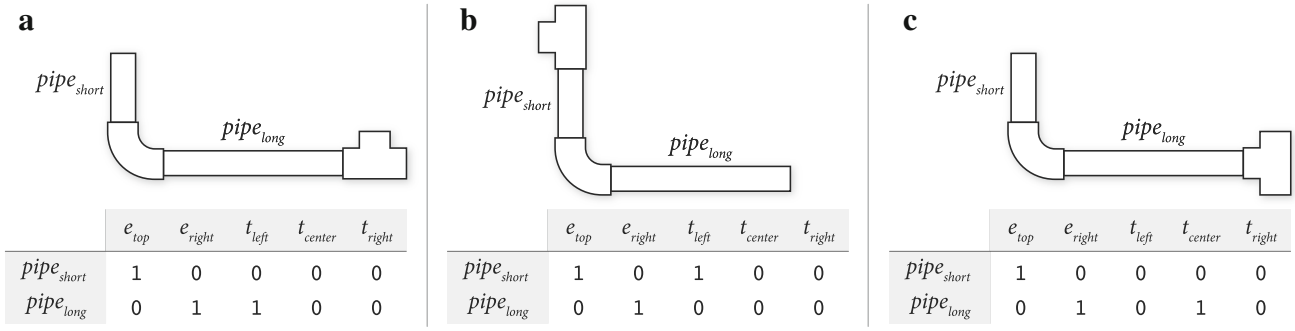
4.2.1 Vision module

The vision module was designed to achieve two goals: to detect the status of the participant's workspace and to process information on the participant's location. Sensing necessary for achieving each of these goals was managed by a separate camera.

To achieve the first goal, the vision module builds a graph of pipe connections, C, following a three-step process: finding the AR-tag glyphs in the frame, associating these glyphs with pieces, and detecting which pieces are connected based on a set of heuristics. In the first step, at the completion of the participant's turn, the frame is searched for AR glyphs using a modified version of Gratf[1] to create a set of glyphs G, where each glyph in G is defined by its type t, its position (x, y), and its rotation θ. Upon discovering a glyph, the algorithm searches known pieces of which type t belongs (i.e., if the glyph belongs to a t-joint, all t-joints are searched) for any pieces that are missing that particular glyph. The glyph is associated with a piece if the algorithm matches the glyph to the piece based on its proximity and rotation properties. If no piece is found, a new piece is created, and the glyph is associated with the new piece. This process results in a set of pieces P where each piece p is characterized by a set of glyphs that are associated with that piece. All of the glyphs for a piece p form a bounding box that gives a rough estimate of the physical boundaries of that piece. Using these coordinates, we can confirm whether any two pieces are connected and subsequently build a graph structure C that reflects the workspace. This structure is organized such that each pipe is represented by a row, and each joint is represented by a set of columns, each of which corresponds to a place on the joint where a pipe can be connected. For example, a t-joint, which has three connectors, would be represented by three columns. When a pipe and a joint are connected, a `1` is placed in the cell at the intersection of that pipe's row and the joint's column for that particular connector. The remaining cells would be `0`s in order to express no connection. Figure 4a illustrates an example matrix, C^*, that represents the correct graph structure of two pipes (short and long), an elbow joint (e_{top} and e_{right} for the top and right connector, respectively), and a t-joint (t_{left}, t_{center} and t_{right} for the left, top, and right connector, respectively).

When the user completes a turn, the correct graph structure C^* is compared against the structure C of the workspace. If the two graphs are isomorphic, then the user has successfully completed the instruction. If the graphs are not isomorphic, however, the robot will discover an inconsistency between p_k and p_k^* during the isomorphism check. The lowest p_k^* which is violated is then passed to the control module. The system determines the lowest p_k^* violated using two pieces of information: the number of each type of piece present in the workspace and a comparison of C against C^*. Counting each type of piece and comparing these counts to the expected counts is used to check violations of p_1. If the check for p_1 passes, then the system continues to determine the lowest p_k^* between p_2 and p_3 that was violated. To identify which p_k^* were violated, the isomorphism check continually saves the observed graph C' that contains the fewest number of errors when compared to the correct graph structure C^*. At the end of the isomorphism check, the C' that was the closest permutation to C^* encountered is considered as C. When comparing C to C^*, a violation of p_2 is observed when a

[1] Gratf: http://www.aforgenet.com/projects/gratf/.



a	e_{top}	e_{right}	t_{left}	t_{center}	t_{right}
$pipe_{short}$	1	0	0	0	0
$pipe_{long}$	0	1	1	0	0

b	e_{top}	e_{right}	t_{left}	t_{center}	t_{right}
$pipe_{short}$	1	0	1	0	0
$pipe_{long}$	0	1	0	0	0

c	e_{top}	e_{right}	t_{left}	t_{center}	t_{right}
$pipe_{short}$	1	0	0	0	0
$pipe_{long}$	0	1	0	1	0

Fig. 4 Illustrations of and graph structures that represent three different pipe configurations that include two pipes (short and long), an elbow joint (e_{top} and e_{right} for the top and right connector, respectively), and a t-joint (t_{left}, t_{center} and t_{right} for the left, top, and right connector, respectively)

1 is misplaced either across rows, indicating that the joint is attached to the wrong pipe, or across sets of columns, indicating that the pipe is connected to the wrong joint. For example, in the matrix C in Fig. 4b, because the t-joint is connected to $pipe_{short}$ instead of $pipe_{long}$, the 1 indicates that the connection of the t-joint is in the wrong row for this C.

A violation of p_3 is observed when the 1 is in the correct set of columns for that particular joint but is in the wrong column within the set of columns, indicating that the joint is in the correct place but rotated incorrectly. In the example C in Fig. 4c, the t-joint is connected to the correct pipe but is rotated incorrectly, connecting at the center connector instead of the left connector. As a result, the 1 is placed in the correct set of columns, i.e., in one of the t-joint columns, but is in the wrong column for that set, i.e., t_{center} instead of t_{left}. Note that this rule is true both for adding a pipe to the incorrect connector of a joint or for adding an incorrectly rotated joint to a pipe.

If the system needs to check multiple instructions at once, the set of pipe connections C is built incrementally, starting with the first instruction that needs to be checked. Because each instruction involves the addition of a new piece to a specific location and with a particular rotation, checking the workspace for the first instruction s_1 will result in the detection of too many pieces, as pieces for instructions through s_n are also on the table. In this case, the module is responsible for systematically eliminating extraneous pieces from C. A piece is defined as extraneous if its removal does not result in a disjoint graph in C and does not reduce the count of that particular piece below what is needed to complete the instruction. Once a modified version of C that results in a correct check of s_1 is found, pieces are added incrementally back to C such that they maintain connectivity between all pieces in C and maintain a set P that is equivalent to the number of each type of piece needed to complete the instruction s_m.

The second goal of the vision module—detecting the participant's location—is checked at every frame. When the participant is within 1 ft. of the workspace, the robot repositions its head so that it is gazing at the table, monitoring the workspace. When the participant is further away (e.g., standing back to check their work, retrieving the piece), the robot raises its head and gazes toward the participant's face. However, if the participant or the robot is talking, or if the robot is checking the workspace in response to a prompt from the user, the robot looks toward the participant or where on the workspace changes have been made, respectively.

4.2.2 Listening module

The listening module detects and categorizes requests from the participant into semantic meanings using the capabilities of the Microsoft Kinect sensor and speech-recognition API. We provided the API with a grammar that included speech acts from our data on human–human instruction that we marked as one of the following semantic meanings:

- *Request for repetition*: (e.g., "What did you say?" "Can you repeat the instructions?")
- *Check for correctness*: (e.g., "Is this the right piece?" "I'm done attaching the pipe.")
- *Check for options*: (e.g., "Which pipe do I need?" "Where does it go?")

Utterances that did not belong to one of these categories, such as confirmation of an instruction, were ignored by the system.

We use a dialogue manager to coordinate responses to each type of query. Each recognized utterance has an associated semantic meaning that indicates the purpose of the utterance. For example, the phrase "What did you say?" is assigned the semantic meaning of "recognition request." These semantic meanings allow the control module to understand the type of utterance processed and to reply to the utterance appropriately given the current state of the participant's workspace. To process requests that refer to the workspace, the system first checks the state of the workspace through the vision module.

For example, asking "Did I do this right?" requires the robot to determine whether the current workspace is correct.

4.2.3 Control module

Decisions regarding the robot's next actions are determined by the control module. It uses input from the vision and dialogue modules and, following a simulation-theoretic approach, makes decisions by comparing this input to actions that the robot expects in response to its instructions. At a high level, the robot provides instruction(s) according to its teaching strategy, waits for user input, either in the form of a question about the instruction(s) or a request for the robot to provide feedback, and then responds to the user's input.

When the robot provides instruction(s), the control module takes into consideration whether summarization and/or grouping will be employed. When neither strategy is used, the robot provides a single instruction and then waits for user input. If using grouping, the robot provides a predetermined number of instructions that range between two and four. When using summarization, the robot provides a summary of the goal of a predefined number of instructions prior to giving these instruction.

The above process is outlined in Algorithm 2. The *state* variable denotes the robot's current state. This variable is frequently updated by the listening module, which provides information on the semantic meaning of the user's speech, and occasionally by the control module. For example, "Could you repeat that?" would transition the *state* variable to "repetitionRequest." For brevity, Algorithm 2 illustrates how a state would be executed under the grouping and/or summarization conditions only for the first state "instruction."

Algorithm 2 Pseudocode for controlling the robot to give instructions, check the workspace, or answer questions from the user

```
bool summarize?
bool grouping?
int stepNum ← 0
String state ← "instruction"
Step[] instructions
if state == "instruction" then
    if summarization? and
        instructions[stepNum].summary != null then
     give instructions[stepNum].summary
    end if
    int count ← 1
    if grouping? then
       int count ← instructions[stepNum].groupingCount
    end if
    for i ← 0; i < count; i++ do
     give instructions[stepNum].instruction
     stepNum++
    end for
     state ← "wait"
else if state == "checkWorkspace" then
    enum visual ← checkWorkspace()
    if visual == "correct" then
     give correctAnswer
     stepNum++
     state ← "instruction"
    else if visual == "wrongPiece" then
     give instructions[stepNum].pieceNeeded
     state ← "wait"
    else if visual == "wrongLocation" then
     give instructions[stepNum].locationNeeded
     state ← "wait"
    else if visual == "wrongRotation" then
     give instructions[stepNum].rotationNeeded
     state ← "wait"
    end if
else if state == "pieceQuestion" then
     give instructions[stepNum].pieceNeeded
     state ← "wait"
else if state == "locationQuestion" then
     give instructions[stepNum].locationNeeded
     state ← "wait"
else if state == "rotationQuestion" then
     give instructions[stepNum].rotationNeeded
     state ← "wait"
else if state == "repetitionRequest" then
     give instructions[stepNum].instruction
     state ← "wait"
end if
```

4.2.4 Dialogue module

After evaluating input from the vision and listening modules, the control module passes three pieces of information to the dialogue module: current instruction, the semantics associated with the speaker's last utterance (if any), and the control module's evaluation of the workspace (if any).

Given this information, the dialogue module initiates the appropriate verbal response, choosing from among predefined dialogue acts based on which task instruction the participant is completing, the current layout of the workspace, and the type of question the participant asked. Not all responses depend on all three pieces of information; for example, requests for repetition of the last instruction are independent of how the workspace is currently configured, and responses to hesitancy are independent of the current workspace and interaction with the participant. However, a request to check if an instruction has been correctly completed requires knowledge of both the instructions completed and the current layout of the workspace.

5 Evaluation

To evaluate the effectiveness of the strategies that we identified from our analysis in human–robot instruction, we

conducted a study that followed the same task setup as our modeling study. Due to a lack of sufficient theory that would predict the effects of these instructional strategies on trainee performance and experience, we chose not to pose any hypotheses and performed an exploratory evaluation. Our analysis aims to provide guidelines as to how future instructional robots should be designed and highlight differences and similarities between the human–human and human–robot instruction.

5.1 Study design

To assess the effectiveness of and tradeoffs between various teaching strategies, we designed a between-participants-design study to compare four different models of teaching strategies that fell along two factors: *grouping* and *summarization*. Grouping defines how many instructions are issued during the instructor's turn. For the purposes of our study, grouping has two levels: *no grouping*, where a single instruction is given during the round, and *grouping*, where a set of two or more instructions are given at once. Summarization defines whether or not the instructor gives a summary of the objective of the next few instructions. In our study, we created two levels of this factor: *no summarization*, where the instructor does not give summaries, and *summarization*, where the instructor offers summaries. We observed the instructor–trainee pairs in our modeling study to exhibit all four combinations of these two factors and created four conditions for our study: (1) *no grouping, no summarization*, (2) *grouping, no summarization*, (3) *no grouping, summarization*, and (4) *grouping, summarization*.

The architecture detailed in the previous section was used in all conditions. Differences between conditions were controlled in the control module that managed decisions on how to structure instructions. Additionally, the dialogue module responded to requests in the grouping level that did not exist in the no grouping level (e.g., repeating multiple instructions).

5.2 Task

All participants were autonomously guided through assembling a set of pipes by the robot in the setup shown in Fig. 5. Participants were given two bins—one for pipes and one for joints—that contained only the pieces necessary for completing the task, mimicking the setup in which different types of parts might be kept at a workshop. Following an introduction, the robot directed the participant in the assembly task by issuing instructions according to the condition to which the participant was assigned, varying the number of instructions provided and whether or not high-level summaries of future instructions were provided. The robot also provided repair as necessary. Following completion of the task, the robot thanked the participant. Completing the task took between 3:57 and 9:20 min ($M = 6{:}44$, $SD = 1{:}23$).

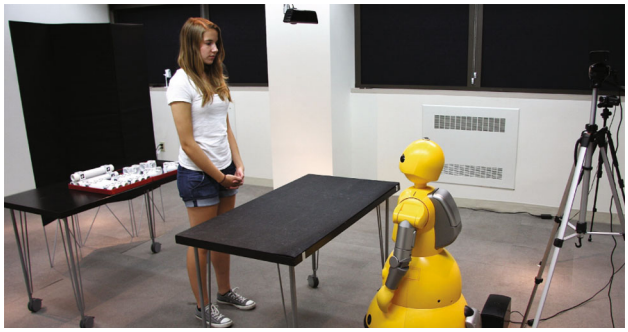


Fig. 5 The setup used in our experimental evaluation. After the robot gave an instruction, the participant retrieved the necessary pieces from behind them and assembled the pieces on the workspace in front of the robot. A camera above the workspace captured the configuration of the pieces

In the *no grouping, no summarization* and *no grouping, summarization* conditions, the robot provided one instruction at a time. Instructions in the grouping condition involved two-to-four instructions at a time, based on the average number of instructions ($M = 3.1$) human instructors gave when they employed grouping. Following grouped instructions always involved assembling spatially connected pieces, such that each instruction in the group asked the user to attach the new piece to the previously added piece. In the *no grouping, summarization* and the *grouping, summarization* conditions, the robot provided a high-level summary of the next few steps prior to giving instructions, while it provided no summary in the other conditions. Summaries were provided at the beginning of each set of grouped instructions, regardless of whether or not grouping was employed, in order to ensure that summarization and grouping applied to the same set of instructions in the *grouping, summarization* condition.

Following instructions, the participant retrieved the pieces to complete the steps and assembled the pieces on the table. If the participant requested repetition or clarification, the robot answered. When the participant asked the robot to check the workspace, it confirmed correct actions or provided repair according to our model. If no repair was needed, it congratulated the participant on completing the task and proceeded to the next instruction or set of instructions.

The resulting pipe-structure included a total of 15 connected pipes and joints. While the resulting structure was a tree that had no cycles, it had no predefined "root" piece, making the computational complexity of checking for isomorphism against the correct structure an NP-hard problem. We significantly reduced the runtime of this operation by exploiting domain knowledge in our data structure in the form of an incidence matrix of connected joints versus pipes. Once all the pipes were connected, checking for graph isomorphism required approximately 10K permutations of

the incidence matrix—far fewer than the hundreds of trillions of checks required without knowledge of the incidence matrix.

Participants started the study standing three feet away from the robot, with a two foot long table between them. A second table was placed five feet behind where the participant started. A single video camera captured the entire interaction for additional data analysis.

5.3 Procedure

Following informed consent, participants were guided into the experiment room. The experimenter explained the task and introduced the participant to the pieces used in the task. After the experimenter exited the room, the robot started the interaction by explaining that it would provide step-by-step instructions for assembling the pipes. The robot then provided instructions until the participant completed the entire structure. At the end of the task, the robot thanked the participant. The participant then completed a questionnaire and received $5.

5.4 Participants

A total of 32 native English speakers between the ages of 18 and 34 ($M = 23, SD = 4.9$) were recruited from the local community. These participants had backgrounds in a range of occupations and majors. All conditions were gender balanced.

5.5 Measures & analysis

We used two objective measures to evaluate participant performance in the task: number of breakdowns and time spent per step. Number of breakdowns was defined as the number of times the participant made a mistake in fulfilling an instruction or asked for repetition or clarification of the instruction. We also measured the time spent per step, expecting a lower number of repairs to indicate a faster time spent per step. These measures were coded from video recordings of the trials. To ensure reliability of the measures, a second experimenter coded for repairs. The inter-rater reliability showed substantial agreement (87 % agreement, Cohen's $\kappa = .83$) (Landis and Koch 1977).

We also used subjective measures that collected data on the participant's impressions of the robot, including likability, naturalness, and competency, the participant's experience with the task, and their rapport with the robot. Participants rated each item in our scales using a seven-point rating scale. A confirmatory factor analysis showed high reliability for all scales, including the likability (10 items, Cronbach's $\alpha = .846$), naturalness (6 items, Cronbach's $\alpha = .842$), competency of the robot (8 items, Cronbach's $\alpha = .896$), participant experience (8 items, Cronbach's $\alpha = .886$), and rapport with the robot (6 items, Cronbach's $\alpha = .809$).

Our analysis of data from these measures involved a two-way analysis of variance (ANOVA), including grouping, summarization, and the interaction between them as fixed-effect factors. For main and interaction effects, we used α levels of .050 and .10 for significant and marginal effects, respectively. We conducted four contrast tests to understand the effects of each factor in the absence or presence of the other factor using a Bonferroni-adjusted α level of .0125 (.05/4) for significance.

5.6 Results

We primarily report marginal and significant effects of the instructional strategies used by the robot on objective and subjective measures and summarize them in Fig. 6.

To ensure that possible errors in the robot's autonomous behavior did not introduce any biases to our data that would jeopardize our ability to distinguish the differential effects of the teaching strategies that the robot used, we examined video recordings of the study for mistakes by the system. Our criteria for removing data included (1) whether or not the robot offered incorrect instruction or repair and (2) whether or not the robot failed more than once to understand a single speech act by the participant. Our examination found no instances of system error regarding the configuration of the pipes in the instructions it gave or the repair it offered, indicating no instances of an incorrect instruction or repair. While the robot failed to understand 21 % of the participants at least once during their entire interaction, no single speech-act was misunderstood more than once, as participants either more clearly reiterated or rephrased their statement. We included the data from this second speech act in our analysis.

To evaluate the effectiveness of the instructional strategies, we measured the number of breakdowns that occurred during the task and the time taken to complete the task.

All of the 13 instances of repair where the participant asked the robot a question, such as "What piece do I need," involved the participant asking the robot to repeat the instruction. The analysis of the data showed that grouping instructions significantly reduced the time spent per step, $F(1, 28) = 13.35$, $p = .001$, $\eta^2 = .313$, while significantly increasing the number of breakdowns, $F(1, 28) = 8.87$, $p = .006$, $\eta^2 = .213$. Summarization had no overall effect on the time spent per step, $F(1, 28) = 0.07$, $p = .793$, $\eta^2 = .002$, or the number of breakdowns, $F(1, 28) = 1.25$, $p = .274$, $\eta^2 = .030$. The analysis also showed a marginal interaction effect between grouping and summarization over the number of breakdowns, $F(1, 28) = 3.47$, $p = .073$, $\eta^2 = .083$, but no interaction effects were found over the time spent per step, $F(1, 28) = 1.29$, $p = .266$, $\eta^2 = .030$. Con-

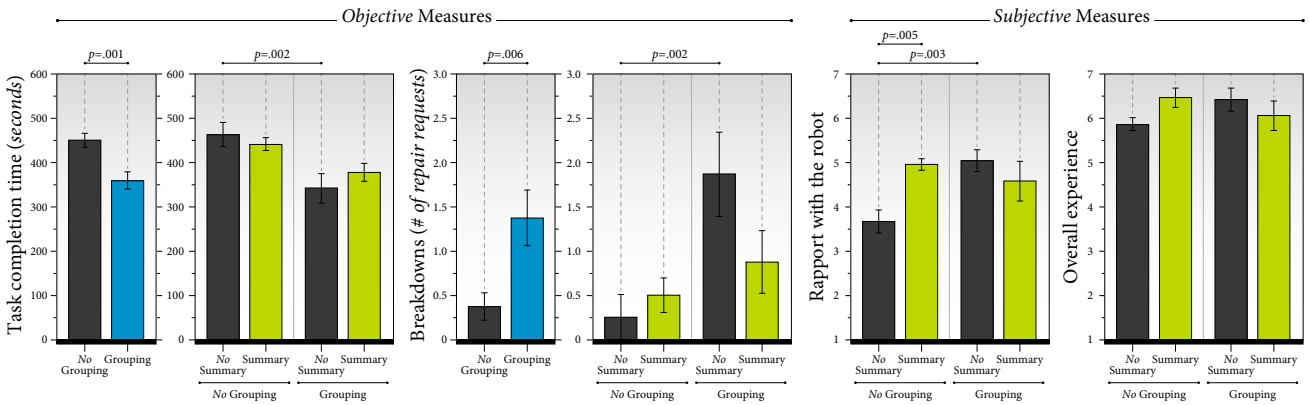


Fig. 6 Results from our evaluation. Significant and marginal results were found for total task time, number of breakdowns encountered, participants' perceived rapport with the robot, and their overall experience with the task

trast tests across conditions showed that, when the robot did not provide a summary, grouping instructions significantly reduced the time spent per step, $F(1, 28) = 11.47$, $p = .002$, $\eta^2 = 269$, but resulted in a significant increase in the number of breakdowns, $F(1, 28) = 11.71$, $p = .002$, $\eta^2 = .282$. This increase was alleviated to some extent by summarization, as participants encountered noticeably fewer breakdowns when the robot also provided a summary along with grouping, $F(1, 28) = 4.44$, $p = .044$, $\eta^2 = .107$, although this effect was not significant at $\alpha = .0125$.

The subjective measures captured the participants' perceptions of the robot, including likability, naturalness, and competency, their rapport with the robot, and their overall experience with the task. The analysis showed an interaction effect between grouping and summarization over the participants' rapport with the robot, $F(1, 28) = 8.76$, $p = .006$, $\eta^2 = .211$. When the robot provided no summary, grouping instructions improved participant rapport with the robot, $F(1, 28) = 10.81$, $p = .003$, $\eta^2 = .260$. When the instructions were not grouped, summarization also improved rapport with the robot, $F(1, 28) = 9.54$, $p = .005$, $\eta^2 = .230$. Consistent with the results on participant rapport, we also found a marginal interaction effect between grouping and summarization over participants' ratings of their overall experience with the task, $F(1, 28) = 3.68$, $p = .065$, $\eta^2 = .115$. Contrast tests showed that when the robot did not group its instructions summarization resulted in an improvement in participants' overall task experience, $F(1, 28) = 2.91$, $p = .099$, $\eta^2 = .091$, although this effect was also not significant at $\alpha = .0125$.

6 Discussion

The data from our objective and subjective results provided a number of findings to guide the design of instructional robots, the implications of which we highlight below.

Our objective results showed that grouping instructions resulted in a tradeoff between the time spent per step and the number of breakdowns that the participants encountered. We found that participants completed the steps in the task significantly faster when the robot grouped its instructions than when the robot provided instructions one-by-one. We observed that when participants received multiple instructions, they retrieved all parts necessary to complete these instructions from the bins at once, proceeded with assembling multiple pieces in a sequence, and sought confirmation of the correctness of the whole sequence from the robot, completing the overall assembly significantly faster. When participants received instructions one-by-one, they instead retrieved pieces one-by-one and proceeded to the next instruction only when the robot confirmed the successful completion of an assembly, which resulted in overall longer time spent per step. Contrary to the improvement in the time spent per step, participants encountered significantly more breakdowns when the robot grouped its instructions than when the robot provided individual instructions. We speculate that grouped instructions required participants to retain a greater amount of information, which might have impaired their understanding or recall of the instructions, resulting in mistakes in the assembly that had to be repaired by the robot.

Further analysis into breakdowns that occurred with grouped instructions showed that 60 % of breakdowns occurred in the first set of instructions, which contained four instructions, 25 % occurred in the second, third, and fifth set of instructions, which all contained three instructions, and 15 % occurred in the fourth set of instructions, which contained two instructions. This distribution of breakdowns indicates an increase in the number of breakdowns as the number of grouped instructions increases, which might indicate a greater cognitive load placed on the participant by the introduction of more pieces into an instruction (Sweller 1988). Additionally, participants may have demonstrated selective attention when the robot provided grouped instruc-

tions, causing them to miss information (Sweller 1988). Our data on the number of breakdowns provided limited support for this explanation; in carrying out grouped instructions, participants encountered fewer breakdowns when the robot provided a summary of subsequent steps ($M = 0.88, SD = 0.99$) than when no summary was provided ($M = 1.88, SD = 1.36$), although this effect was not significant at α level .0125. The summary provided by the robot might have consolidated the participants' understanding of the grouped instructions. However, some of the breakdowns that occurred early in the interaction may have been caused by the participant acclimating to the task or the task involving a greater variety of pieces to choose from at the beginning.

Our analysis of the subjective measures showed a significant interaction effect between grouping and summarizing on participant rapport with the robot. We found that participants reported higher rapport with the robot when it grouped instructions with no summary than when the robot used neither grouping nor summarization. This improvement might be due to the quicker, less monotonous experience that the robot offered when it delivered instructions all at once and spent no time on summarizing them. The results also showed that participants reported higher rapport with the robot when the robot provided a summary of subsequent steps along with individual instructions than when it neither grouped its instructions nor provided a summary. Consistent with the interaction effect on participant rapport with the robot, we also found a marginal interaction effect between grouping and summarizing on their overall experience with the task, although the contrast tests did not show significant differences at α level .0125. We speculate that, when the robot provided a summary of what was ahead in the task, as a summary involved information on upcoming steps, participants might have felt more informed and perceived the robot as more invested, although this information did not improve task performance.

6.1 Comparison to the human–human data

In this section, we compare the results of our human–robot and human–human evaluations in order (1) to build a more complete model to which future robot implementations can aspire and (1) to establish a basis to better compare and discuss findings from the human–robot study. Although differences in the specifics of the pipe-building task, the task setup, and the capabilities of the robot versus the human instructor prevents us from performing statistical tests across data sets, we are able to compare the effects identified by our analyses in the two studies as well as qualitatively examine the human–human data in order to gain additional insights.

Our comparison revealed many similarities in the results from the human–human and human–robot evaluations. In both scenarios, grouping was significant in reducing the time spent per step, but it still increased the total number of breakdowns. In both the human–human and human–robot analyses, summarization had no effect on the time spent per step. However, while summarization had no effect on the total number of breakdowns in the human–robot evaluation, summarization did marginally reduce the number of breakdowns in the human–human study.

The similarities in the findings from the two studies suggest that our system is capable of autonomously leading trainees through the pipe building task in a way that is similar to how a human instructor offers instruction. This autonomous system also includes models that enable the robot to recognize progress through each step of the task and to detect and correct breakdowns in the task or in the interaction. Additionally, results from each data set indicate that our system successfully utilizes the grouping strategy in the way that human instructors do, but that our characterization and/or implementation of summarization is not yet as effective as human instructors' use of this strategy. Summarization appears to be a more complex strategy than what our system demonstrated; we observed that human instructors offered summaries as they felt necessary during the task, while the robot instructor employed summarization in the same way for all participants. We speculate that the summaries were more effective with the human dyads due to the higher complexity of the task and the more nuanced instructions human instructors provided; human instructors effectively generate useful summarizations as needed during the task in reaction to challenges the trainee is having or based on challenges they themselves may have faced in learning the task. We speculate that human instructors were more effective in providing useful summarizations due to the higher complexity of the task, requiring them to respond to challenges that the trainee may be facing or to provide more nuanced instructions based on challenges they themselves may have faced in learning the task. Future work should examine when human instructors decide to offer summaries and how they formulate their summaries in order to inform the development of a model for deciding when and how a robot should summarize future task steps.

6.1.1 Additional insights from the human–human data

In addition to our statistical analyses, a qualitative inspection of the human–human video data revealed additional instructional strategies and mannerisms employed by human instructors. We did not incorporate these behaviors into our current system, because they either occurred infrequently in our human–human data or were difficult to replicate with the simplified pipe-building task that we developed for the human–robot evaluation. We discuss each of these behaviors and their potential for future work below.

Reference to subtasks: The pipe task used in the human–human data collection involved a number of repeated subtasks. These subtasks ranged from actions with multiple steps, such as the addition of a pipe always involving the use of a washer and nut, to repeated structural components, such as a symmetrical structure where sides mirror one another. Over the course of the task, participants developed an implicit understanding of the subtasks required to complete a step. For example, the instructor might tell the trainee to attach two pipes later in the training, and the trainee would know that the washer and nut must be attached before the pipes were joined. Future work should consider the role that subtasks play in completing the task and explore how task planning algorithms may represent subtasks in task models and utilize them in instruction.

Use of gestures: Instructors occasionally employed gestures to supplement their verbal communication with the trainee. For example, if the trainee was unsure of which piece was needed or the placement of the piece, the instructor pointed toward either the piece or the location in question. Additionally, instructors on occasion used iconic gestures—gestures that represent concrete objects or events in discourse (McNeill 1992)—when providing guidance about the orientation or placement of a pipe. For example, if a trainee was unsure of how a pipe should be oriented relative to the existing structure and was currently holding the pipe vertically, the instructor shaped their hand in a way that mimicked the orientation of the piece that trainee was using and then rotated their hand to indicate the correct orientation, similar to the "poising" gesture described by Clark (2005). Our prior work explored the rich space for the use of pointing gestures under different environmental conditions and demonstrated that pointing improves common ground in human–robot interaction in task settings (Sauppé and Mutlu 2014b). Future work should consider when gestures may be beneficial during instruction and how appropriate iconic gestures can be used given the task and communicative goals.

Detection of intent: During the task, we observed instructors to preemptively offer additional instruction or correction to the trainee on the current step. For example, when the trainee appeared confused about which piece to acquire next, the instructor gestured toward the needed piece. Prior work on human–robot interaction explored how human social cues, such as gaze, can be used by partners to infer cognitive state or intent (Sakita et al. 2004). While we utilized information on how long instructors waited before offering correction to build our model of hesitation-triggered repair, future work should consider how to accurately infer intent from a broader set of visible social cues and when to interrupt and aid the user in order to enable a more "proactive" form of repair found in previous research on human–human interactions (Clark 1996).

Identification of pieces: While grouping and summarization were the most common strategies instructors used in delivering instructions, a less common strategy was the instructor verbally identifying all of the pieces necessary to complete the task step before giving the instruction. For example, the instructor might mention six pieces, indicating that the trainee should separate those pieces from the rest, and then provide instructions for one or more task steps that use only this subset of pieces. We speculate that instructors utilized this strategy in order to simplify their instructions by reducing the number of clarifications they would have to offer to identify the pieces from a larger set. Future work should collect additional data on this strategy and examine its effectiveness.

Incremental task guidance: For some of the steps in the task, instructors did not give all the information necessary to complete the step correctly. For example, we observed instructors to give information on where a joint belongs but not the proper orientation of the joint. These instructors then waited until the joint was attached and then gave orientation information. From our data, it was unclear whether or not instructors intentionally omitted details that may be revealed later in the process, believed the detail to be unnecessary, or forgot about the detail until the resulting structure did not match their expectation. Future work might examine these and other possible motivations for withholding instruction details and whether this approach has any impact on task outcomes.

Student controlling interaction: When instructors employed summarization in the delivery of instructions, trainees occasionally continued the structure assembly beyond the instructions given by the instructor, relying only on the summary in performing task steps. For example, because the structure assembled by the human dyads was symmetrical, we observed some instructors to first explain the overall construction and the symmetrical form of the completed structure. These instructors then guided the trainee through the assembly of one side of the structure who then on occasion continued through to the completion of the other side on their own without prompts or instructions from the instructor. This behavior required instructors to continuously monitor their trainees for any potential errors to be able to interrupt the trainee and provide correction in a timely manner.

6.2 Design implications

These results have a number of implications for the design of instructional robots. Our results suggest that, despite resulting in more mistakes, grouping significantly improves task completion times, making it ideal for settings in which faster task completion are critical and mistakes are not costly. Furthermore, coupling summarization with grouping alleviates some of the mistakes caused by providing multiple instruc-

tions at once. However, there are many scenarios where providing instructions one-by-one might be preferable. For example, with more complex tasks or students who might have trouble keeping up with the robot's instructions (e.g., novices), providing instructions one-by-one might help the student complete the task with fewer breakdowns. Additionally, in situations where mistakes could be dangerous or costly, individual instruction might reduce the chance of these mistakes occurring. In these scenarios, including summaries of upcoming instructions might also improve student rapport with the robot.

The analysis of the human–human data suggests that robots will need to be equipped with fast and accurate methods for assessing trainee status and adapting to their needs. This analysis also demonstrated that summarization, when contextualized appropriately, has the potential to alleviate some of the breakdowns caused by grouping. To implement this behavior, robots will need to be capable of gauging the difficulty of task steps as well as the trainee's ability to complete them. Our qualitative analysis of the human–human data highlights additional strategies, such as references to subtasks or the use of gestures, that may be employed at opportune moments to aid in conveying instructions.

6.3 Limitations

The work presented here has three key limitations. First, although our model considers two structural components of instruction-giving, there are other elements that we did not observe or that occurred infrequently in our modeling study and thus did not include in our model. Analyses of human interactions in a more diverse set of instructional scenarios may enable the development of richer models of instruction. Second, while our repair model offered repair when prompted, the system did not proactively offer repair due to the difficulty of accurately discerning when mistakes occurred. The structure of the task and available methods for perception made it difficult to continuously update a model of the workspace and determine whether it was being modified, as participants obstructed the camera's view when modifications were occurring. Third, our evaluation focused on testing only the immediate effects of the proposed instructional strategies on student performance and perceptions. We plan to extend our work to explore a more diverse set of instructional scenarios, instructions that are distributed over time, and long-term effects of the proposed strategies on task-based instruction.

6.4 Future work

The limitations of the work presented here highlight opportunities for future research into how robots should give instructions. Two immediate extensions are (1) enhancing instructions by integrating nonverbal behaviors and (2) enabling the robot to provide proactive help. While we chose to focus our current investigation on speech, prior work demonstrated the important role nonverbal behaviors can play in conveying information or intent, such as using gestures to identify objects (Brooks and Breazeal 2006; Sauppé and Mutlu 2014b) or spaces (Hato et al. 2010). These behaviors can enhance the robot's verbal instructions and improve its effectiveness as an instructor. Additionally, proactively offering help can significantly improve trainee performance and experience. Achieving proactive repair will require the system to integrate several additional cues from the trainee, such as the trainee's gaze cues, and the task space, such as what pieces are being manipulated, with a data-driven model of common mistakes trainees make in order to predict errors before they occur.

In addition to building enhanced instructional capabilities for the robot, future work may explore how the design process we followed in this work can be improved. For example, many of the models presented here, such as our model of how different types of task breakdowns triggered different forms of repair, can be learned from data on human instructor–trainee interactions. Similarly, user states such as hesitation can be recognized using predictive models trained on a richer set of cues from the user and the task space.

7 Conclusion

As robots move into roles that involve providing users with task guidance, such as teaching in labs and assisting in assembly, they need to employ strategies for effective instruction. In this paper, we described two key instructional strategies—grouping and summarization—based on observations of human instructor–trainee interactions in a pipe-assembly task. We implemented these strategies on a robot that autonomously guided its users in this task and evaluated their effectiveness in improving trainee task performance and experience in human–robot instruction. Our results showed that, when the robot grouped instructions, participants completed the task faster but encountered more breakdowns. We also found that summarizing instructions increased participant rapport with the robot. However, comparisons with results from the human instructor–trainee interactions indicates that summarization is a more nuanced strategy than our system implemented. Future work should further explore summarization strategies for robots. Our findings show that grouping instructions results in a tradeoff between task time and breakdowns and that summarization has some benefits under certain conditions, suggesting that robots selectively use these strategies based on the goals of the instruction.

Acknowledgments We thank Brandi Hefty, Jilana Boston, Ross Luo, Chien-Ming Huang, and Catherine Steffel for their contributions to and National Science Foundation Awards 1149970 and 1426824 and Mitsubishi Heavy Industries, Ltd. for their support of this work. Some of the findings from the human–human and human–robot data presented here have been published in the *Proceeding of Robotics: Science and Systems* (Sauppé and Mutlu 2014a) and included in a book chapter in *Robots that Talk and Listen* (Markowitz 2015).

References

Alfieri, L., Brooks, P., Aldrich, N., & Tenenbaum, H. (2011). Does discovery-based instruction enhance learning? *Journal of Educational Psychology*, *103*(1), 1–18.

Andrist, S., Spannan, E., & Mutlu, B. (2013). Rhetorical robots: Making robots more effective speakers using linguistic cues of expertise. In *Proc. HRI'13* (pp. 341–348).

Bicho, E., Erlhagen, W., Louro, L., & Costa e Silva, E. (2011). Neuro-cognitive mechanisms of decision making in joint action: A human-robot interaction study. *Human Movement Science*, *30*(5), 846–868.

Blaylock, N., Allen, J., & Ferguson, G. (2003). Managing communicative intentions with collaborative problem solving. In K. JCJ & R. Smith (Eds.), *Current and new directions in discourse and dialogue* (pp. 63–84). Berlin: Springer.

Boucher, J. D., Pattacini, U., Lelong, A., Bailly, G., Elisei, F., Fagel, S., et al. (2012). I reach faster when I see you look: Gaze effects in human–human and human–robot face-to-face cooperation. *Frontiers in Neurorobotics*, *6*, 1–11.

Brooks, A. G., & Breazeal, C. (2006). Working with robots and objects: Revisiting deictic reference for achieving spatial common ground. In *Proceedings of the 1st ACM SIGCHI/SIGART conference on human-robot interaction* (pp. 297–304). New York: ACM.

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

Clark, H. H. (2005). Coordinating with each other in a material world. *Discourse Studies*, *7*(4–5), 507–525.

Foster, M., Giuliani, M., Isard, A., Matheson, C., Oberlander, J., & Knoll, A. (2009). Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proc. IJCAI'09* (pp. 1818–1823).

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences*, *2*(12), 493–501.

Gonsior, B., Wollherr, D., & Buss, M. (2010). Towards a dialog strategy for handling miscommunication in human-robot dialog. In *Proc. RO-MAN'10*

Gray, J., Breazeal, C., Berlin, M., Brooks, A., & Lieberman, J. (2005). Action parsing and goal inference using self as simulator. In *Proc. RO-MAN'05*

Grosz, B., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Grosz, B., & Sidner, C. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Hato, Y., Satake, S., Kanda, T., Imai, M., & Hagita, N. (2010). Pointing to space: Modeling of deictic interaction referring to regions. In *Proceedings of the 5th ACM/IEEE international conference on human-robot interaction* (pp. 301–308). New York: IEEE Press

Hirst, G., McRoy, S., Heeman, P., Edmonds, P., & Horton, D. (1994). Repairing conversational misunderstandings and non-understandings. *Speech Communication*, *15*(3), 213–229.

Hoey, J., Poupart, P., Boutilier, C., & Mihailidis, A. (2005). POMDP models for assistive technology. In *Proceedings of the AAAI 2005 fall symposium*.

Huang, C. M., & Mutlu, B. (2012). Robot behavior toolkit: Generating effective social behaviors for robots. In *Proc. HRI'12* (pp. 25–32).

Kanda, T., Sato, R., Saiwaki, N., & Ishiguro, H. (2007). A two-month field trial in an elementary school for long-term human-robot interaction. *IEEE Transactions on Robotics*, *23*(5), 962–971.

Koulouri, T., & Lauria, S. (2009). Exploring miscommunication and collaborative behaviour in HRI. In *Proc. SIGDIAL'09*.

Landis, J., & Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lee, M., Kiesler, S., Forlizzi, J., Srinivasa, S., & Rybski, P. (2010). Gracefully mitigating breakdowns in robotic services. In *Proc. HRI'10*.

Markowitz, J. (2015). *Robots that talk and listen*. Boston: Walter de Gruyter.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

Nicolescu, M., & Mataric, M. (2003). Linking perception and action in a control architecture for human-robot domains. In *Proc. HICSS'03*.

Norrick, N. (1991). On the organization of corrective exchanges in conversation. *Journal of Pragmatics*, *16*(1), 59–83.

Reigeluth, C., Merrill, M., Wilson, B., & Spiller, R. (1980). The elaboration theory of instruction: A model for sequencing and synthesizing instruction. *Instructional Science*, *9*(3), 195–219.

Sakita, K., Ogawara, K., Murakami, S., Kawamura, K., & Ikeuchi, K. (2004). Flexible cooperation between human and robot by interpreting human intention from gaze information. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004 (IROS 2004) Proceedings* (Vol. 1, pp. 846–851). New York: IEEE

Sauppé, A., & Mutlu, B. (2014a). Effective task training strategies for instructional robots. In *Proceedings of the 10th annual robotics: science and systems conference*.

Sauppé, A., & Mutlu, B. (2014b). Robot deictics: How gesture and context shape referential communication. In *Proceedings of the 2014 ACM/IEEE international conference on human-robot interaction* (pp. 342–349). ACM.

Schegloff, E., Jefferson, G., & Sacks, H. (1977). The preference for self-correction in the organization of repair in conversation. *Language*, *53*, 361–382.

Seedhouse, P. (1999). The relationship between context and the organization of repair in the l2 classroom. *International Review of Applied Linguistics in Language Teaching*, *37*(1), 59–80.

Staudte, M., & Crocker, M. (2009). Visual attention in spoken human-robot interaction. In *Proc. HRI'09* (pp. 77–84).

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285.

Tanaka, F., & Movellan, J. (2006). Behavior analysis of children's touch on a small humanoid robot: Long-term observation at a daily classroom over three months. In *Proc. RO-MAN'06*.

Tanaka, R., & Kimura, T. (2009) The use of robots in early education: A scenario based on ethical consideration. In *Proc. RO-MAN'09*.

Tomasello, M., Conti-Ramsden, G., & Ewert, B. (1990). Young children's conversations with their mothers and fathers: Differences in breakdown and repair. *Journal of Child Language*, *17*(01), 115–130.

Torrey, C., Powers, A., Marge, M., Fussell, S., & Kiesler, S. (2006). Effects of adaptive robot dialogue on information exchange and social relations. In *Proc. HRI'06* (pp. 126–133).

Torrey, C., Powers, A., Fussell, S., & Kiesler, S. (2007). Exploring adaptive dialogue based on a robot's awareness of human gaze and task progress. In *Proc. HRI'07*.

Trafton, J., Cassimatis, N., Bugajska, M., Brock, D., Mintz, F., & Schultz, A. (2005). Enabling effective human-robot interaction using perspective-taking in robots. *IEEE Transactions on Systems, Man, and Cybernetics, Part A*, *35*(4), 460–470.

Zahn, C. (1984). A reexamination of conversational repair. *Communications Monographs*, *51*(1), 56–66.

Allison Sauppé is an assistant professor of computer science at the University of Wisconsin–La Crosse. She received her Ph.D. degree in Computer Sciences from the University of Wisconsin–Madison in 2015. As part of the Wisconsin Human-Computer Interaction Laboratory, her work explores human–robot interaction, modeling human behavior, and enabling effective human–robot collaboration. She received her M.S. in Computer Sciences from the University of Wisconsin–Madison in 2011 and B.S. in Computer Science and Software Engineering from Rose-Hulman Institute of Technology in 2009.

Bilge Mutlu is an associate professor of computer science at the University of Wisconsin–Madison where he directs the Wisconsin Human-Computer Interaction Laboratory. He received his Ph.D. degree from Carnegie Mellon University's Human-Computer Interaction Institute in 2009. His background combines training in interaction design, human-computer interaction, and robotics with industry experience in product design and development. Dr. Mutlu is a former Fulbright Scholar and the recipient of the NSF CAREER award and several paper awards and nominations, including HRI 2008, HRI 2009, HRI 2011, UbiComp 2013, IVA 2013, RSS 2013, HRI 2014, and CHI 2015. His research has been covered by national and international press including the New Scientist, MIT Technology Review, Popular Science, Discovery News, Science Nation, and Voice of America. He has served in the Steering Committee of the HRI Conference and the Editorial Board of IEEE Transactions on Affective Computing, co-chairing the Program Committees for HRI 2015 and ICSR 2011, the Program Sub-committees on Design for CHI 2013 and CHI 2014, and the Organizing Committee for HRI 2017. More information on Dr. Mutlu and his research program can be found at http://bilgemutlu.com and http://hci.cs.wisc.edu.