Taking the Actor’s Perspective Enhances Action Understanding and Learning

Bridgette A. Martin (martin@psych.stanford.edu)
Sandra C. Lozano (scl@psych.stanford.edu)
Barbara Tversky (bt@psych.stanford.edu)
Department of Psychology, Bldg. 420 Jordan Hall
Stanford, CA 94305 USA

Abstract

Encoding observed action according to its hierarchical, goal-subgoal structure makes it easier to learn, and leads learners to orient objects and actions from the actor’s spatial perspective. The present pair of studies tested whether adopting the actor’s perspective during encoding leads observers to encode actions more hierarchically. Participants observed and segmented an action sequence that they later performed. In Study 1, participants described the action of each segment as they segmented, either from the actor’s perspective or from their own. In Study 2, participants described the action from the actor’s perspective or from the perspective of a passive observer in the scene. In both studies, participants who described the actions from the actor’s perspective segmented more hierarchically and learned more effectively. Encoding actions from the actor’s perspective thus enhances understanding of goal-subgoal organization.

Introduction

Learning actions by observing others is often the most efficient way to acquire complex skills (Caroll & Bandura, 1990). A key factor in learning observed actions is the ability to infer their underlying intentional structure. For example, people plan tasks like building a sandcastle according to a goal-subgoal hierarchy: an overarching goal (make the sandcastle) is broken down into component subgoals (build the towers, dig the moat, etc.), that are in turn broken down into component subgoals (scoop sand into a bucket, pack it down, etc) (Newell & Simon, 1972).

Recent studies by Martin, Lozano, & Tversky (2004, 2005) showed that encoding observed actions according to their hierarchical organization facilitates learning, because it involves representing observed actions in the same form used for action plans, thus easing the translation between observation and execution. In these studies, participants watched a film of an actor assembling objects, and pressed a key to segment the film into large and small action units on two separate viewings. Later, participants performed the observed assembly task themselves.

Previous work has shown that people encode action hierarchically: they segment large and small action units that are hierarchically related, corresponding to goals and subgoals, respectively (Zacks, Tversky, & Iyer, 2001; Martin, Tversky, & Lang, 2004). But observers differ in how consistently they segment hierarchically. Martin et al. (2004, 2005) found that observers who segmented more hierarchically learned the assembly task better. Telling participants that actions were hierarchically organized and encouraging them to look for hierarchically related actions increased hierarchical segmentation, and consequently learning. Also, participants who described the actions while segmenting them or who segmented larger actions first segmented more hierarchically and learned better than participants who segmented silently or who segmented smaller actions first.

Surprisingly, observers who segmented more hierarchically tended to copy the actor’s perspective when performing the assembly task, positioning objects on their left that had been on the actor’s left, for example. Participants who segmented less hierarchically tended to maintain an observer’s perspective in performing the assembly task, positioning objects on their right that had been on the actor’s left. These findings did not just reflect individual differences: manipulations that increased hierarchical segmentation—encouraging observers to infer hierarchical organization, requiring them to describe the actions or to identify larger action units—also increased the tendency to copy the actor’s perspective during assembly.

Why did participants who segmented more hierarchically adopt the actor’s perspective when performing the assembly task? One possibility is that participants who segmented more hierarchically copied the actor’s perspective because they originally encoded the actions from this perspective. Many of the participants who copied the actor’s perspective were ones who had described the actions during the segmentation task. Analysis of their descriptions showed that described space from the actor’s perspective (e.g., she takes the side board on her left) more often than from their own (e.g., she takes the side board on the right). These results are surprising; common intuition dictates that the easiest, most natural spatial perspective is one’s own. At least since Piaget and Inhelder’s (1956) work on perspective taking, it has been assumed that taking another person’s perspective is inherently difficult and will necessarily increase the cognitive load of the observer.

Furthermore, among participants who described, describing actions from the actor’s perspective predicted greater degree of hierarchical segmentation and better performance on the later assembly task. These results suggest that encoding the hierarchical organization of actions involves understanding those actions from the actor’s perspective. According to this hypothesis, instructing observers to think about actions from the actor’s perspective
should lead to better hierarchical encoding than instructing observers’ to maintain their own perspective.

**Study 1: Actor- vs. Self-Perspective**

**Method**

**Participants and Design** Forty Stanford University undergraduates participated in exchange for course credit. A 2 x 2 x 2 x 2 Mixed Factorial design was used. Segmentation level (*fine, coarse*) was varied within participants; and perspective condition (*actor-perspective, self-perspective*), segmentation order (*coarse first, fine first*), and awareness of the later assembly task (*aware, unaware*) were varied between participants.

**Stimuli** Participants viewed two films, one for practice and one for test. The practice film was created by Zacks et al. (2001) and showed a woman assembling a saxophone. The test film showed a woman assembling a television (TV) cart (see Figure 1). Films were presented on a 21-inch, flat screen computer monitor. Response times were recorded on a Macintosh G4 computer attached to a keyboard, using a program written in PsyScope 1.2.5 (Cohen, MacWhinney, Flatt, & Provost, 1993).

**Procedure** Participants were asked to view two films and divide those films into separate events, by pressing the spacebar whenever one event ended and another event began. Each time they pressed the spacebar, participants were asked to briefly describe what happened in the segment they had just observed, from their own or from the actor’s perspective. Participants were shown a picture of a woman placing a saxophone on a table (from the practice film) and were given examples of how this picture could be described either from an actor- or self-perspective.

Participants then viewed the saxophone assembly film created by Zacks et al. (2001) to practice segmenting and describing from their assigned perspective. Participants were asked to mark whatever events felt natural and meaningful to them.

Participants then segmented and described the test film, which was 6 minutes and 35 seconds long and showed an actor assembling the TV cart in the same room where participants were tested. Half of the participants were asked to mark off the smallest units that seemed natural and meaningful to them (*fine first*); the other half were asked to mark off the largest units that seemed natural and meaningful (*coarse first*). Participants then segmented the test film a second time according to the opposite unit-size instructions. Viewing the films twice was necessary to create a measure of hierarchical segmentation (Martin et al., 2004, 2005; Zacks et al., 2001). Before segmenting the test film, participants in the *aware* condition were told that they would later assemble the object they saw in the test film.

**Results and Discussion**

**Does Perspective Affect Hierarchical Encoding?**

Hierarchical encoding was measured by assessing the degree of hierarchical organization in participants’ segmentation. This was accomplished using the technique developed by Martin et al. (2004, 2005). Key presses made during coarse segmentation were called coarse breakpoints. Key presses made during fine segmentation were called fine breakpoints. For each participant, degree of hierarchical segmentation equaled the total number of coarse breakpoints that fell after their closest fine breakpoint, divided by the total number of coarse units. If two or more coarse breakpoints shared a common closest fine breakpoint, then that fine breakpoint was paired with the coarse breakpoint it was closest to and the remaining coarse breakpoints were left unpaired.

Participants in the *actor-perspective* condition segmented more hierarchically (*M* = .74, *SEM* = .04) than participants in the *self-perspective* condition (*M* = .55, *SEM* = .03), *F*(1, 32) = 11.53, *p* < .01. Participants who segmented *coarse first* segmented more hierarchically (*M* = .71, *SEM* = .05) than participants who segmented *fine first* (*M* = .59, *SEM* = .03), *F*(1, 32) = 4.77, *p* < .05. Awareness of the later assembly task did not affect hierarchical segmentation and no interactions were significant.

**Does Perspective Affect Learning?** For each participant, assembly performance was coded for total time and total errors. On average, participants completed the TV cart assembly in about 10 minutes (*SEM* = 35.33 s), and made approximately 2 errors (*SEM* = 0.22).
Of participants who described from an actor-perspective, 97% assembled the television cart by taking the actor’s perspective: they built the cart in the same orientation and usually stood on the same side of the table as the actor. Of the participants who had described the film from a self-perspective, 75% assembled the cart from a self-perspective: they oriented the television cart pieces the opposite direction as the actor and usually stood on the opposite side of the table, $X^2 = 23.41, p < .001$ (see Figure 2). Neither segmentation order nor awareness had a reliable effect on participants’ later assembly perspective.

The perspective participants described affected both assembly time and errors. Participants who described the film from an actor-perspective assembled the cart 27% faster ($M = 508.75$ s, $SEM = 32.60$ s) than participants who described from a self-perspective ($M = 695.40$ s, $SEM = 56.07$ s), $F(1, 32) = 8.58, p < .01$. Participants in the actor-perspective condition also made half as many errors ($M = 1.50$, $SEM = 0.18$) as those who described from a self-perspective ($M = 2.95$, $SEM = 0.33$), $F(1,32) = 16.65, p < .001$.

As Figure 3 shows, there was an interaction between segmentation order and perspective condition, $F(1, 32) = 7.15, p < .05$, such that actor-perspective describers did equally well whether they segmented coarse first ($M = 1.7$ errors, $SEM = 0.30$) or fine first ($M = 1.3$, $SEM = 0.21$). Self-perspective describers, however, made fewer errors if they segmented coarse first ($M = 2.2$, $SEM = 0.35$) than fine first ($M = 3.7$, $SEM = 0.44$).

Replicating findings of Martin et al. (2004, 2005), hierarchical segmentation predicted later assembly performance: participants who segmented more hierarchically made fewer assembly errors, $r(38) = -0.50, p < .01$.

Figure 3: Mean number of assembly errors, as a function of perspective condition and segmentation order.

Did changes in hierarchical segmentation mediate the effects of perspective condition on assembly performance? To address this question, a mediation analysis was performed, according to the techniques of Baron and Kenny (1986).

According to a linear regression, perspective condition predicted hierarchical segmentation ($B = 0.59$, $SE = 0.04$, $t(1) = 15.17, p < .001$) and assembly errors ($B = -6.61$, $SE = 1.23$, $t(1) = -5.39, p < .001$). Hierarchical segmentation also predicted assembly errors ($B = -8.42$, $SE = 3.34$, $t(1) = -2.52, p < .05$). When hierarchical segmentation was controlled for, the effect of perspective condition on assembly errors was no longer significant ($B = -0.067$, $SE = 0.28$, $t(1) = -0.24, p = .81$). Thus, hierarchical segmentation mediated effects of perspective on learning. In other words, adopting the actor’s perspective was beneficial to action learning because it encouraged hierarchical segmentation.

**Does Perspective Change the Nature of Descriptions?** To further validate our measure of hierarchical encoding, participants’ descriptions were analyzed for links between hierarchical description and segmentation. During fine segmentation, participants’ task was to divide actions into the smallest meaningful units. Thus, one would expect them to provide descriptions corresponding to substeps of the observed task. However, some participants offered summary statements (e.g., “She attached the top shelf”) in addition to fine-unit substeps (e.g., “She inserted the first screw,” “She inserted the second screw,” etc.). These observers found a way both to segment fine units and to meaningfully group them as they happened. The number of summary statements participants provided correlated with hierarchical segmentation, $r(38) = 0.41, p < .01$, replicating findings of Martin et al. (2004, 2005), and confirming that this measure is a valid assessment of hierarchical encoding.

Participants in the actor-perspective group provided more summary statements ($M = 2.45$, $SEM = 0.38$) than participants in the self-perspective group ($M = 1.00$, $SEM = 0.31$), $F(1, 32) = 10.92, p < .01$. This result confirms that taking the actor’s perspective leads to more hierarchical encoding of observed actions. There was also an interaction between perspective condition and segmentation order, $F(1,32) = 10.92, p < .01$, such that people describing from a
self-perspective included fewer summary statements if they segmented coarse first ($M = 0.4$, SEM = 0.10) than if they segmented fine first ($M = 1.6$, SEM = 0.20); but participants who described from an actor-perspective included more summary statements if they segmented coarse first ($M = 3.3$, SEM = 0.30), rather than fine first ($M = 1.6$, SEM = 0.31). No other effects or interactions were reliable.

Participants’ verbal descriptions were also analyzed for how often and from which perspective they described spatial information. Each description corresponding to a breakpoint was assigned to one of four mutually exclusive categories: nonspatial (e.g., “She inserted the screw”); spatial from neutral-perspective (e.g., “She inserted the screw from the side”); spatial from an actor-perspective (e.g., “She inserted the screw on her left”); or spatial from self-perspective (e.g., “She inserted the screw on my right”). This analysis yielded two key insights, which are evident from Figure 4.

First, regardless of whether participants were assigned to describe events from an actor- or self-perspective, the content of their descriptions was remarkably similar. Participants described equal amounts of spatial information relative to nonspatial information in both conditions, $t(38) = -0.67$, $p = .43$. The proportions of descriptions that participants encoded from their assigned spatial perspective (self for self-perspective describers vs. actor for actor-perspective describers) were also equivalent, $t(38) = -0.72$, $p = .48$. Later differences in performance were thus due to the specific perspective that participants encoded, not to differences in how much they attended to space or the general nature of what they were describing.

Second, over 6% of self-perspective participants’ descriptions were from the wrong (actor) perspective. While it was common for self-perspective participants to erroneously use the actor’s perspective at least once ($M = 1.44$, SEM = 0.41), self-perspective participants almost never described the wrong (self) perspective ($M = 0.2$, SEM = 0.13), $t(38) = 2.26$, $p < .001$.

Self-perspective describers were also more likely to qualify their perspective ($M = 2.11$, SEM = 0.72) than actor-perspective describers ($M = 0.10$, SEM = 0.10), $t(38) = 2.89$, $p < .001$. In other words, they were more likely to explicitly state the perspective they were using (e.g., “She inserted the screw on the right, but that’s only if it’s my right that we’re talking about”). Combined, these results suggest that adopting the actor’s perspective was the more natural, preferred perspective when attempting to understand action.

Even within the actor-perspective condition, describing the actor’s perspective improved learning. Participants in this group who described the actor’s perspective more often later built the cart faster, $r(18) = -.37$, $p < .05$, and made fewer assembly errors, $r(18) = -.50$, $p < .01$. By contrast, in the self-perspective condition, using more self-perspective descriptions led to slower assembly times, $r(18) = .45$, $p < .01$ and increasing numbers of assembly errors, $r(18) = .42$, $p < .01$.

Study 2: Actor- vs. Observer-Perspective

Does taking the actor’s perspective really enhance encoding of goal-subgoal organization? Or could it be that adopting any perspective other than their own gets observers more engaged in the task, leading to more attention and better encoding. Adopting the actor’s perspective required a 180° spatial transformation of all observed actions and objects, whereas describing a self-perspective required no such mental rotation. Conceivably the added cognitive effort of performing this mental rotation could have made the task more engaging and this, not perspective-taking per se, might accounts for the results of Study 1.

The aim of Study 2 was to address this concern. Because the TV Carts from Study 1 were also discontinued, we also took the opportunity to generalize the findings from Study 1 to a different assembly task. Participants were shown a film containing an actor and an observer, both of whom were rotated 90° from participants viewing the film. Participants’ job was to describe actions from the perspective of either the actor or the observer in the film. If simulating any perspective other than one’s own can motivate observers to encode goal-subgoal structure, then the two groups should have equivalent hierarchical segmentation and assembly performance.

Method

Participants and Design Sixteen Stanford University undergraduates participated in exchange for course credit. A $2 \times 2 \times 2$ Mixed Factorial design was used. Segmentation level (fine, coarse) was varied within participants; perspective condition (actor-perspective, observer-perspective), segmentation order (coarse first, fine first), and actor position (left, right) were varied between participants.

Stimuli All participants viewed a practice film and test film. The practice film showed a female observer watching a female actor make coffee. The test film contained the same observer and actor from the practice film; but this film
showed the actor assembling two camels and a heart using red, yellow, green, and blue blocks made by Duplo®. The actor made approximately equal numbers of left- and right-handed actions. In all films, both the observer and actor stood opposite of each other and were rotated 90° to the camera (see Figure 5).

Procedure Prior to testing, participants completed the Vandenberg Mental Rotation Test (MRT), a measure of spatial ability (Vandenberg & Kuse, 1978). Aside from this, the procedure was identical to that used in Study 1, except that when describing each film, all participants were asked to adopt a perspective that was offset by 90° from their viewing perspective, either the actor’s or the observer’s. All participants were unaware of the later assembly task during segmentation.

Results and Discussion

Does Perspective Affect Hierarchical Encoding? Replicating findings from Study 1, participants who described from an actor-perspective segmented more hierarchically (M = .72, SEM = .08) than observer-perspective describers (M = .35, SEM = .07), F(1, 12) = 11.62, p < .01.

Does Perspective Affect Learning? On average, participants completed the assembly task in about 9 minutes (SEM = 105 s) and made about 8 errors (SEM = 1.78). Spatial ability (as measured by MRT scores) did not predict assembly time or errors (p > .05 for both); it was also uncorrelated with any other measures.

Of the participants who described from the actor’s perspective, 100% performed the assembly task from the actor’s perspective, meaning that they built the cart in the same orientation and stood on the same side of the table as the actor. Of the participants who described from an observer-perspective, 63% assembled the cart from the observer’s perspective: they oriented the blocks as they looked to the observer in the film and stood on the observer’s side of the table, X² = 9.29, p < .01.

Replicating findings from Study 1, participants who described from an observer-perspective made about 4 times as many assembly errors (M = 12.25, SEM = 2.45) as those who described from an actor-perspective (M = 3.37, SEM = 1.41), F(1, 12) = 13.96, p < .01 (see Figure 6). Participants who segmented fine units first made over twice as many errors (M = 11.13, SEM = 2.83) as those who segmented coarse units first (M = 4.50, SEM = 1.58), F(1, 8) = 7.78, p < .02. As in Study 1, hierarchical segmentation predicted fewer errors on the later assembly task, r(16) = -.52, p < .05 and marginally faster assembly time, r(14) = -.44, p = .09. As in Study 1, hierarchical segmentation fully mediated the effects of perspective condition on assembly errors.

Does Perspective Change the Nature of Descriptions? Analysis of descriptions replicated findings from Study 1. Participants who described from an actor-perspective included more summary statements in their descriptions (M = 1.63, SEM = 0.32) than did those who described from an observer-perspective (M = 0.13, SEM = 0.13), F(1, 12) = 19.64, p < .01. No other effects or interactions were reliable. The number of summary descriptions again correlated positively with hierarchical segmentation, r(14) = 0.52, p < .05.

The content of the descriptions in the two groups was again remarkably similar. Participants described equal amounts of spatial information relative to nonspatial information, regardless of perspective, t(14) = -0.96, p = .35. The proportions of statements that they encoded from their assigned perspective (observer for observer-perspective describers vs. actor for actor-perspective describers) were also equivalent, t(14) = -0.17, p = .43.

Participants in the observer-perspective condition were more likely to use the incorrect perspective (M = 1.50, SEM = 0.87) than those in the actor-perspective condition (M = 0.00, SEM = 0.00), t(14) = 1.73, p < .01. Observer-perspective describers were also more likely to qualify their perspective (M = 0.63, SEM = 0.26) than those in the actor-perspective condition (M = 0.00, SEM = 0.00), t(14) = 2.38, p < .01.

The amount of spatial information that a participant included from their assigned (actor or observer) perspective predicted how well they performed on the later assembly task. Participants who used more actor-perspective
descriptions later made fewer assembly errors, $r(14) = -.54$, $p < .05$. By contrast, using more observer-perspective descriptions led to marginally slower assembly times, $r(14) = .45$, $p = .08$ and more assembly errors, $r(14) = .86$, $p < .001$.

General Discussion

The present studies examined the relationship between hierarchical encoding of observed action and the perspective from which actions are understood. Participants watched an object assembly task, segmented it into the largest and smallest units that made sense, described each segment, and then performed the same task they observed in the film. Describing the film from the actor’s perspective reliably increased degree of hierarchical segmentation and subsequent learning compared to describing the film from a self or observer perspective. Furthermore, describing from any perspective but the actor’s led to more description errors and more qualifications. This suggests that the more natural, preferred perspective for encoding actions is the actor’s.

These results leave open the question of why describing observed action from the actor’s perspective is preferred over one’s own, and why this leads to improved encoding of hierarchical organization. One possibility is that encoding goal-subgoal structure requires people to think about actions as if they were the ones performing them. In other words, people might simulate observed actions to gain insight into their organization.

It is well-documented that people can mentally simulate actions, and that this simulation prepares them to act. Merely imagining skilled performance can constitute a kind of practice, as many athletes, musicians, and surgeons know (Pascuale-Leone, 2001). A large number of empirical findings suggest that people simulate observed actions in a similar way (for a review, see Grèzes & Decety, 2001).

Simulation has been proposed over the years as at least one way that people connect themselves to others and infer their goals and intentions (Gallese, 2001; Gallese & Goldman, 1998; Gordon, 1995). The underlying idea is that when people perform actions and observe their outcome, they gain insight into action relations and their causal connection to the outcome. In the same way, action simulation might provide insight into how actions relate to each other and to higher-level goals.

Findings from Martin et al. (2004, 2005), as well as findings from the present studies, are consistent with the idea that observers simulate actions to infer their organization. Observers who attend to action organization spontaneously encode those actions from the actor’s perspective (Martin et al., 2004, 2005), and observers who are instructed to encode the actor’s perspective spontaneously attend to action organization.

Acknowledgments

We are grateful to Jane Solovyeva and Sanjay Kairam for their assistance with stimuli creation and data collection. Portions of this research were supported by Office of Naval Research, Grants Number NOO014-PP-1-O649 and N000140110717 to Stanford University.

References


