Journal of Experimental Psychology: General

A Dissociation Between Conceptual Prominence and Explicit Category Learning: Evidence From Agent and Patient Event Roles

Lilia Rissman and Gary Lupyan

Online First Publication, November 11, 2021. http://dx.doi.org/10.1037/xge0001146

AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

https://doi.org/10.1037/xge0001146

A Dissociation Between Conceptual Prominence and Explicit Category Learning: Evidence From Agent and Patient Event Roles

Lilia Rissman and Gary Lupyan
Department of Psychology, University of Wisconsin – Madison

We investigate whether linguistic categories have the same structure as categories used to conceptualize the world outside of language. We focus on the event roles Agent and Patient (in the sentence *Murray ate the ice cream*, Murray is the Agent and the ice cream is the Patient). These categories appear to be tightly linked across language and cognition: they are encoded robustly in the world's languages and have been argued to be highly prominent conceptually, even part of innate core knowledge. This view predicts (a) that Agent and Patient categories will be readily accessible to adults in explicit categorization tasks and (b) that these categories have similar structure across semantic and conceptual domains. We tested these predictions across four experiments in which adult speakers of English had to induce Agent and Patient categories from visual illustrations of events (e.g., one figure kicking another). We found that 25% to 40% of participants failed to induce the categories, suggesting that prominent concepts are not always easily accessed for conscious reasoning. At the same time, for those participants who did induce the categories, they generalized these categories in ways predicted by previous analyses of English syntax. This finding supports the view that Agent and Patient are domain-general, spanning both conceptual and linguistic representation, though not necessarily used by participants in explicit categorization tasks.

Keywords: agency, categorization, concepts, event cognition, thematic roles

To what extent do the categories people use in language have the same structure as the categories used in making sense of the world outside of language? On the one hand, conceptual and semantic categories appear to parallel each other in a variety of ways (see Strickland, 2017; Ünal et al., 2020). For example, the Goal of a motion event (e.g., *Dionne threw the ball to home plate*) has been shown to be more prominent than the Source (e.g., *Dionne threw the ball from first base*) in both linguistic and nonlinguistic tasks and for users of diverse languages (Lakusta & Landau, 2012; Lakusta et al., 2017; see Rissman & Majid, 2019, for review). On the other hand, individual linguistic structures in different languages vary substantially in how they carve up conceptual space: for example, the Chinese verb *huà* labels a set of events that English speakers separate into *painting* versus *drawing* events. So the relationship between conceptual and semantic structure cannot be described as a mapping from the concept DRAW to the word *draw* (see Lupyan & Lewis, 2019; Malt & Majid, 2013). Given both types of evidence (for parallelism vs. differentiation), the tightness of the linkage between conceptual and semantic representation is still a matter of debate.

We address this question here by focusing on representations of event roles (sometimes called "thematic roles"): for example, in *Murray petted the cat*, Murray is characterized as an Agent and the cat as a Patient in many analyses (Fillmore, 1968; Gruber, 1965; Levin & Rappaport-Hovav, 2005). These roles are encoded robustly in language—almost without exception, the languages of the world distinguish Agent and Patient roles morpho-/syntactically (see Rissman & Majid, 2019, for review). In English active sentences, for example, Agents appear in the syntactic position of Subject and Patients in the syntactic position of Object—*Murray petted the cat* has to mean that Murray is the one doing the petting, by virtue of the fact that Murray is the Subject. The linguistic robustness of the Agent/Patient distinction is also evident from the observation that morphosyntactic structures for distinguishing Agents from Patients emerge rapidly in new sign languages (Ergin et al., 2018; Flaherty, 2014; Goldin-Meadow & Mylander, 1998; Padden et al., 2009; Sandler et al., 2005).

One possibility is that Agent and Patient roles are encoded robustly in language because they are important conceptually (see Jackendoff, 1983, 1990). As we review below, a variety of empirical evidence attests to the conceptual prominence of these roles—so much so that Agency has been argued to be part of universal, core knowledge (Carey, 2009; Fillmore, 1968; Spelke & Kinzler, 2007). The previous literature thus suggests a tight linkage between linguistic and conceptual representations of Agents and Patients.

Lilia Rissman https://orcid.org/0000-0002-3796-2719
Gary Lupyan https://orcid.org/0000-0001-8441-7433

This research was supported by NSF-PAC 1734260. Thank you to Tyler McCarthy for help with stimuli preparation and Kyara Rozman for help with data collection. Thank you to all study participants. This research was presented at the 2020 meetings of the Cognitive Science Society and the Psychonomics Society as well as the 2021 Dubrovnik Conference on Cognitive Science. Stimuli, data files and analysis scripts are available at: https://osf.io/a5cev/. We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to Lilia Rissman, Department of Psychology, University of Wisconsin – Madison, 1202 West Johnson Street, Madison, WI 53706, United States. Email: lrissman@wisc.edu

This view makes several predictions that have not yet been tested. First, if event roles are highly prominent conceptually, then they should be readily available to adults in explicit categorization tasks. For example, dimensions of quantity and size are thought to be conceptually prominent, and Ferrigno et al. (2017) found that adults were more than 90% accurate in learning to categorize displays of dots into *few* and *many* categories (where participants could categorize on the basis of quantity, cumulative area, or both). Second, if event roles are tightly linked across linguistic and conceptual representation, then these categories should be structured in similar ways across domains. We tested these predictions through a set of experiments in which English-speaking adults categorized pictures illustrating an Agent acting on a Patient. To the extent that these predictions are not met, it may cast doubt on the theory that a small number of core categories underlie both linguistic and conceptual processes.

The Conceptual Prominence of Agent and Patient Roles

Within the first year of life, infants represent event participants in terms of properties such as intentionality and causality that are thought to be definitive of Agent and Patient roles (see Carey, 2009; Csibra & Gergely, 2007; Kelso, 2016; for review). By 14 months of age, children can map the role of "chaser" in a chasing event to a nominal label (e.g., *a tacok*), abstracting across the perceptual characteristics of individual chasers and chasing trajectories (Yin & Csibra, 2015). Children as young as two years can map a transitive sentence such as *Elmo blicked Cook*ie *Monster* to a video of Elmo causally affecting Cookie Monster, indicating that children are have learned semantic generalizations about the roles encoded by grammatical Subject and Object (Arunachalam & Waxman, 2010; Lidz et al., 2003; Naigles, 1990; Noble et al., 2011; Savage et al., 2003).

Agent and Patient roles are examples of role-governed categories, like the English nouns *guest*, *thief*, and *friend*: their meaning depends on their relation to other individuals in a situation (Markman & Stilwell, 2001). In a sense, Agent and Patient are role-governed categories *par excellence*. In novel linguistic contexts, English-speaking adults have been shown to rapidly create new role-governed categories, indicating the prominence of role-governed categories in language (Goldwater et al., 2011).

Abstract role representations have also been argued to operate automatically when adults view even very briefly presented events. For example, Hafri et al. (2018) found that when participants were asked to locate a target individual of a particular gender or shirt color in an event, participants were slower when the target individual differed in its Agent/Patient status from the target in the previous trial. This indicates priming of role knowledge across different events. Hafri et al. (2013) found that when adults viewed images of events for only 37 ms, they could reliably answer questions about the event and the roles of the participants in the event (e.g., "is the girl performing the action?"). This ability suggests that event apprehension is aided by an abstract schema for encoding event roles. Given this set of results, Hafri and colleagues argue that extraction of event roles from visual scenes is rapid and spontaneous.

The findings reviewed thus far—that infants represent event roles and use them to learn, that roles are encoded robustly across languages, and that roles are processed automatically in perception—have been taken as evidence that event roles are part of core knowledge (Strickland, 2017). Another possible influence of core knowledge on cognition is that core knowledge affects explicit category learning. In categorization tasks, people are confronted with stimuli that differ on multiple dimensions (e.g., pictures of human faces) and must sort the stimuli according to one or more of these dimensions. Category learning involves projecting a hypothesis space of potentially relevant dimensions and moving through this space until a solution is reached. If core knowledge of event roles has a robust influence on implicit learning and perceptual processing, it may also have a robust influence on explicit category learning. In other words, the prior literature suggests that event roles will be among the salient dimensions populating learners' spaces of hypotheses. In Experiments 1–4, we test this prediction, asking whether English-speaking adults can construct Agent/Patient categories from visual scenes.[1]

Do Linguistic and Conceptual Role Categories Share Similar Structure?

The evidence reviewed thus far provides strong support for the hypothesis that Agents and Patients are linguistically prominent because they are prominent in conceptual structure. Under the theory that these roles are part of core cognition, there may be a single set of categories underlying both linguistic and conceptual processes. The view that event roles are tightly linked across semantic and conceptual domains is supported by studies investigating thematic role hierarchies (Grimshaw, 1990; Levin & Rappaport-Hovav, 2005; Ünal et al., 2021). For example, adults across a range of languages are more likely to mention Goals than Sources when describing motion events (Lakusta & Landau, 2005, 2012; Narasimhan et al., 2012; Papafragou, 2010). This asymmetry is reflected in a variety of linguistic and nonlinguistic data: patterns of extension of morphosyntactic role markers (Kabata, 2013), the gestural utterances of child homesigners (Zheng & Goldin-Meadow, 2002), adults' memory in change-detection tasks (Lakusta & Landau, 2012; Papafragou, 2010; Regier & Zheng, 2007), and infants' preferential looking (Lakusta & Carey, 2015; Lakusta et al., 2007, 2017; Tatone et al., 2015). This suite of evidence indicates that the Goal of a motion event is represented more robustly than the Source across both semantic and conceptual representation. This evidence also suggests that the asymmetry between Goals and Sources is universal.

A possible counterargument to the claim that event roles are universal is the observation that the mapping between morphosyntactic structures and role categories varies across languages (Bornkessel et al. 2006; Bowerman & Brown, 2008; Croft, 2012). For

[1] The expectation that universal conceptual knowledge translates into explicit category learning appears to be widely shared. We conducted an informal survey where we asked 49 people recruited on Twitter to estimate how well English speakers would construct Agent/Patient categories given a task such as in Experiments 2–4. About half of the same were researchers in cognitive/developmental psychology or linguistics. We then asked the respondents how much they agreed with the statement "Some types of conceptual knowledge are shared by all adults, regardless of culture or education." We found that people who endorsed this statement more strongly had higher expectations for how accurately participants would perform in the categorization task, $r(47) = .33$, $p < .05$, and for participants' ability to describe the difference between Agent and Patient categories, $r(47) = .43$, $p < .01$.

example, while English allows a range of inanimate event participants to appear as Subject (e.g., *the wind knocked over the sign; the stone broke the window*), in many languages, the Subject position is restricted to animates (see Wolff et al., 2009 for review).[2] Preferences for producing agentive language can also vary across languages: when describing accidental events such as a man crushing a can by stepping on it, English speakers are more likely to produce transitive descriptions (e.g., *the man crushed the can*) than Spanish speakers (Fausey & Boroditsky, 2011) or Japanese speakers are (Fausey et al., 2010). As described by Rissman and Majid (2019), a nativist view of event roles is only tenable under the assumption that roles have prototype structure, and it is the prototypes that are universal. From a nativist perspective, the specific pattern manifest in English, whereby natural forces such as wind are expressible as Subjects, would exemplify linguistic variation at the periphery of semantic and conceptual categories with a shared, universal prototype. If, however, the properties characterizing role prototypes differ across semantic and conceptual categories, this would suggest that role prototypes can be language-specific, casting doubt on the nativist view.

If event roles are tightly linked across semantic and conceptual domains, the prediction follows that Agent and Patient categories have similar internal structure in each domain. In the linguistics literature, roles are often analyzed as having prototype structure: for example, that the prototypical Agent is sentient, intentionally causing an event to occur (Ackerman & Moore, 2001; Dowty, 1991; Grimm, 2011; Hopper & Thompson, 1980; Lakoff & Johnson, 1980; Luraghi, 1995; Primus, 1999). Dowty (1991) analyzes how the arguments of English verbs map to Subject and Object position, proposing a range of properties that characterize Proto-Agents and Proto-Patients. For example, causing an event to occur is a Proto-Agent property. In Experiments 2–4, we ask whether these properties which are relevant to the linguistic expression of Agents and Patients in English also characterize the internal structure of the categories that English speakers extract from visual events.

Approach

We asked two questions: (a) Do English speakers explicitly categorize visual events in terms of Agent and Patient roles? (b) Are the Proto-Agent and Proto-Patient properties that are relevant to English syntax also definitive of the categories that English speakers extract from visual events? We investigated these questions in five experiments. In Experiment 1, participants completed a free sorting task: they were given a variety of pictures of Agents acting on Patients and were asked to group these pictures into two piles. This task reveals the event dimensions that participants find most salient for categorization, our main interest being whether participants spontaneously categorize pictures in terms of Agent and Patient roles. In Experiments 2–4, we trained participants to sort pictures into two categories based on Agent and Patient roles with trial-by-trial feedback, examining the ease with which people extracted the Agent and Patient categories. We addressed our second question by analyzing how participants in Experiments 2–4 generalized the categories they learned to new pictures of Agents acting on Patients (i.e., in a test phase following the training phase). In the test phase, the prototypicality of the Agents and Patients varied according to previous linguistic analyses of role prototypicality, particularly Dowty (1991). If the categories that participants extract from visual events are structured in terms of the Proto-properties proposed by Dowty (1991), then we expect that these Proto-properties will predict participants' accuracy, decision confidence, and reaction time (RT) at test. Finally, in Experiment 5, we asked whether participants are able to categorize visual scenes based on their valence, assessing the robustness of our method for training category formation.

We used a common set of 64 visual scenes for Experiments 1–5. These scenes showed cartoon illustrations of two human-like figures interacting. Scenes varied as to whether one figure was acting upon the other rather than the two figures jointly engaged in an action, a dimension we refer to as "transitivity" (see Experiment 1 Design and Materials). Scenes also varied as to their valence—whether the interaction was positive or negative. In Experiment 1, participants sorted scenes that were highly transitive but varied in their valence. In Experiments 2–4, participants were trained on scenes that were both transitive and negative (Experiment 2), scenes that were transitive but varied in their valence (Experiment 3), or scenes that spanned the entire range of stimuli (Experiment 4). This variation allows us to examine how generalization is impacted by the variability (diversity) of the training set.

Experiment 1: Free Sorting

Method

Participants

Forty-two native English-speaking undergraduates at the University of Wisconsin–Madison were tested (N_{female} = 21; age range = 18–20, median age = 18). Participants received course credit for completing the study. In this and subsequent experiments, participants gave informed consent; all studies were approved by the University of Wisconsin–Madison Institutional Review Board.

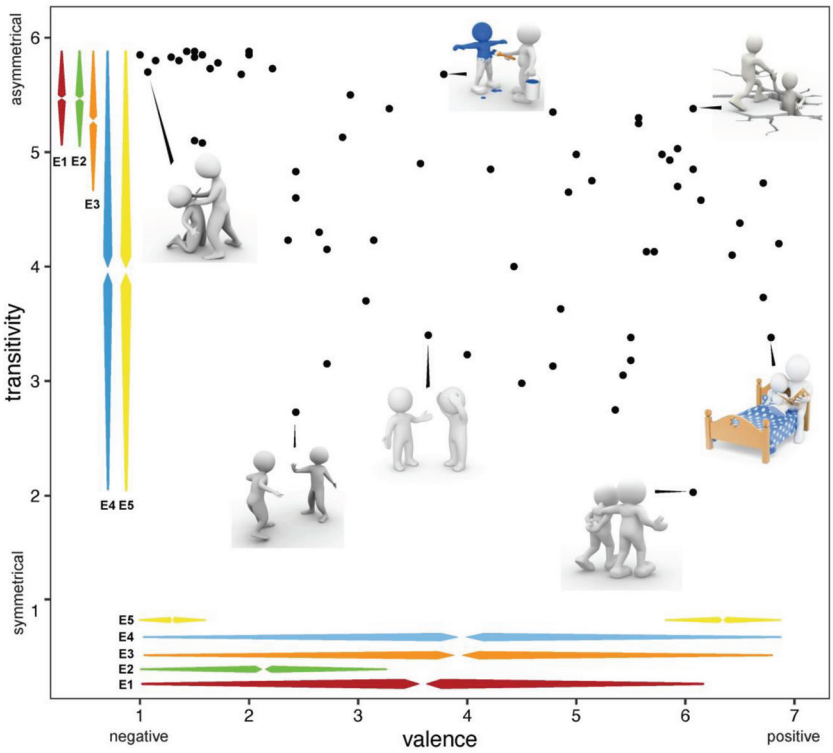
Design and Materials

We obtained a variety of cartoon illustrations from an online stock images database. We prenormed these illustrations based on a prompt about "transitivity." Forty adult English speakers on Amazon Mechanical Turk were instructed that "transitive" events involve one person initiating an action and the other person receiving the action, where the initiator intentionally causes a change in the second person. Participants then decided on a scale from 1 (*symmetrical*) to 6 (*asymmetrical*) whether each scene was transitive. We also prenormed the scenes for their valence—whether the interaction between the figures was positive or negative. Fourteen English speakers on Amazon Mechanical Turk rated the valence of each scene on a 1 (*negative*) to 7 (*positive*) scale. Figure 1 shows the distribution on these two dimensions of the 64 scenes that were used across Experiments 1–5, as well as examples of the

[2] Although English allows a wider range of participant types to appear as Subject than many other languages do, it is not the case that animacy is irrelevant to Subjecthood in English. For example, sentences with inanimate Subjects are more acceptable to the degree that an external causative entity can be pragmatically inferred (Schlesinger, 1989).

Figure 1

Distribution of 64 Scenes Used in Experiments 1–5 With Respect to Valence (1 = Negative; 7 = Positive) and Transitivity (1 = Symmetrical; 6 = Asymmetrical)



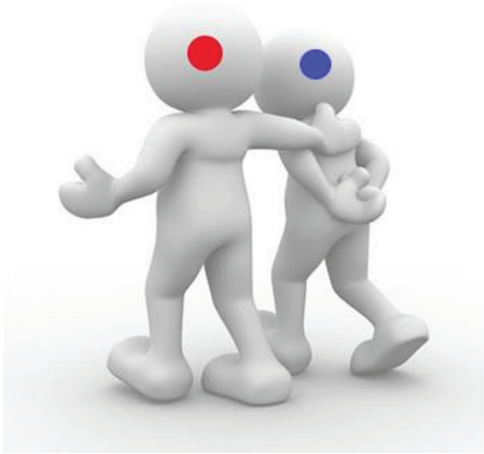
Note. Each point corresponds to a scene. Arrows indicate which scene each example picture corresponds to. Colored brackets show the transitivity and valence distributions for Experiment 1 scenes and Experiments 2–5 training scenes (e.g., the red bracket corresponds to the distributions of Experiment 1 scenes). See the online article for the color version of this figure.

scenes themselves. The figures lacked human features such as faces, hair and clothing; we reasoned that removing such features would help participants attend to the relations between the figures as the basis for categorization.

A hallmark of event representation is that events can be construed in different ways: for example, the same visual scene could be construed as an event of chasing or an event of fleeing, depending on one's perspective (DeLancey, 1991; Fisher et al., 1994; Kuchinsky, 2009). Differences in construal can affect whether the participants in the scene are interpreted as Agents or Patients—the Patient of a chasing event is the Agent of a fleeing event. To quantify such differences in construal, we collected descriptions of each of the 64 scenes from 15 English speakers on Amazon Mechanical Turk. For this description task, one of the figures was marked with a red dot and the other with a blue dot (see Figure 2). Participants described what they saw happening in the picture. In half of the scenes viewed by each participant, the Agent was marked with a red dot; in the other half, the Agent was marked with a blue dot. We coded each description as to whether the Agent in the description matched our assessment of which figure was the Agent. For example, our intuition is that in Figure 2, the figure with the red dot is the Agent. Most speakers agreed with this assessment, producing descriptions like *the person with the red dot hugs and talks to the person with the blue dot.* However, some participants described an alternate construal in which the figure with the blue dot is the Agent, as in *the person with the blue dot is listening to the person with the red dot.* Sentences such as these were coded as mismatches. Symmetrical constructions such as *blue and red dots are walking* were also coded as mismatches. Averaging across the descriptions for each scene, we calculated the proportion of descriptions where our assessment of the Agent matched the Agent in the description ("Description Match"). Description Match values serve as a measure of the Agent/Patient ambiguity of the scenes—mean values for Experiments 1–5 are shown in Table 1. Appendix A lists transitivity, valence, and Description Match values for each scene.

For Experiment 1, we selected 24 scenes that each had a transitivity rating of 5.1 or above, the goal being that the figures clearly exemplified Agent and Patient roles. Mean Description Match for these scenes was 99.7%. In each picture, either the Agent or the Patient was marked with a red dot (see Figure 3). For each scene, we generated four pictures, counterbalancing (a) whether the Agent or the Patient was marked with a red dot and (b) whether the Agent appeared on the left or right side of the image. Each participant saw each of the 24 scenes once and was given either the Agent-dot or Patient-dot version of each scene (12 pictures for each version). The scenes that made up the sets of Agent-dot and Patient-dot pictures were selected at random. The pictures were

Figure 2
Example Stimulus Image From Sentence Description Task



Note. For each image, participants were asked "What is happening in the picture?." See the online article for the color version of this figure.

printed on cardstock and were 3" × 3" in size. Stimuli, data files, and analysis scripts for all experiments are available at: https://osf.io/a5cev/ (Rissman & Lupyan, 2020). The design and hypotheses for this and subsequent experiments were not preregistered.

Procedure

Participants were tested individually while seated at a desk. The experimenter shuffled the 24 cards depicting an Agent/Patient interaction and gave them to the participant in a single pile. The experimenter then told the participant:

> Your task is to sort these pictures into two piles based on how they are similar to each other. The pictures in each pile should share something in common with each other that is not shared by the pictures in the other pile. You can sort the pictures however you would like. Your two piles do not need to have equal numbers of pictures. You can take as much time as you would like.

After the first sort, the experimenter asked the participant to type an explanation for their sort on a computer. This procedure was repeated two additional times; for the second and third sorts, the experimenter asked the participant to sort the cards in a different way than they had previously. The experimenter recorded which cards were placed in which piles after each sort. If, after these three sorts, the participant did not produce an Agent versus Patient sort, the experimenter sorted the cards into Agent/Patient piles and presented them to the participant, saying: "Here is another way of sorting the pictures. What makes one pile different from the other pile?" The experimenter asked the participant to type their explanation on the computer. After the experiment, the experimenter debriefed the participant regarding the purpose of the study.

Results

Participants spontaneously sorted the pictures along 14 different dimensions. Table 2 shows, for each dimension, the proportion of participants who produced that dimension at some point over the course of the three sorts. For each dimension, Table 2 also shows an example explanation provided by one of the participants. Separating the cards into Agent and Patient piles was the second most common strategy overall, with 52% of participants spontaneously producing this sort.

Figure 4 shows the frequency of the five most common dimensions for each of the three sorts. For the first and second sorts, the most common strategy was to distinguish "harming" scenes, such as one figure punching another, from "helpful" scenes, such as one figure performing CPR on another. For the third sort, the most common strategy was to distinguish scenes where the two figures were interacting with a prop (e.g., one figure chasing another with a hammer) from scenes where no prop was present.

Across all participants, 48% never used the Agent versus Patient strategy. When these participants were given Agent and Patient piles and were asked to explain how the cards were sorted, only 50% of these participants correctly identified the Agent versus Patient dimension. The remaining 50% either gave incorrect explanations or simply said that they did not know what the relevant dimension was.

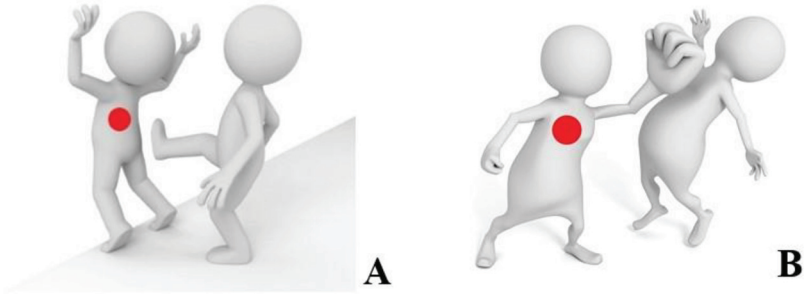
Discussion

Based on previous experimental work showing that children and adults represent visual events in terms of abstract Agent and Patient roles, we expected that these roles would be among the dimensions that adults would use when sorting the pictures. This prediction was only partially met. On the one hand, 24% of participants used the Agent versus Patient strategy as their first sort—for these participants, the event roles were highly prominent conceptually. On the other hand, only 52% of participants used this strategy spontaneously across three sorts. Most strikingly, 24% of participants were unable to explain the Agent versus Patient strategy even when it was provided to them. For this latter group of participants, it appears that event roles were not available as dimensions relevant to the task. In Experiments 2–4, we ask how prominent

Table 1
Number of Unique Scenes and Minimum, Maximum, and Mean Construal Ambiguity (Description Match) Values for the Scenes Used in Each Experiment

Experiment	*N*	Min	Max	*M*
Experiment 1	24	0.933	1	0.997
Experiment 2 training	20	0.933	1	0.997
Experiment 3 training	28	0.867	1	0.983
Experiment 4 training	60	0.533	1	0.938
Experiment 5 training	24	0.733	1	0.978

Figure 3
Example Stimulus Images From Experiment 1



Note. In panel B the red dot marks the Agent; in panel A, the red dot marks the Patient. See the online article for the color version of this figure.

the Agent versus Patient distinction is when supported by explicit corrective feedback in the context of a category learning task.

Experiment 2

The Agent versus Patient distinction was a salient categorization dimension for only about half of the participants in Experiment 1. In Experiment 2, we ask whether a higher proportion of participants will show sensitivity to this dimension in the context of a trained category learning task. We also ask whether Dowty's Proto-properties predict how participants generalize the Agent/Patient distinction when presented with novel scenes.

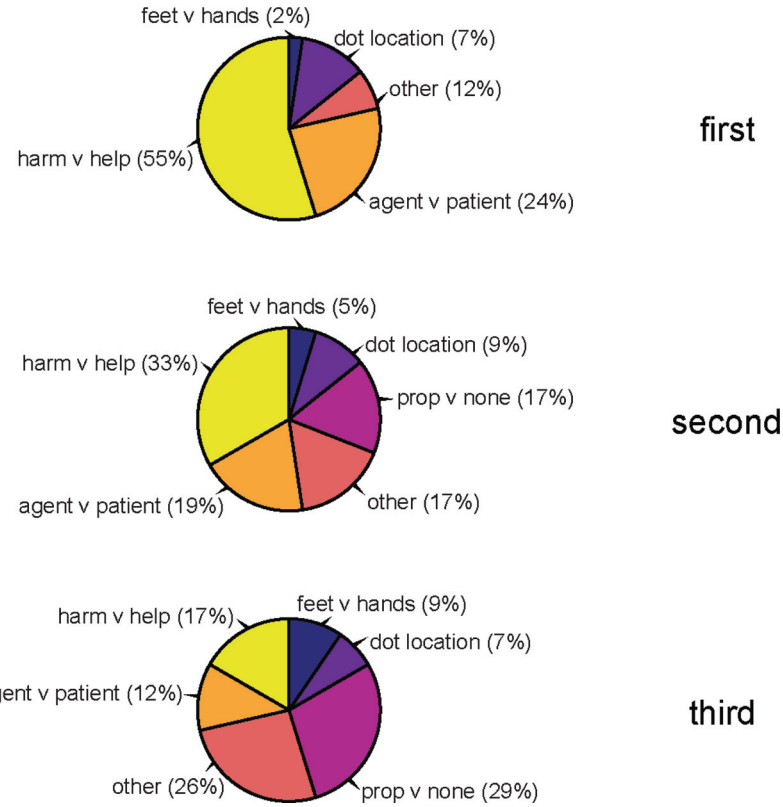
In linguistic theory, the set of constraints determining how a verb's arguments are expressed syntactically is called *argument realization* (see Levin & Rappaport-Hovav, 2005, for review). The category of event participants that can appear as the Subject of an English active transitive sentence appears to be a cluster concept rather than a category defined in terms of necessary and sufficient

Table 2
Proportion of Participants Who Used Each Strategy at Some Point Over Three Free Sorts

Sort type	Example explanation	Proportion of participants
Harm vs. Help	One pile of pictures were all violent actions, the other pile of pictures was all nonviolent acts	0.93
Agent vs. Patient	I sorted the cards based on if the figure with a red dot present was doing an action to another figure in one pile and if an action was being done to the figure with a red dot in the other pile.	0.52
Prop vs. None	This time, I sorted them based on whether there was some kind of prop involved. If there was a prop involved, I made one pile. If there was no prop, I made another pile.	0.45
Dot location	I sorted them by where the red dot was on the figures body. one group had the dot on the head, the other on the figures chest.	0.26
Feet vs. Hands	I sorted the pictures based on which pictures involved actions using hands and a different pile that involved actions using feet or legs.	0.17
Touching vs. Not	Based on if the two characters in the photos were touching each other	0.12
Red only vs. Multicolored	The first group was the majority of the cards which only had red, the other group had less cards that also had blue on the cards.	0.10
Figures: same vs. diff size	One pile had cards in which both figures were the same large size (presumably adults) and the other had cards in which at least one figure was small (presumably children).	0.07
Kick/punch vs. Not	I sorted one pile into pictures that portrayed people kicking and punching people and the other pile into all the pictures that didn't show kicking and punching.	0.07
Completed vs. Not	One pile was that the action was already received, and the person is reacting to the it and the second pile is when the person is in the process of receiving the action and hasn't reacted, yet.	0.02
Dot: left vs. right	I sorted the pictures by the placement of the person with the red dot: one stack where the person with the red dot was on the left, the other stack was for when the person was on the right.	0.02
Vertical vs. Horizontal	whether the action was horizontal or vertical	0.02
Upper body vs. Lower body	based on if the action was performed with upper or lower body	0.02
Motion vs. Static	I sorted the pictures with one pile with pictures that showed both figures in motion or engaging in the activity and the other pile with pictures that showed one figure not in motion or not engaging in the activity	0.02

Figure 4

Proportion of Sorts Falling Into the Five Most Common Strategies Across Three Rounds of Sorting



Note. See the online article for the color version of this figure.

conditions. Sentience, for example, has been argued to be relevant to argument realization in English, allowing *Josh* to surface as Subject in *Josh believed the news.* For an argument to appear as Subject, however, sentience is neither necessary (cf. *the rock broke the window*) nor sufficient (cf. *the sound frightened the teenagers*). Data such as these led Dowty (1991) to propose a Proto-property theory of English argument realization. In his proposal, Proto-Agent properties such as sentience and intentionality guide which of a verb's arguments appears as Subject: an argument need not have all of the Proto-Agent properties to appear as Subject, just more Proto-Agent properties than the other argument(s) of the verb. Table 3 shows the set of Proto-Agent and Proto-Patient properties proposed by Dowty, as well as examples of verbs where the Subject/Object has that Proto-property.

Josh believed the news is an acceptable English sentence because *Josh* has the Proto-Agent property of sentience, whereas *the news* has no Proto-Agent properties. The argument with the most Proto-Patient properties surfaces as Object, predicting argument realization for verbs with more than two arguments or for verbs whose arguments are matched with respect to their Proto-Agent properties.

The Proto-properties in Table 3 have been validated experimentally by Kako (2006) who asked English speakers to evaluate Dowty's Proto-properties in sentences with nonce content words (e.g., *the grack mecked the zarg*). Speakers rated the likelihood of each Proto-property being true of each argument (e.g., *how likely is it that the zarg chose to be involved in mecking?*). Kako found that overall, Subjects were rated higher on Proto-Agent properties than Objects and Objects were rated higher on Proto-Patient properties than Subjects. Across a series of experiments with both nonce and real verbs, Kako also found evidence for the validity of each individual Proto-property (Proto-Agent properties rated more highly for Subject than Object; Proto-Patient properties rated more highly for Object than for Subject). At the same time, the effect was stronger and more consistent for some properties than for others: volitional involvement (i.e., choosing to be involved) was a stronger Proto-Agent property than independent existence.

Dowty's Proto-properties have been further validated by Reisinger et al. (2015). These researchers extended Kako's rating task to real sentences from the PropBank corpus (Palmer et al., 2005), asking annotators to evaluate the Proto-properties for more than 9,000 arguments and 5,000 verb tokens. Their results were similar to those in Kako (2006): Proto-Agent properties were rated higher for Subjects, and Proto-Patient properties were rated higher for Objects. In addition, each Proto-property was individually predictive, with the exceptions of movement and being stationary. The contrast between Subject and Object ratings was particularly high for the Proto-Agent properties of volition and causation ("instigation" in Reisinger et al.'s terminology). In sum, there is strong support for many of the properties in Table 3 as being definitive of

Table 3
Proto-Agent and Proto-Patient Properties Proposed by Dowty (1991)

Proto-Agent properties	Proto-Patient properties
Volitional involvement in the event or state	Undergoes change of state
Mary is ignoring John	*Mary erased the error*
Sentience (and/or perception)	Incremental theme
John believes Mary	*John crossed the driveway*
Causing an event or change of state in another participant	Causally affected by another participant
His loneliness causes his unhappiness	*Mary broke her bat*
Movement (relative to the position of another participant)	Stationary relative to movement of another participant
The rolling tumbleweed passed the rock	*The bullet entered the target*
Exists independently of the event named by the verb	Does not exist independently of the event, or not at all
John needs a new car.	*Mary build a house*

Note. The underlined word/phrase indicates the argument that has the Proto-property in the preceding line.

Agent and Patient categories in English, as far as syntactic argument realization is concerned.

In Experiment 2, we tested whether the Proto-properties proposed by Dowty predict participants' accuracy, decision confidence, and RT when participants are tasked with generalizing the Agent/Patient categories they induced during training. We trained participants on scenes rated as highly transitive, as described in Experiment 1 Design and Materials—in other words, on scenes involving prototypical Agents and Patients. At test, participants viewed scenes that varied in terms of their transitivity. As the test phase required participants to generalize beyond the particular items they viewed in training, we assume that participants' patterns of generalization reveal their implicit knowledge of Agent and Patient categories. If the Proto-properties are predictive for nonlinguistic stimuli, this would suggest that the Agent and Patient roles have similar internal structure across semantic and conceptual domains. In Experiment 2, participants were trained only on negatively-valenced scenes but viewed both positive and negative scenes at test. Since valence was the most salient dimension of the stimuli in Experiment 1, we suspected that including both positive and negative scenes during training would make it more difficult for participants to induce the Agent/Patient categories, as participants must ignore the valence dimension.

Method

Participants

We tested 202 adult native English speakers on Amazon Mechanical Turk (N_{female} = 83, N_{male} = 117, age range = 21–70, median age = 34). An additional 23 participants were tested but were excluded for failing to pass at least 11 of 12 attention checks. Participants received $1.50 for completing the study and received a $.50 bonus if their overall training accuracy was $\geq$ 70%.

Design and Materials

Participants categorized 64 cartoon illustrations of one human-like figure interacting with another (see examples in Figure 1). In this and subsequent experiments, participants viewed 24 training trials: the Agent-dot and Patient-dot variants of each of 12 different scenes. In Experiment 2, participants viewed 52 test trials. The training and test trials were structured in the following way: the 12 training scenes were selected randomly from a set of 20 scenes that were highly transitive and negatively valenced. The 52 test trials contained three types of scenes: (a) the remaining eight transitive/negative scenes that were not part of the training set, (b) eight scenes that were transitive and positive, and (c) 36 scenes that were rated as having lower transitivity and varied valence. We structured the stimuli in this way to ensure that the test scenes varied in terms of both their transitivity and their valence. We do not analyze test performance in terms of these discrete categories (e.g., transitive/positive), however. Instead, we analyze test performance in terms of a range of continuous scene dimensions, valence being one of them. These continuous dimensions are described in Stimuli Norms. Across all participants, each of the 20 transitive/negative scenes was viewed in both the training and test phases. Appendix B lists which scenes were included in the training and test phases of Experiments 2-5.

In this and subsequent experiments, only one dot variant was shown for each of the test scenes (Agent or Patient). In addition, dot variant (Agent-dot vs. Patient-dot) and Agent side (left/right) were counterbalanced in both the training and test phases.

Procedure

For the training phase, participants were instructed that they should assign each picture to one of two categories: Category A and Category B. They were told: "in each picture, one individual will be marked with a red dot. The individual with the red dot determines which picture goes in which category." Participants pressed the 'a' key for Category A and the 'b' key for Category B. In each trial of the training phase, participants were given feedback (correct vs. incorrect). If a response was incorrect on a given trial, participants needed to press the correct key to proceed to the next trial. Whether Category A corresponded to Agent or Patient was counterbalanced between participants. At the end of 24 training trials, participants were asked "how would you describe the categories A and B that you learned?" and were asked to rate their confidence in their response on a scale from 1 to 5.

For the test phase, participants were told that they would see different pictures and that they would need to choose whether the pictures belonged to the A/B categories they learned during the first phase. After making each response, participants were asked to indicate their confidence on a 1–5 scale. No feedback was given on test trials. At the end of the test trials, participants were asked whether their understanding of the categories A and B had changed and if so, how.

Coding

We coded participants' explanations of the A/B categories as correct if it included any mention of a difference in roles between

Agents and Patients. For example, the following explanations were coded as correct: "A is the person being attacked. B is the person attacking," "Category A is someone who is an aggressor, violent, or bully. Category B is someone who is a victim," "aggressor and victim." Explanations were coded as incorrect if they characterized a dimension of the events irrelevant to roles, for example: "violent and mean," "A is not touching, B is touching." Explanations were coded as uninformative if they did not characterize any properties of the events, for example: "unsure," "it was awesome."

Stimuli Norms

Each of the 64 scenes was normed on 11 dimensions; these are shown in Table 4, along with the number of speakers who provided the norms and the specific prompt question they were asked. The raters were English speakers tested on Amazon Mechanical Turk. Raters viewed both Agent-dot and Patient-dot versions of the images, excepting the valence norm, where neither figure was marked with a dot. Pictures were rated on a scale from 1 to 5, except for the valence norm, where raters used a scale from 1 to 7.

We collected data for three Proto-Agent properties (volition, causation, and movement) and three Proto-Patient properties (change, affectedness, and being stationary). We did not test the Proto-Agent property of sentience, because we expected that participants would use an animate/inanimate distinction as a primary basis for categorization, making it more difficult for them to learn the role categories. As we showed illustrations of events with two figures, we also did not test the properties of independent existence/lack of independent existence. Following Kako (2006) and Reisinger et al. (2015), we did not test incremental theme as a Proto-Patient property. An argument is an incremental theme if the stage of completion of the event maps homomorphically to the degree of change of the argument—for example, in *Christian ate the apple*, the apple is an incremental theme. This property depends on the semantics of the verb and is therefore difficult to test for the visual illustrations we showed.

We also collected data on five dimensions of the stimuli that were not part of Dowty's original proposal but could be relevant to the categories that participants construct given illustrations of events. The first is the valence of the event: participants' role categories could specifically encode whether the Agent is acting in a negative or positive way (e.g., one figure kicking another vs. one figure performing CPR on another). Following Hafri et al. (2013), we also collected data on the body postures of the Agent and Patient in each event. These authors found that participants performed worse in a visual role recognition task when the postures of the Agent and Patient mismatched their semantic role (e.g., a Patient with outstretched arms that was leaning toward the Agent). Such perceptual features may also shape the categories that English speakers construct from visually depicted events.

Following the analytical approach in Kako (2006) and Reisinger et al. (2015), we calculated a difference score for each scene on each Proto-property by subtracting the Agent score from the Patient score for that scene. For example, the Agent in the scene in Figure 3A received a score of 4.4 on the Intention dimension ("To what extent did the dot-figure choose to be involved in the interaction?") whereas the Patient in the scene received a score of 1.3 on this dimension; therefore, the Intention difference score for this scene is 3.1. Note that because we subtract the Agent score from the Patient score for all Proto-properties, many of the scenes have negative difference scores for the Proto-Patient properties. For the four body posture norms in Table 4, we calculated Agent–Patient difference scores for each norm and then averaged the four difference scores, resulting in a single Body Posture score for each scene. Figure 5 shows the correlations between each pair of norms across all 64 scenes.

Results

We analyzed the results from Experiments 2–5 using mixed-effects regression models with by-subject and by-scene random intercepts and, as appropriate, by-scene random slopes. We computed these models using R (R Core Team, 2017) and the *lme4* package (Bates et al., 2014). For linear models, we used the *lmerTest* package (Kuznetsova et al., 2017) and Satterthwaite approximation to compute *p*-values for fixed effects (see Luke, 2017).

Training Phase: Did Participants Induce Agent/Patient Categories?

We assessed whether participants induced the Agent/Patient distinction through two measures: accuracy on the last eight (of 24) training trials and whether a correct explanation was provided at the end of the training phase. *M* accuracy on the last eight training trials was 75.8% (95% CI [72.3%, 79.3%]). Figure 6 shows training accuracy broken down by explanation type (correct, incorrect, uninformative) and shows the proportion of participants who provided each type of explanation. In Experiment 2, 59% of participants provided correct explanations after the training phase. Most participants in Experiment 2 who gave correct explanations had training accuracy of 75% or higher; most participants who did not give a correct explanation had chance-range accuracy. Several participants gave a correct explanation but scored 50% or below—these responses may reflect mistakes such as switching the mapping between role and category. There were also a few participants who gave incorrect or uninformative explanations but had 100% accuracy. We think the most likely cause of this pattern is participants who did not put in a good faith effort in producing the description, but it is also possible that they learned the categories well enough to produce perfect responses without being able to describe them clearly in writing.

To assess whether training accuracy differed across explanation types, we fit a mixed-effects logistic regression model predicting accuracy from Explanation type (baseline = uninformative). Participants who gave correct explanations were more accurate than participants who gave uninformative explanations ($b = 2.5$, $SE = .20$, $p < .001$). Accuracy did not differ between people who gave incorrect and uninformative explanations ($b = -.14$, $SE = .29$, $p = .55$). Training accuracy was worse for participants giving incorrect than correct explanations ($b = -2.7$, $SE = .24$, $p < .001$; setting baseline level to "correct"). To summarize the training phase results, we found that accuracy was higher than expected by chance, and that most participants who scored accurately were able to explicitly explain the Agent and Patient categories.[3] With some exceptions, participants who did not provide correct

[3] In Experiment 2 and subsequent experiments, we found no effect of gender, age, or highest level of education on participants' training accuracy or on whether participants provided a correct explanation for the categories.

Table 4
Norms Collected for Stimuli Used in Experiments 1-5

Norm category	Variable name	Prompt	*N* raters
Valence	Valence	How positive or negative is the interaction between the two figures?	14
Proto-agent	Intention	To what extent did the dot-figure choose to be involved in the interaction?	33
Proto-agent	Causation	To what extent did the dot-figure cause the interaction to happen?	27
Proto-agent	Movement	To what extent does the dot-figure change location during the interaction?	28
Proto-patient	Change	How much does the dot-figure change as a result of the interaction?	31
Proto-patient	Affectedness	How much is the dot-figure affected by the other figure?	33
Proto-patient	Stationary	To what extent is the dot-figure stationary during the interaction?	30
Body posture	Head	To what extent is the head of the dot-figure facing toward the other figure?	24
Body posture	Body	To what extent is the body of the dot-figure facing toward the other figure?	24
Body posture	Limbs	To what extent are the arms or legs of the dot-figure outstretched toward the other figure?	24
Body posture	Leaning	To what extent is the dot-figure leaning toward the other figure?	23

characterizations of the categories had near-chance accuracy. That these participants passed the attention checks rules out the possibility that their low performance was attributable to failing to engage with the task entirely.

Test Phase: How Did Participants Generalize the Categories?

Overall Test Performance. Test accuracy was 87.7% (95% CI [85.0%, 90.4%]) for participants giving correct explanations, 55.7% (95% CI [50.9%, 60.5%]) for incorrect explanations and 53.2% (95% CI [49.0%, 57.4%]) for uninformative explanations.[4] A mixed-effects logistic regression showed that test accuracy did not differ between participants giving incorrect and uninformative explanations ($b = .12$, $SE = .24$, $p = .62$). Because individuals giving incorrect versus uninformative explanations did not differ in their test accuracy in this or any of Experiments 3–5, we combined the data from these two groups in subsequent analyses of overall test performance.

To test whether Dowty's Proto-properties predict generalization in the test phase, we analyzed the test data from Experiment 2 including participants who gave correct explanations after training as well as participants who gave noncorrect explanations, but whose training accuracy was at least 87.5% (seven out of eight trials correct).[5] For this and subsequent experiments, we refer to the group of participants who met at least one of these two criteria as "learners."

For the set of learners in Experiment 2 ($N = 125$), mean test accuracy was 86.8% (95% CI [84.1%, 89.5%]). For the nonlearners, test accuracy did not differ from chance: $M = 52.3\%$ (95% CI [49.4%, 55.2%]. Figure 7 shows means and individual participant data for each dependent variable (accuracy, confidence rating, and RT) for the set of learners in each experiment. Owing to our manipulation of the training items across Experiments 2–5, slightly different sets of scenes were viewed at test in each experiment—Figure 7 only includes the 55 scenes that were common across Experiments 2–5, enabling more meaningful comparisons. See the OSF repository for this article for figures that include all the items and data for nonlearners.

For this and subsequent analyses of the test phase, we excluded trials where the RT exceeded 10 seconds, which was roughly double the median RT across all experiments.[6] We did not set a minimum RT inclusion threshold. In addition, in our analyses of test confidence and RT, we only included trials where participants responded correctly.

Do the Proto-Properties Predict Test Performance? We now turn toward the question of whether Dowty's Proto-properties predict generalization in the test phase. If the Proto-properties in Table 3 are relevant to the Agent and Patient categories that participants extract in the training phase, we predict that at test, participants will be faster, more accurate, and have higher confidence for scenes that are more prototypical with respect to these properties. We modeled test accuracy, confidence, and RT in a stepwise fashion, where only those predictors with significant model coefficients were retained after each step. Table 5 shows the order in which the predictors were entered. We first included predictors that we did not expect to predict performance: whether the red dot was on the left or right side of the scene (Side) and whether the red dot was on the Agent or Patient (Target). We then included the measure of how often raters' descriptions matched our intuition about which figure in the scene was the Agent (Description Match). This measure controls for the likelihood that different participants construed the Agents and Patients differently in a given scene.

We next computed the similarity between each training item and test item seen by each participant (Train-test Similarity). We represented each scene as a vector encoding the values for each norm in Table 4 (coded as the difference between the Agent and Patient for all norms except Valence). We then computed the Euclidean distance between the vectors representing each test item and each training item. We computed two variants of Train-test Similarity: the median distance between a given test item and all the training items seen by that participant, and the minimum distance between a given test item and the training items (that is, how similar is the most similar item). If both median and minimum distance were (individually) significant predictors in a given model, we used the variant with the largest effect size. To the extent that participants are more likely to be correct on a test item if it is more similar to the training items, this similarity measure should be positively associated with test performance. If the Proto-properties

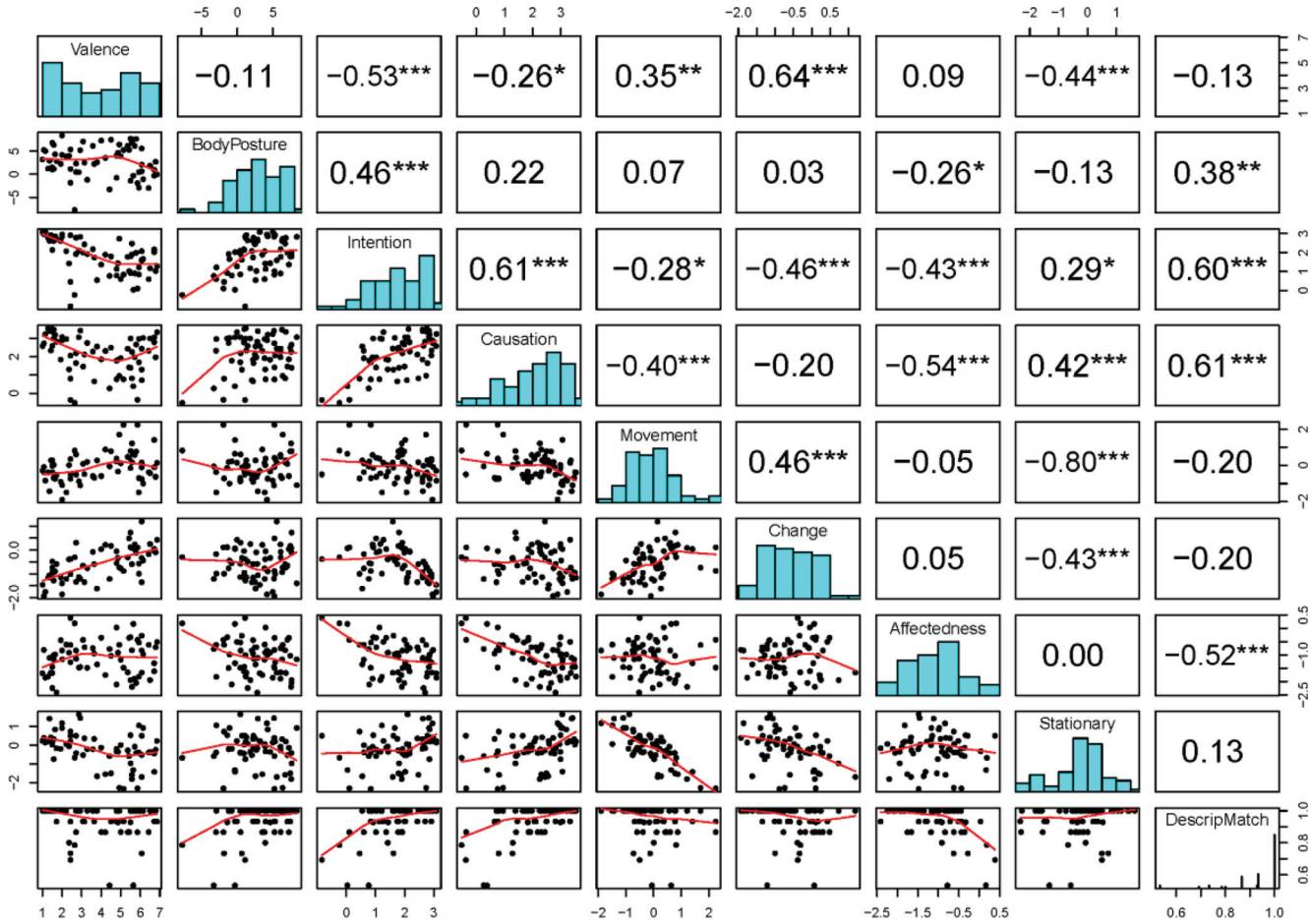
[4] Among participants giving incorrect/uninformative explanations at the end of the training phase, a few gave correct explanations at the end of the test phase: 2% of participants in Experiment 2, 13% of participants in Experiment 3, and 7% of participants in Experiment 4.

[5] The probability of getting seven out of eight trials correct by guessing is 3% assuming a binomial distribution.

[6] Trials that exceeded the 10 second threshold constituted the following percentages of test trials across experiments: 8% (Experiment 2), 7% (Experiment 3), 7% (Experiment 4), 6% (Experiment 5).

Figure 5

Relationships Between Each Pair of Norms



Note. Distributions of each norm are shown on the diagonal. For all norms except Valence and DescripMatch, the value for each scene is the Agent minus Patient difference score. Greater valence indicates more positive valence; greater BodyPosture indicates that the Agent has a more active posture than the Patient; greater Intention indicates that the Agent chooses to be involved in the action more than the Patient; greater Causation indicates that the Agent causes the action more than the Patient; greater Movement indicates that the Agent moves more than the Patient; greater Change indicates that the Agent changes as a result of the action more than the Patient; greater Affectedness indicates that the Agent is affected more than the Patient; greater Stationary values indicate that the Agent is more stationary than the Patient; greater DescripMatch indicates that our choice as to which figure is the Agent is more consistent with English speakers' descriptions. * $p < .05$. ** $p < .01$. *** $p < .001$. See the online article for the color version of this figure.

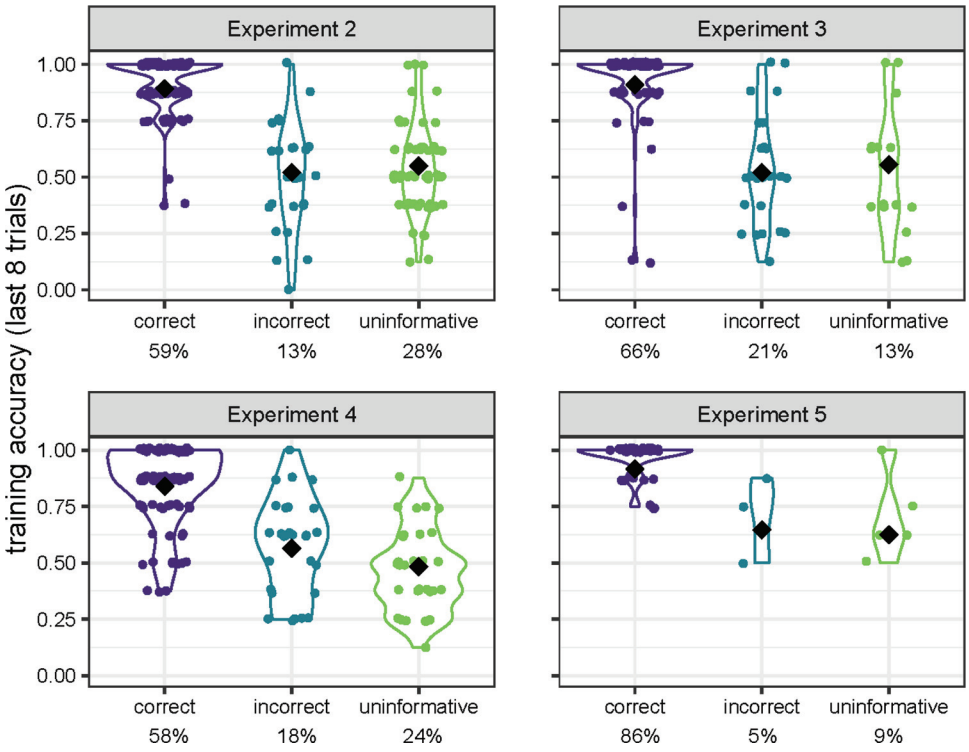
have broad predictive power, we expect they will predict test performance beyond what is contributed by Train-test Similarity. We then added the Valence rating, including Valence as both a linear and a quadratic predictor because Valence may have different effects at the ends than in the middle of the continuum, particularly in Experiment 5. To estimate the effects of Body Posture, we included an aggregate rating of the four body posture norms in Table 4. As the strongest test of whether the Proto-properties are predictive, we included the Proto-properties last in each model.

We standardized (z-scored) all predictors in Table 5 except for the categorical predictors Side and Target. The outcome variables confidence and RT were also standardized. We used logistic regression to model accuracy and linear regression to model confidence and RT. The coefficient estimates for the best-fitting models of test accuracy, confidence, and RT are shown in Figures 8–10, respectively.

Not surprisingly, Description Match—an index of the Agent/Patient ambiguity of the scenes—was associated with higher accuracy, higher confidence, and lower RTs. Participants were also more accurate and more confident when a test item was more globally similar to the scenes that that participant saw during training. Crucially, we also found that some of the Proto-properties predicted performance: the more intentional the Agent was relative to the Patient, the greater were participants' accuracy, confidence, and speed. Larger differences on the Causation dimension predicted better accuracy and larger differences on the Affectedness dimension predicted reduced confidence and slower reaction times. These effects persisted even when Train-test Similarity was controlled for. We found no evidence that the Proto-properties Movement and Stationary predicted performance.

We also found that attributes of the scenes beyond the Proto-properties had an influence: participants were less confident when

Figure 6
Training Accuracy by Explanation Type for Each Experiment



Note. Each dot represents an individual participant; diamonds indicate the mean. The percentage of participants who provided each type of explanation is listed under the label for each experiment. See the online article for the color version of this figure.

the scenes were more positive (Valence), likely reflecting a transfer cost from the training items which included only negatively valenced scenes. Participants were also more confident for scenes at the extreme ends of the valence continuum (most positive and most negative) compared with scenes in the middle, as expected if they are attuned to valence even as they rely on event roles. Reaction times were also slower when the dot was on the Patient than on the Agent (an effect of Target). We found no evidence that the Body Posture attribute influenced participants' generalization.

Discussion

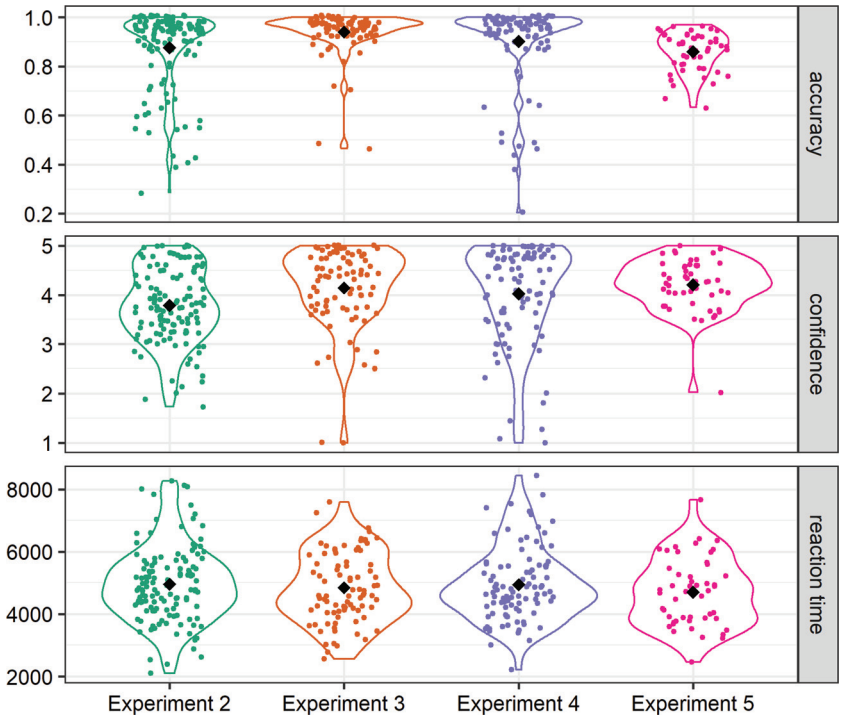
The literature reviewed in the introductory section suggests that Agent and Patient categories are conceptually prominent—we asked whether they are also prominent when people hypothesize what dimensions of visual scenes are relevant to a categorization task. Many of the participants in Experiment 2 did quickly induce the Agent/Patient distinction from the training trials and were able to articulate this distinction verbally. However, 41% of participants gave either incorrect or uninformative explanations for the categories. As a group, these participants had chance-level performance. This finding suggests that for some individuals, event roles do not have a high level of prominence in the context of an explicit categorization task. For participants who did induce the categories, they generalized their category knowledge in ways predicted by Dowty's theory of Proto-Roles. The difference in Intention between the Agent and the Patient was a particularly consistent predictor, even when Train-test Similarity was controlled for.

In Experiment 2, participants were trained on scenes that were relatively homogenous semantically: They all had high transitivity (that is, asymmetry) and were all negatively valenced. In Experiment 3, we ask how expanding the training stimuli to include both negative and positive events affects participants' ability to induce the categories. Expanding the training stimuli also allows us to test the generality of the test phase findings from Experiment 2. If participants' nonlinguistic event categories are structured in terms of the Proto-properties, then the Proto-properties should predict test performance regardless of how the Agent/Patient categories are instantiated during training.

Experiment 3

The most frequent sorting strategy in Experiment 1 was to separate the pictures into harming versus helping categories. Given the salience of valence, we wondered whether including both positive and negative scenes as training items would lead to poorer performance in inducing the Agent/Patient categories—that is, whether representing the positive/negative (that is, helping/harming) meaning of the scenes comes at the expense of representing the figures in terms of their Agent/Patient roles. At the same time, including a wider range of valence—insofar as it increases the

Figure 7
Mean Test Accuracy, Mean Test Confidence, and Mean Test Reaction Time for Each Experiment



Note. Data are shown for "learners" (i.e., those who either gave correct explanations after training or had training accuracy of at least 87.5%, or both). Data are shown only for those 55 scenes that were common across the test phases in all four experiments. Each dot represents an individual participant; diamonds indicate the mean. See the online article for the color version of this figure.

semantic coverage of the training scenes—may translate to superior generalization performance.

Method

Participants

We recruited 119 adult native speakers of English on Amazon Mechanical Turk (N_{female} = 61, N_{male} = 58, age range = 18–68, median age = 33). An additional five participants were tested but excluded for failing attention checks. The participants from Experiment 3 had not taken part in Experiment 2. Participants received $1.50 for completing the study and received a $.50 bonus if their overall training accuracy was ≥ 70%.

Design and Materials

The design was the same as in Experiment 2 except that across the 12 training scenes, six were highly transitive and negative and six were highly transitive and positive (see Figure 1). Each set of six scenes was selected randomly from a set of 14 scenes. As in Experiment 2, the eight transitive/negative and eight transitive/positive scenes that were not encountered during training were viewed at test. We also constructed multiple stimuli orders such that each of these 28 scenes was viewed at both training and test. At test, participants also viewed 28 scenes that were lower in transitivity, completing a total of 44 test trials.

Procedure and Coding

The procedure and coding were the same as in Experiment 2.

Results

Training Phase: Did Participants Induce Agent/Patient Categories?

Figure 6 shows mean training accuracy by explanation type for Experiment 3. Given the importance of scene valence in guiding participants' categories in Experiment 1, we predicted that participants would have more trouble learning the Agent/Patient distinction when the training set included both positive and negative scenes (Experiment 3) than when the training set included only negative scenes (Experiment 2). Following the training trials, 66% of participants in Experiment 3 gave correct explanations of the categories. To test whether explanation success was greater than in Experiment 2, we fit a logistic regression model with the fixed effect Experiment, predicting correct versus incorrect/uninformative noncorrect explanations. The proportion of correct explanations did not differ across experiments (b = .30, SE = .24, p = .21).

Table 5
Order in Which Factors Were Included in Regression Models

Order	Factor
1	Side
2	Target
3	Description Match
4	Train-test Similarity (median or minimum distance)
5	Valence
6	Body posture
7	Proto-agent: intention
8	Proto-agent: causation
9	Proto-agent: movement
10	Proto-patient: change
11	Proto-patient: affectedness
12	Proto-patient: stationary

Mean accuracy on the last eight training trials in Experiment 3 was 79.1% (95% CI [74.1%, 84.1%]), which was in fact numerically higher than training accuracy in Experiment 2 (75.8%). Modeling training accuracy with both Experiment (baseline = Experiment 2) and Explanation type (baseline = uninformative) as predictors, we found an effect of Explanation type such that participants giving correct explanations were more accurate than participants giving uninformative explanations ($b = 2.7$, $SE = .23$, $p < .001$). As before, there was no difference in accuracy between people giving incorrect and uninformative explanations ($b = -.15$, $SE = .28$, $p = .60$). There was no main effect of Experiment ($b = -.04$, $SE = .34$, $p = .91$) or interaction between Experiment and Explanation type (correct: $b = .22$, $SE = .43$, $p = .60$; incorrect: $b = .18$, $SE = .48$, $p = .71$). Training accuracy across Experiments 2 and 3 was lower for participants giving incorrect than correct explanations ($b = -2.8$, $SE = .29$, $p < .001$). These results indicate that inducing the categories was not more difficult when the training set included both positive and negative scenes.

Test Phase: How Did Participants Generalize the Categories?

Overall Test Performance. For participants giving correct explanations ($N = 78$), test accuracy was 93.6%, 95% CI [91.6%, 95.6%]; for those giving incorrect explanations ($N = 25$) it was 58.9%, 95% CI [51.1%, 66.7%], and for those giving uninformative explanations ($N = 16$), it was 51.6%, 95% CI [41.2%, 62.0%]. To test whether test accuracy differed between Experiments 2 and 3, we combined the data from participants giving incorrect/uninformative explanations and fit a model with Experiment (baseline = Experiment 3) and Explanation type (baseline = incorrect/uninformative) as fixed effects, including only the 56 scenes that were common to the test phases in both experiments. Test accuracy was higher for participants giving correct than incorrect/uninformative explanations ($b = 2.24$, $SE = .15$, $p < .001$). Although there was no main effect of Experiment ($b = .13$, $SE = .20$, $p = .51$), there was an interaction such that the difference in test accuracy between participants giving correct versus incorrect/uninformative explanations was greater in Experiment 3 than in Experiment 2 ($b = .63$, $SE = .26$, $p < .05$). This result confirms the prediction that participants who learned broader categories—those including both negative and positive scenes—had better generalization.

Do the Proto-Properties Predict Test Performance? Turning to the question of whether the Proto-properties predict test accuracy, confidence, and RT, we analyzed the test data only for the set of learners (that is, participants who gave correct explanations or whose training accuracy was at least 87.5%, or both). Mean test accuracy for the learners ($N = 85$), was 93.2% (95% CI [91.3%, 95.1%]). For the nonlearners (29% of the participants), test accuracy did not differ from chance: $M = 49.5\%$ (95% CI [44.9%, 54.1%]). Figures 8–10 show coefficient estimates for the best-fitting models of accuracy, confidence rating, and RT.

As in Experiment 2, Description Match (an index of the ambiguity of the scenes) was predictive for all three measures. In addition, accuracy was higher for test scenes that were more globally similar to the training scenes viewed by a particular participant. As in Experiment 2, Proto-properties predicted test performance. Intention was predictive for all three measures: the greater the difference between the Agent's intention and the Patient's intention, the better participants performed on the test trials. Causation marginally predicted accuracy. Unlike for Experiment 2, we observed an effect of Change: the more change the Agent underwent relative to the Patient, the faster were participants' reaction times. We found no effect of the Proto-properties Movement or Stationary, as in Experiment 2.

We found several effects in Experiment 3 that we did not predict: an effect of dot placement (Target) such that accuracy and confidence were higher when the dot was on the Agent, and an effect of Side such that accuracy was lower when the Agent was on the right side of the image. As in Experiment 2, we found no effect of Body Posture.

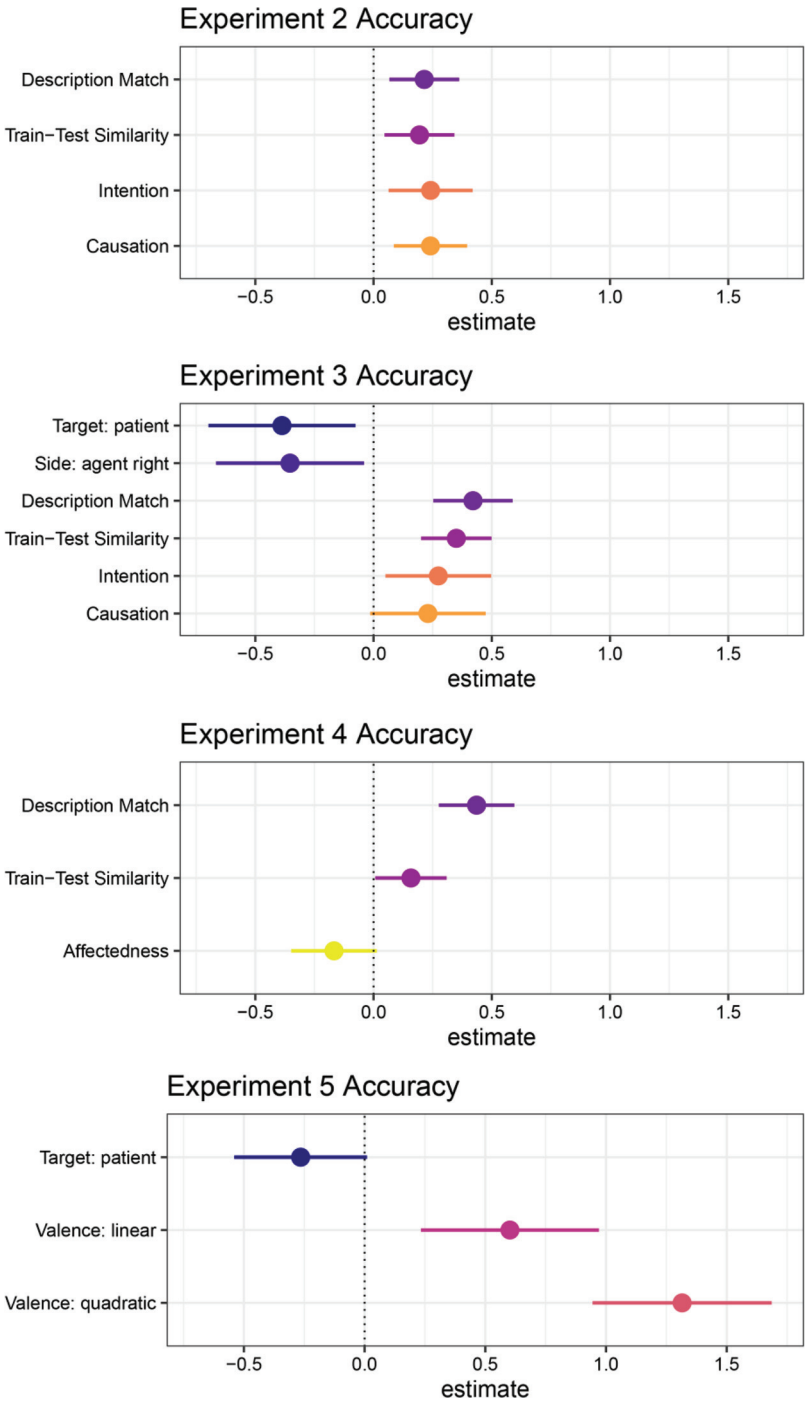
In Experiment 3, we found a small main effect of Valence on confidence, but this time in the opposite direction: participants were more confident when the scenes were more positive. Confidence was also greater for scenes with more extreme valence values. As a direct test of whether the valence of the scenes affected participants' confidence in different ways across experiments, we combined the data from Experiments 2–3 and modeled confidence with the following predictors: Description Match, Train-test Similarity, Valence, Intention, Affectedness, Experiment, and the interaction between Experiment and Valence. Overall, confidence was higher in Experiment 3 than Experiment 2 ($\beta = .32$, $SE = .10$, $p < .01$), confidence was higher for scenes that were more negative ($\beta = -.14$, $SE = .021$, $p < .001$), and confidence was higher for scenes with more extreme valence values ($\beta = .076$, $SE = .017$, $p < .001$). There was a significant interaction between the linear Valence term and Experiment such that more positively valenced scenes increased confidence more in Experiment 3 than in Experiment 2 ($\beta = .18$, $SE = .017$, $p < .001$). There was no significant interaction between the quadratic Valence term and Experiment ($\beta = .0056$; $SE = .015$, $p = .72$).

Discussion

As in Experiment 2, we found that many participants were able to induce the Agent/Patient categories from the training items, but a sizable percentage (29%) were not. For those who did learn to criterion, they generalized more broadly than participants in Experiment 2 did. In addition, whereas participants who were trained on only negative scenes (Experiment 2) reported reduced confidence for scenes that were more positive, we found the

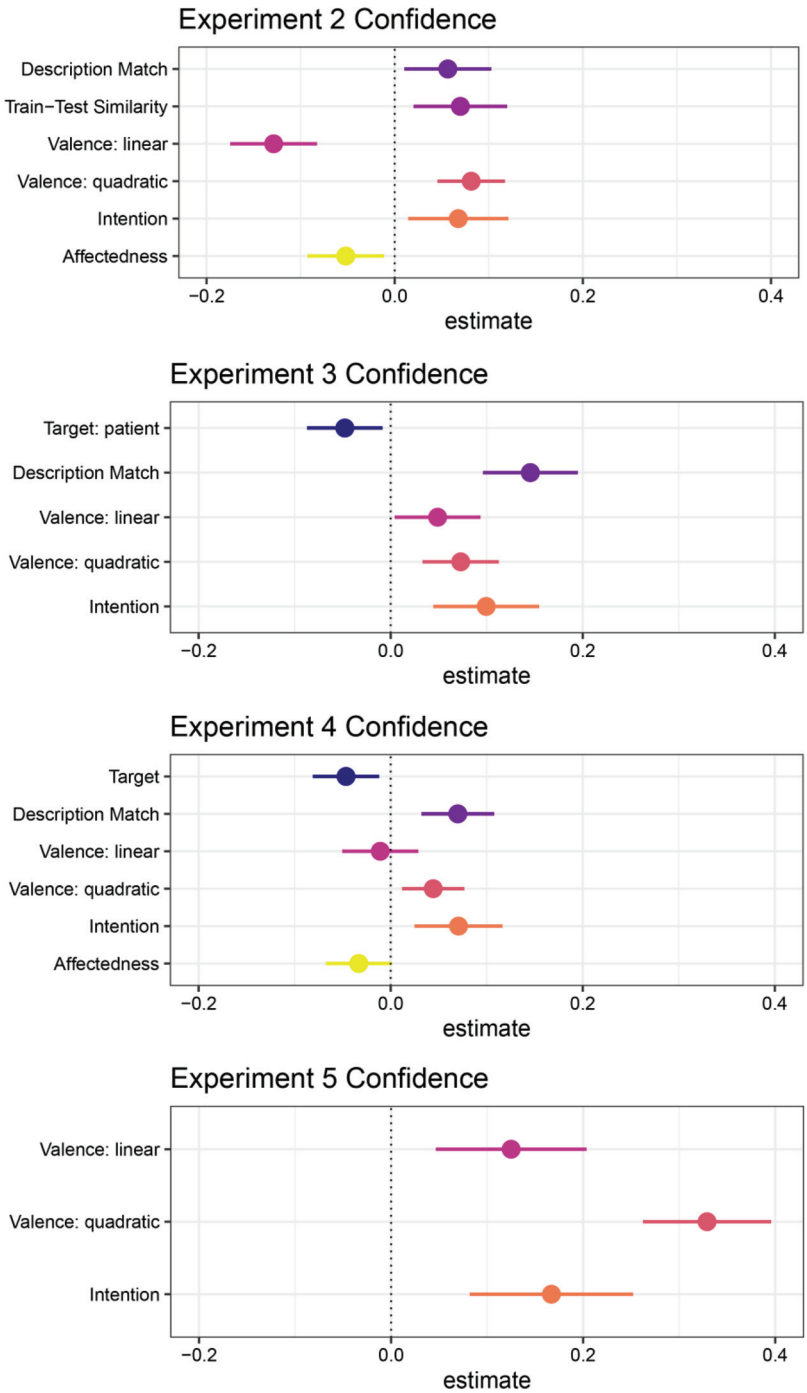
Figure 8

Coefficients for Best-Fitting Models of Test Accuracy



Note. Error bars denote 95% confidence intervals. For Train-test Similarity in Experiments 2 and 4, the metric is median distance. For Train-test Similarity in Experiment 3, the metric is minimum distance. p-values for the coefficients listed are all less than .05 with three exceptions: for Causation as a predictor in Experiment 3, $p = .07$; for Affectedness as a predictor in Experiment 4, $p = .07$; for Target as a predictor in Experiment 5, $p = .06$. See the online article for the color version of this figure.

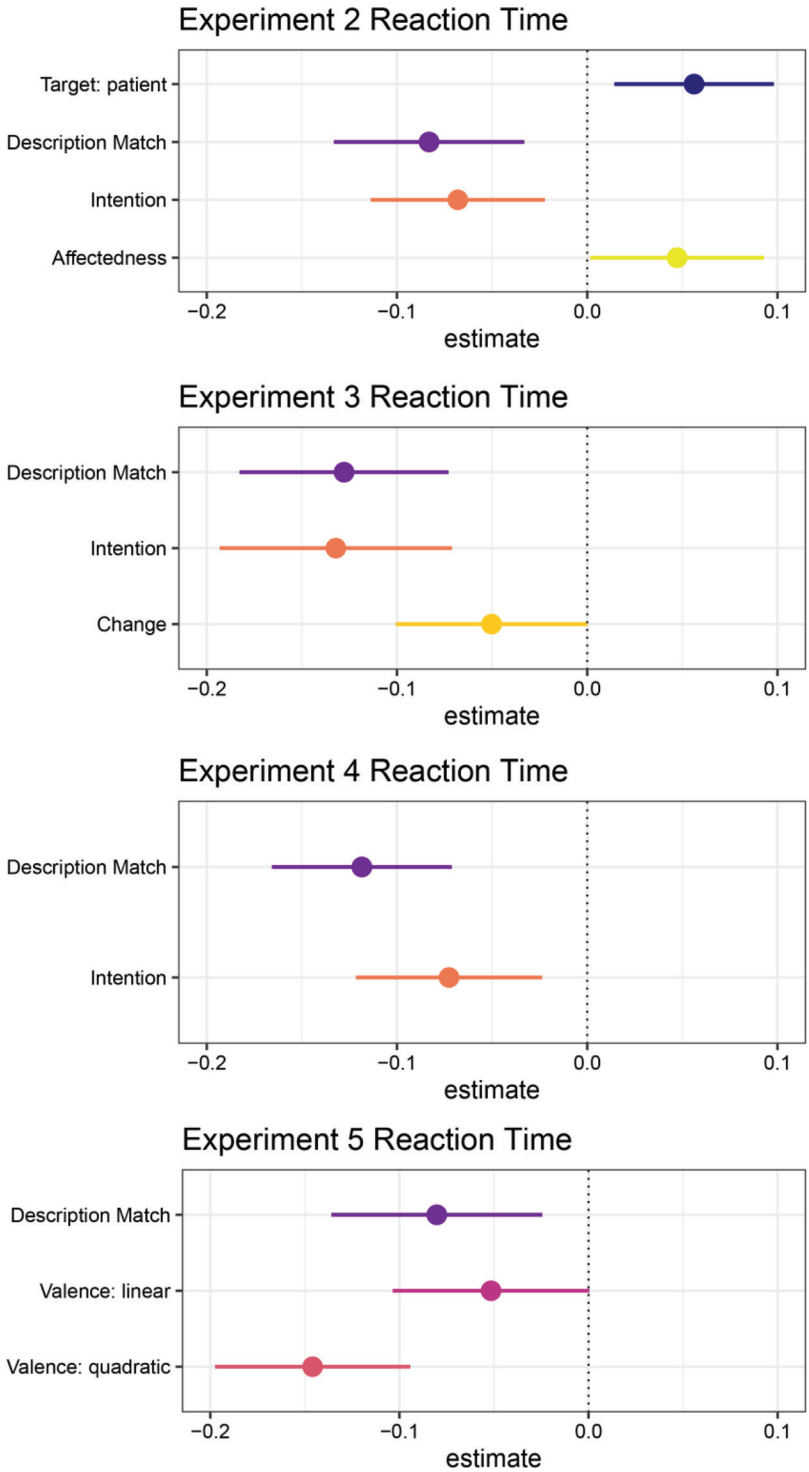
Figure 9
Coefficients for Best-Fitting Models of Test Confidence



Note. Error bars denote 95% confidence intervals. For Train-test Similarity as a predictor in Experiment 2, the metric is median distance. p values for the coefficients listed are all less than .05, with one exception: for Affectedness as a predictor in Experiment 4, $p = .06$. In Experiment 4, the linear term for Valence was not significant. See the online article for the color version of this figure.

Figure 10

Coefficients for Best-Fitting Models of Test Reaction Time



Note. Error bars denote 95% confidence intervals. p values for the coefficients listed are all less than .05, with two exceptions: for Change as a predictor in Experiment 3, $p = .06$; for the linear Valence term a predictor in Experiment 5, $p = .06$. See the online article for the color version of this figure.

reverse effect for Experiment 3. We found that some of the Proto-properties predicted test performance—the Intention dimension was a particularly reliable predictor.

In Experiments 2 and 3, we trained participants on prototypical instances of Agents and Patients (that is, scenes with high transitivity). Participants' ability to induce the Agent and Patient categories may have depended on seeing only prototypical exemplars of these categories during training. At the same time, it is possible that the Proto-properties would be less predictive of test accuracy if participants had been trained on a broader set of exemplars. In Experiment 4, we tested these possibilities by training the categories through scenes that varied both in terms of their transitivity and their valence. We predicted that participants would have more difficulty learning the categories when they were trained on scenes with less prototypical Agents and Patients than when the training only included prototypical category members, as in Experiment 3. We also predicted that participants who learned this broadest set of categories would show better test performance than participants in Experiment 3.

Experiment 4

Method

Participants

We recruited 152 adult native speakers of English on Amazon Mechanical Turk (N_{female} = 57, N_{male} = 93, age range = 20 – 70, median age = 35). An additional six participants were tested but were excluded for failing to pass at least 11 of 12 attention check trials. The participants from Experiment 4 had not taken part in Experiments 2 or 3. Participants received $1.50 for completing the study and received a $.50 bonus if they completed at least 70% of training trials accurately.

Design and Materials

The design was the same as in Experiments 2–3 except for the distribution of transitivity and valence across the training and test phases (see Figure 1). We selected 60 of the scenes from Experiment 2 and divided them into four quadrants: high transitivity [4.6–5.9]/negative [1–3.8], high transitivity [4.6–5.4]/positive [4.2–6.7], low transitivity [2.7–4.3]/negative [2.4–3.6], and low transitivity [2.0–4.4]/positive [4.0–6.9]. In the training phase, each participant viewed three scenes from each of these quadrants, selected at random (that is, 12 scenes viewed during training). The 48 scenes that were not viewed during training were viewed at test. We constructed multiple orders of the stimuli such that, across all participants, each of the 60 scenes was viewed during both training and test.

Procedure and Coding

The procedure and coding were the same as in Experiments 2 and 3.

Results

Training Phase: Did Participants Induce Agent/Patient Categories?

Figure 6 shows mean training accuracy by explanation type for Experiment 4. 58% of participants gave correct explanations. We fit a logistic regression model, predicting correct versus incorrect/uninformative explanations from Experiment 3 versus 4 (baseline = Experiment 3). The proportion of participants giving correct explanations did not differ across experiments ($b = -.32$, $SE = .25$, $p = .20$).

M accuracy on the last eight training trials of Experiment 4 was 70.6% (95% CI [66.6%, 74.6%]). We modeled training accuracy across Experiments 3–4 with both Experiment (baseline = Experiment 3) and Explanation type (baseline = uninformative) as predictors. As in Experiments 2 and 3, participants giving correct explanations were more accurate than participants giving uninformative explanations ($b = 2.9$, $SE = .38$, $p < .001$) and accuracy did not differ between people giving incorrect and uninformative explanations ($b = .03$, $SE = .39$, $p = .94$). We found no main effect of Experiment ($b = -.26$, $SE = .37$, $p = .49$) and only a marginal interaction between Experiment and Explanation type such that participants giving correct explanations had lower training accuracy in Experiment 4 than Experiment 3 ($b = -.75$, $SE = .45$, $p = .09$). A model where the baseline level of Explanation type was "correct" showed that training accuracy was lower for participants giving correct than incorrect explanations ($b = -2.90$, $SE = .33$, $p < .001$). These analyses indicate that seeing both prototypical and nonprototypical Agents and Patients during training did not hinder participants' ability to induce the categories.

Test Phase: How Did Participants Generalize the Categories?

Overall Test Performance. Test accuracy was 91.4% (95% CI [88.1%, 94.7%]) for participants giving correct explanations ($N = 88$), 55.9% (95% CI [48.5%, 63.3%]) for participants giving incorrect explanations ($N = 28$) and 52.0% (95% CI [48.9%, 55.1%]) for participants giving uninformative explanations ($N = 36$). To test whether test accuracy differed between Experiments 3 and 4, we fit a model with Experiment (baseline = Experiment 3) and Explanation type (baseline = incorrect/uninformative) as fixed effects, including only the 55 scenes common across the test phases in both Experiment 3 and 4. Although we predicted that a broader set of training stimuli would lead to better generalization, we found no difference in test accuracy between experiments ($b = -.21$, $SE = .27$, $p = .37$). As before, participants giving correct explanations had higher test accuracy than participants giving incorrect/uninformative explanations ($b = 3.06$, $SE = .23$, $p < .001$). There was no interaction between Experiment and Explanation type ($b = -.19$, $SE = .31$, $p = .54$).

Do the Proto-Properties Predict Test Performance? Turning to the predictions made by the Proto-Role theory, Figures 8–10 show coefficient estimates for the best-fitting models of accuracy, confidence rating, and RT. As in Experiments 2–3, we analyzed the test data for participants who were classified as learners (either producing a correct explanation or scoring 87.5% on training trials, or both). Mean test accuracy for the learners ($N = 93$) was 90.4% (95% CI [87.1%, 93.7%]). For the nonlearners, test accuracy did not differ from chance: $M = 52.0\%$ (95% CI [48.8%, 55.2%]).

As in Experiments 2 and 3, participants were more accurate, more confident, and faster when the scenes were less ambiguous (Description Match). Participants were also more accurate for test scenes that were more globally similar to the scenes they saw at training. We found evidence for the predictiveness of the Proto-properties, though weaker evidence than what was found in Experiments 2 and 3. Larger Intention scores predicted higher confidence and faster

responses in Experiment 4 but not higher accuracy. In addition, neither Causation nor Change predicted any of the dependent variables. At the same time, we found additional evidence for the predictiveness of Affectedness—it was associated with lower accuracy and lower confidence. As in Experiment 3, we found an effect of dot placement (Target) such that participants were more confident when the dot was on the Agent. We also found the effect of Valence such that confidence was higher at the ends of the Valence continuum than in the middle. The difference in Body Posture between the Agent and the Patient did not predict performance. In addition, differences in performance were not associated with differences in whether the Agent was moving more or was more stationary than the Patient.

Discussion

As in Experiments 2 and 3, we found in Experiment 4 that many participants induced the Agent/Patient categories, but many participants did not. The breadth of the categories that participants needed to learn did not appear to affect performance in the training phase—training accuracy did not differ between Experiment 4 and Experiment 3. In addition, the greater semantic coverage of the training scenes did not affect test accuracy, which was more than 90% for participants giving correct explanations in both Experiments 3 and 4. As in Experiments 2 and 3, we found evidence for some but not all of the Proto-properties: only Intention and Affectedness predicted participants' generalization.

A possible objection to the conclusion that inducing the Agent/Patient categories is difficult for some is that the task may somehow obscure participants' abilities. In Experiments 2–4, we tested workers on Amazon Mechanical Turk, a context in which workers are paid a flat rate and some workers may try to complete tasks as quickly as possible. In addition, although the participants we analyzed all passed attention checks, the task itself may be completed with a minimal level of attention. To test whether our method can elicit robust learning, we conducted a final experiment in which participants were asked to categorize the scenes not on the basis of Agent/Patient but on the basis of valence (negative versus positive). As all participants in Experiment 1 spontaneously sorted the pictures by valence, we expected that these categories would be easier for participants to learn the Agent/Patient categories.

Experiment 5

Method

Participants

We recruited 56 adult native speakers of English through Amazon Mechanical Turk (N_{female} = 28, N_{male} = 27, age range = 24–74, median age = 32). An additional six participants were tested but were excluded for failing to pass at least 11 of 12 attention check trials. The participants from Experiment 5 had not taken part in Experiments 2-4. Participants received $1.50 for completing the study and received a $.50 bonus if they completed at least 70% of training trials accurately.

Design and Materials

From the set of 64 scenes tested in Experiment 2, we selected 12 scenes that were highly positive (mean rating of 5.9 or higher) and 12 scenes that were highly negative (mean rating of 1.6 or lower). In the training phase, participants saw six positive scenes and six negative scenes selected at random from each group of 12 scenes. As in Experiments 2–4, participants in the training phase saw both the Agent-dot and Patient-dot variants of each scene, for a total of 24 training trials. The six positive and six negative scenes not viewed during training were viewed at test. The test phase also included the 40 scenes with intermediate valence ratings. As in Experiments 2–4, participants saw only one dot-variant (Agent-dot vs. Patient-dot) for each of the 52 test scenes. Dot-variant and Agent side were counterbalanced in both the training and test phases. Whether the 'A' key corresponded to the positive or negative category was also counterbalanced.

Procedure and Coding

The procedure was the same as in Experiments 2–4 except the instructions omitted any mention of the red dots. The explanations given after the training phase were coded as to whether they conveyed the difference in valence between the categories. Examples of explanations coded as "correct" include: "A is being nice, b is being mean," "Category A are good deeds, category B are bad deeds," and "A is hurting someone B is helping someone."

Results

Training Phase: Did Participants Induce Valence Categories?

Figure 6 shows mean training accuracy by explanation type for Experiment 5. Correct explanations were provided by 86% of participants. *M* accuracy on the last eight training trials was 93.0% (95% CI [89.5%, 96.5%]). A logistic regression showed that the proportion of participants giving correct explanations was higher in Experiment 5 compared with Experiment 2: $b = -1.45$, $SE = .41$, $p < .001$; Experiment 3: $b = -1.15$, $SE = .43$, $p < .01$ and Experiment 4: $b = -1.47$, $SE = .42$, $p < .001$. To assess whether learning was more successful for valence-based categories than for Agent/Patient categories, we fit a mixed-effects logistic regression with four levels of the fixed effect experiment (baseline = Experiment 5). Training accuracy was higher in Experiment 5 than in Experiments 2–4: Experiment 2: $b = -1.90$, $SE = .35$, $p < .001$; Experiment 3: $b = -1.52$, $SE = .37$, $p < .001$; Experiment 4: $b = -2.30$, $SE = .36$, $p < .001$.

Test Phase: How Did Participants Generalize the Categories?

We modeled accuracy, confidence, and RT as in Experiments 2–4, including only participants who gave correct explanations or answered at least 87.5% of the last eight training trials correctly (N = 50). For this set of learners, test accuracy was 86.4% (95% CI [84.2%, 88.6%]). For nonlearners, test accuracy was 52.0% (95% CI [46.7%, 57.3%]). We did not expect that the Proto-properties would predict how participants generalized the valence categories. Figures 8–10 shows the coefficient estimates for the best-fitting model of each dependent variable.

As expected, test performance was largely determined by the quadratic Valence term: participants were more accurate, more confident, and faster for scenes at the ends of the Valence

continuum than for scenes closer to the middle of the valence continuum. Controlling for distance from center (that is, the quadratic predictor), performance on all three measures was better for more positive scenes. Reaction times were faster when the scenes were less ambiguous (Description Match). Two unexpected results were that higher Intention scores predicted higher confidence, and accuracy was worse when the dot was on the Patient.

Discussion

Participants readily learned valence-based categories even though learning these categories required participants to learn to ignore the red dots on the images: participants showed significantly higher training performance than participants learning Agent/Patient categories. This ameliorates the concern that the stimuli were somehow difficult to process or distinguish from one another or that the training procedure was somehow confusing. As expected, the Valence norm was the best predictor of people's generalization, with performance being best for the most positive and negative scenes and lower for scenes closer to the positive/negative boundary, which is understandably subjective.

General Discussion

In this study, we asked whether English speakers explicitly categorize visual events in terms of Agent and Patient roles. We also asked whether the Proto-Role properties relevant to English syntax are relevant to categories extracted from visual events. Across four experiments, we found that many English speakers explicitly encode Agent/Patient categories but many do not. We also found that some but not all of the Proto-properties predicted how participants generalize the Agent/Patient categories they induced. These results suggest that despite the prominence of event roles in event perception and language production and comprehension, many people have trouble using these categories in explicit categorization tasks. Our results also suggest that Agent and Patient are domain-general categories, influencing syntactic organization and visual event apprehension in comparable ways.

Is Core Knowledge Accessible to Awareness?

As described at the outset, core knowledge of event roles has been argued to influence a diverse range of behavior: how children learn language, how meanings are expressed morphosyntactically across the languages of the world, and how adults perceive events. Another possible influence of core knowledge on behavior is that adults robustly access core concepts for the purpose of explicit categorization. As far as core knowledge of Agent/Patient roles is concerned, our results cast doubt on the robustness of this influence. In Experiments 2–4, between 29% and 39% of participants failed to induce the Agent/Patient categories compared with just 11% of participants in Experiment 5, who had to ignore the Agent/Patient distinction while learning to attend to the valence of the scenes. In addition, the in-lab sorting task in Experiment 1 showed that 26% of participants were unable to recognize the Agent/Patient categories even when this way of sorting the cards was presented to them. This difficulty is surprising given that our participants are adult proficient English speakers who rely on Agent and Patient categories every time they produce or comprehend transitive sentences such as *Murray ate the ice cream*. Together, the sorting and supervised categorization tasks suggest that abstracting the Agent/Patient distinction from nonverbal scenes is surprisingly nontrivial, at least in comparison to valence, which nearly all participants effortlessly extracted. Despite evidence that abstract roles are extracted rapidly and automatically when we process visual events (Hafri et al., 2013, 2018), this automatic extraction does not appear to guarantee conscious awareness in explicit categorization tasks.

A few participants may have learned the categories implicitly—10 participants across Experiments 2–5 had 100% accuracy on the last eight training trials but did not provide correct explanations. These participants' descriptions ranged from wholly uninformative ("it was easy") to approaching correct ("a not doing something b doing somthing [sic]"). Although some of these participants may have learned the Agent/Patient categories implicitly, others may not have communicated everything they learned.

One interpretation of our results is that Agent/Patient roles are less conceptually prominent than previously thought. This interpretation, however, is difficult to square with the wealth of literature reviewed earlier that even infants use these categories to interpret events and learn language. This interpretation also fits uneasily with our finding that participants' induction of the Agent/Patient categories was largely resilient to changes in which scenes were used in training. Participants did not have greater difficulty learning the categories when presented with both negative and positive scenes during training than when presented only with negative scenes (Experiment 2 versus 3). In addition, participants were able to induce the categories when the training scenes featured both prototypical and nonprototypical Agents and Patients (that is, both high-transitivity and low-transitivity scenes; Experiment 4). Finally, for 24% of participants in Experiment 1, sorting the cards in terms of Agent/Patient categories was their first sorting strategy.[7]

Our results seemingly present a paradox, with some participants failing to induce the Agent/Patient categories, but others representing them as highly prominent. One possible explanation is that Agent/Patient categories are automatically hypothesized as relevant but other dimensions of the scenes are even more prominent. Indeed, valence appears to be a more prominent dimension than event roles: almost all participants spontaneously sorted the scenes by valence in Experiment 1, and 89% of participants qualified as "learners" in the valence task in Experiment 5. Social cognition has been argued to be one of the systems of core cognition (Spelke & Kinzler, 2007) and valence—recognizing whether an interaction between two people is positive or negative—is an essential part of social cognition. Valence has a biological basis that is manifest, for example, through animals' sensitivity to distress calls (Seyfarth et al., 1980) and awareness of emotional valence has been observed in both pigs (Luna et al., 2021) and horses (Briefer et al., 2017). It may well be that some systems within core cognition are more conceptually prominent than others and that valence is more prominent than event roles. We do not believe, however, that our

[7] We do not know whether they initially considered sorting by valence and rejected it in favor of Agent/Patient, or whether valence occurred to them as a sorting criterion only after Agent/Patient.

results are fully explained by a view in which valence and event roles are both automatically hypothesized but valence is more prominent. Participants in Experiments 2–4 were given ample evidence that valence was not the correct dimension for categorization—for some participants, hypothesizing the relevance of event roles appeared to present a particular difficulty that is unexpected if these roles automatically populate learners' hypothesis spaces.

Why were some participants so much better at spontaneously inducing event roles from visual events? A full investigation of the factors responsible for the large individual differences we observed here requires further study, but we speculate on a few. In Experiment 1, we found that participants had different sorting preferences, reminiscent of the finding that adults vary as to whether they categorize nouns according to thematic similarity (for example, *bees—honey*) or taxonomic similarity (for example, *bees – mosquitos*; Lin & Murphy, 2001; Mirman & Graziano, 2012). Lin and Murphy suggest that such differences may be driven by variation in participants' experience and thinking style. Future research can explore the extent to which individuals' differing approaches to categorization align across different semantic domains. Differences in categorization strategies may be even wider for individuals living in different cultural environments.

Another possibility is that given tasks such as ours, some participants default to a more abstract reasoning style. For example, presented with sets of either all identical or nonidentical images, some participants readily induce the abstract distinction same versus different. Others do not, instead focusing on more specific properties of the images making up the sets (Castro & Wasserman, 2013; Wasserman & Castro, 2012). Further insight into the factors responsible for differences in explicitly inducing event roles can also be gained by investigating whether participants who succeed at learning the Agent/Patient categories are also more successful at tasks involving other abstract relational reasoning (see Gentner & Asmuth, 2019; Jamrozik & Gentner, 2020). It is worth noting that success in extracting the categories was not explained by participants' level of education (see footnote 3).

Individual differences may also be the result of differences in the likelihood that participants rely on a naming strategy. In past work, we have shown that nameability is a powerful predictor of category learning. For example, categories based on more nameable colors and shapes were learned more quickly and accurately than categories with the same logical structure, but with less nameable parts (Zettersten & Lupyan, 2020). English lacks conventional, widely-known labels for Agent/Patient. In fact, across the 33 Agent/Patient explanations elicited in Experiment 1, participants used 17 different predicates to characterize the categories (for example, *do*, *perform*, *receive*, *give*, *harm*, *help*, and *control*). It is possible that participants who were more likely to succeed were those that had more consistent Agent/Patient labels available at their disposal. In contrast to event roles, valence is highly nameable and has been argued to be universally encoded in the emotional lexicon of different languages (Jackson et al., 2019). The relative ease of naming the scenes as good versus bad may help explain both the much higher categorization accuracy in Experiment 5, and the greater ability to explain what the categorical distinction entails.

Evaluating the Proto-Role Theory for Visual Events

Our second question was whether Agent/Patient categories are structured in similar ways across linguistic and conceptual representation. As described earlier, there is strong evidence, particularly in the area of event cognition, that semantic and conceptual domains are organized in parallel. In linguistic theory, Agent/Patient roles are often analyzed as having prototype structure. In Experiments 2–4, we found evidence that the Proto-properties explaining English argument realization also explain generalization of Agent/Patient categories constructed from nonlinguistic stimuli.

Intention was the most reliable Proto-property, predicting performance in eight of nine analyses.[8] This is consistent with the judgment studies in Kako (2006) and corpus ratings in Reisinger et al. (2015): intentionality was strongly and consistently associated with the Subject argument of a sentence. The predictiveness of intentionality is also consistent with the finding that adults are biased to interpret the Subject of an ambiguous sentence (e.g., *Dionne bumped into Murray*) as intentional (Strickland et al., 2014). Affectedness was the second most reliable Proto-property, emerging as a significant predictor in four out of nine analyses (marginally significant for Experiment 4).[9] Causation predicted accuracy in Experiment 2 and marginally predicted accuracy in Experiment 3—Reisinger et al. also found that causation was strongly associated with the Subject argument. We found no evidence that Movement or being Stationary predicted performance. Strikingly, this is also what Reisinger et al. found in their corpus analysis—that these properties were not significantly associated with either the Subject or the Object. The set of significant effects that we observed across Experiments 2–4 emerged even when we controlled for ambiguity in the stimuli (Description Match), global similarity between each test scene and the particular training scenes that a participant saw (Train-test Similarity), and Valence. Given that these three factors often correlated with the Proto-properties themselves (see Figure 5), our analysis constitutes a strong test of the predictiveness of the Proto-properties. One point of divergence between our results and previous linguistic studies was for the Proto-Patient predictor Change, which was significant in only a single analysis. We found that the Agent–Patient difference score for Change (i.e., how much the dot-figure changed as a result of the interaction) marginally predicted RT in Experiment 3, and in the opposite direction from what we predicted: reaction times were faster when the Agent was more changed as a result of the interaction than the Patient.

Our results support the proposal that Agent and Patient have prototype structure and that this structure is similar in linguistic and nonlinguistic tasks. Our results are consistent with the nativist view sketched by Rissman & Majid (2019) that Agent and Patient are universal categories. Cross-linguistic work is needed to further test this view, as the similarity in category structure that we observed across linguistic and nonlinguistic tasks may be specific to English. We note that the categories induced across Experiments 2–4 do not appear to be identical: participants in Experiment 2 (trained only on negative scenes) reported less confidence for more positive scenes at test, a pattern which was reversed in Experiment 3 (where participants were trained on both positive and negative scenes). This finding suggests that although participants may be tapping into universal Agent/Patient knowledge

[8] Nine ~ three experiments * three dependent variables per experiment.

[9] Kako (2006) and Reisinger et al. (2015) did not test Affectedness.

when completing the generalization trials, this knowledge is modulated by particular properties of the training items.

One somewhat unexpected finding was that participants did better with scenes that had the dot on the Agent than scenes that had the dot on the Patient. Given the symmetry of the design, this is surprising: if someone learned that the distinction was about Agents and Patients, if the dot is on one, then it is not on the other. However, this kind of *psychological* asymmetry is consistent with results suggesting that Agent is a more stable, coherent category than Patient (Cohn & Paczynski, 2013; Hafri et al., 2013; White et al., 2017). If participants begin by formulating an explicit hypothesis about what is special about the dotted figure, it may be that it is easier to formulate (and/or test) this hypothesis about the more stable coherent category of Agent compared with a more diffuse category of Patient. We never observed Body Posture predicting generalization—this result presents a contrast with Hafri et al. (2013), where postures such as outstretched arms predicted participants' ability to detect the Agent. Postures may be relevant to prototypicality when an image is presented for only a brief amount of time, as in the Hafri task, but not relevant when participants have ample time to study the images.

Conclusion

Our studies yielded a surprising finding: that prominent Agent/Patient concepts, which may be innate, are nonetheless difficult for some participants to access during explicit category induction tasks. One implication of this disconnect is that theories of core knowledge may be more limited than previously thought regarding the scope of behavioral phenomena that they can make predictions about. For those participants who did induce the Agent/Patient categories, their patterns of generalization provide strong evidence that Agent/Patient categories are domain-general, with similar prototype structure across linguistic and nonlinguistic domains.

References

Ackerman, F., & Moore, J. (2001). Proto-properties and grammatical encoding. *Stanford Monographs in Linguistics*. CSLI.

Arunachalam, S., & Waxman, S. R. (2010). Meaning from syntax: Evidence from 2-year-olds. *Cognition*, *114*(3), 442–446. https://doi.org/10.1016/j.cognition.2009.10.015

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed-effects models using S4 classes (R package version 1.1–7). http://CRAN.R-project.org/package=lme4.

Bornkessel, I., Schlesewsky, M., Comrie, B., & Friederici, A. (Eds.). (2006). *Semantic role universals and argument linking: Theoretical, typological and psycholinguistic perspectives*. Mouton de Gruyter. https://doi.org/10.1515/9783110219272

Bowerman, M., & Brown, P. (2008). *Crosslinguistic perspectives on argument structure: Implications for learnability*. Erlbaum.

Briefer, E. F., Mandel, R., Maigrot, A.-L., Briefer Freymond, S., Bachmann, I., & Hillmann, E. (2017). Perception of emotional valence in horse whinnies. *Frontiers in Zoology*, *14*(1), 8. https://doi.org/10.1186/s12983-017-0193-1

Carey, S. (2009). *The origin of concepts*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195367638.001.0001

Castro, L., & Wasserman, E. A. (2013). Humans deploy diverse strategies in learning same-different discrimination tasks. *Behavioural Processes*, *93*, 125–139. https://doi.org/10.1016/j.beproc.2012.09.015

Cohn, N., & Paczynski, M. (2013). Prediction, events, and the advantage of agents: The processing of semantic roles in visual narrative. *Cognitive Psychology*, *67*(3), 73–97. https://doi.org/10.1016/j.cogpsych.2013.07.002

Croft, W. (2012). *Verbs: Aspect and causal structure*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199248582.001.0001

Csibra, G., & Gergely, G. (2007). 'Obsessed with goals': Functions and mechanisms of teleological interpretation of actions in humans. *Acta Psychologica*, *124*(1), 60–78. https://doi.org/10.1016/j.actpsy.2006.09.007

DeLancey, S. (1991). Event construal and case role assignment. In L. Sutton, C. Johnson, & R. Shields (Eds.), *Proceedings of the 17th annual meeting of the Berkeley linguistics society* (pp. 338–353). Berkeley Linguistics Society.

Dowty, D. (1991). Thematic proto-roles and argument selection. *Language*, *67*(3), 547–619. https://doi.org/10.1353/lan.1991.0021

Ergin, R., Meir, I., Aran, D. I., Padden, C., & Jackendoff, R. (2018). The development of argument structure in central taurus sign language. *Sign Language Studies*, *18*(4), 612–639. https://doi.org/10.1353/sls.2018.0018

Fausey, C. M., & Boroditsky, L. (2011). Who dunnit? Cross-linguistic differences in eye-witness memory. *Psychonomic Bulletin & Review*, *18*(1), 150–157. https://doi.org/10.3758/s13423-010-0021-5

Fausey, C. M., Long, B. L., Inamori, A., & Boroditsky, L. (2010). Constructing agency: The role of language. *Frontiers in Psychology*, *1*, 162. https://doi.org/10.3389/fpsyg.2010.00162

Ferrigno, S., Jara-Ettinger, J., Piantadosi, S. T., & Cantlon, J. F. (2017). Universal and uniquely human factors in spontaneous number perception. *Nature Communications*, *8*(1), 13968. https://doi.org/10.1038/ncomms13968

Fillmore, C. J. (1968). The case for case. In E. W. Bach & R. T. Harms (Eds.), *Universals in linguistic theory* (pp. 1–88). Holt, Rinehart and Winston.

Fisher, C., Hall, D. G., Rakowitz, S., & Gleitman, L. (1994). When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *Lingua*, *92*(0), 333–375. https://doi.org/10.1016/0024-3841(94)90346-8

Flaherty, M. E. (2014). *The emergence of argument structural devices in Nicaraguan Sign Language*. The University of Chicago.

Gentner, D., & Asmuth, J. (2019). Metaphoric extension, relational categories, and abstraction. *Language, Cognition and Neuroscience*, *34*(10), 1298–1307. https://doi.org/10.1080/23273798.2017.1410560

Goldin-Meadow, S., & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature*, *391*(6664), 279–281. https://doi.org/10.1038/34646

Goldwater, M. B., Markman, A. B., & Stilwell, C. H. (2011). The empirical case for role-governed categories. *Cognition*, *118*(3), 359–376. https://doi.org/10.1016/j.cognition.2010.10.009

Grimm, S. (2011). Semantics of case. *Morphology*, *21*(3-4), 515–544. https://doi.org/10.1007/s11525-010-9176-z

Grimshaw, J. B. (1990). *Argument structure*. MIT Press.

Gruber, J. (1965). *Studies in lexical relations* [Dissertation/Thesis]. Massachusetts Institute of Technology, Department of Modern Languages.

Hafri, A., Papafragou, A., & Trueswell, J. C. (2013). Getting the gist of events: Recognition of two-participant actions from brief displays. *Journal of Experimental Psychology: General*, *142*(3), 880–905. https://doi.org/10.1037/a0030045

Hafri, A., Trueswell, J. C., & Strickland, B. (2018). Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition*, *175*, 36–52. https://doi.org/10.1016/j.cognition.2018.02.011

Hopper, P. J., & Thompson, S. A. (1980). Transitivity in grammar and discourse. *Language*, *56*(2), 251–299. https://doi.org/10.1353/lan.1980.0017

Jackendoff, R. (1983). *Semantics and cognition* (Vol. 8). MIT press.

Jackendoff, R. (1990). *Semantic structures*. MIT Press.

Jackson, J. C., Watts, J., Henry, T. R., List, J.-M., Forkel, R., Mucha, P. J., Greenhill, S. J., Gray, R. D., & Lindquist, K. A. (2019). Emotion

semantics show both cultural variation and universal structure. *Science*, *366*(6472), 1517–1522. https://doi.org/10.1126/science.aaw8160

Jamrozik, A., & Gentner, D. (2020). Relational labeling unlocks inert knowledge. *Cognition*, *196*, 104146. https://doi.org/10.1016/j.cognition.2019.104146

Kabata, K. (2013). Goal–source asymmetry and crosslinguistic grammaticalization patterns: A cognitive-typological approach. *Language Sciences*, *36*, 78–89. https://doi.org/10.1016/j.langsci.2012.03.021

Kako, E. (2006). Thematic role properties of subjects and objects. *Cognition*, *101*(1), 1–42. https://doi.org/10.1016/j.cognition.2005.08.002

Kelso, J. A. S. (2016). On the self-organizing origins of agency. *Trends in Cognitive Sciences*, *20*(7), 490–499. https://doi.org/10.1016/j.tics.2016.04.004

Kuchinsky, S. E. (2009). *From seeing to saying: Perceiving, planning, producing*. University of Illinois at Urbana-Champaign.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Lakusta, L., & Carey, S. (2015). Twelve-month-old infants' encoding of goal and source paths in agentive and non-agentive motion events. *Language Learning and Development*, *11*(2), 152–157. https://doi.org/10.1080/15475441.2014.896168

Lakusta, L., & Landau, B. (2005). Starting at the end: The importance of goals in spatial language. *Cognition*, *96*(1), 1–33. https://doi.org/10.1016/j.cognition.2004.03.009

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517–544. https://doi.org/10.1111/j.1551-6709.2011.01220.x

Lakusta, L., Spinelli, D., & Garcia, K. (2017). The relationship between pre-verbal event representations and semantic structures: The case of goal and source paths. *Cognition*, *164*, 174–187. https://doi.org/10.1016/j.cognition.2017.04.003

Lakusta, L., Wagner, L., O'Hearn, K., & Landau, B. (2007). Conceptual foundations of spatial language: Evidence for a goal bias in infants. *Language Learning and Development*, *3*(3), 179–197. https://doi.org/10.1080/15475440701360168

Levin, B., & Rappaport-Hovav, M. (2005). *Argument realization*. Cambridge University Press. https://doi.org/10.1017/CBO9780511610479

Lidz, J., Gleitman, H., & Gleitman, L. (2003). Understanding how input matters: Verb learning and the footprint of universal grammar. *Cognition*, *87*(3), 151–178. https://doi.org/10.1016/S0010-0277(02)00230-5

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General*, *130*(1), 3–28. https://doi.org/10.1037/0096-3445.130.1.3

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

Luna, D., González, C., Byrd, C. J., Palomo, R., Huenul, E., & Figueroa, J. (2021). Do domestic pigs acquire a positive perception of humans through observational social learning? *Animals*, *11*(1), 127. https://doi.org/10.3390/ani11010127

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

Luraghi, S. (1995). Prototypicality and agenthood in Indo-European. In H. Andersen (Ed.), *Historical linguistics 1993* (pp. 259–259). John Benjamins. https://doi.org/10.1075/cilt.124.21lur

Malt, B. C., & Majid, A. (2013). How thought is mapped into words. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 583–597. https://doi.org/10.1002/wcs.1251

Markman, A. B., & Stilwell, C. H. (2001). Role-governed categories. *Journal of Experimental & Theoretical Artificial Intelligence*, *13*(4), 329–358. https://doi.org/10.1080/09528130110100252

Mirman, D., & Graziano, K. M. (2012). Individual differences in the strength of taxonomic versus thematic relations. *Journal of Experimental Psychology: General*, *141*(4), 601–609. https://doi.org/10.1037/a0026451

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(2), 357–374. https://doi.org/10.1017/S0305000900013817

Narasimhan, B., Kopecka, A., Bowerman, M., Gullberg, M., & Majid, A. (2012). Putting and taking events: A crosslinguistic perspective. In A. Kopecka & B. Narasimhan (Eds.), *Events of 'Putting' and 'Taking': A crosslinguistic perspective* (pp. 1–20). John Benjamins. https://doi.org/10.1075/tsl.100.03nar

Noble, C. H., Rowland, C. F., & Pine, J. M. (2011). Comprehension of argument structure and semantic roles: Evidence from English-learning children and the forced-choice pointing paradigm. *Cognitive Science*, *35*(5), 963–982. https://doi.org/10.1111/j.1551-6709.2011.01175.x

Padden, C., Meir, I., Sandler, W., & Aronoff, M. (2009). Against all expectations: Encoding subjects and objects in a new language. In D. Gerdts, J. Moore, & M. Polinsky (Eds.), *Hypothesis A/Hypothesis B: Linguistic explorations in honor of David M. Perlmutter* (pp. 383–400). MIT Press.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106. https://doi.org/10.1162/0891201053630264

Papafragou, A. (2010). Source-goal asymmetries in motion representation: Implications for language production and comprehension. *Cognitive Science*, *34*(6), 1064–1092. https://doi.org/10.1111/j.1551-6709.2010.01107.x

Primus, B. (1999). *Cases and thematic roles: Ergative, accusative and active*. Max Niemeyer Verlag. https://doi.org/10.1515/9783110912463

R Core Team. (2017). R: A language and environment for statistical computing. https://www.R-project.org/.

Regier, T., & Zheng, M. (2007). Attention to endpoints: A cross-linguistic constraint on spatial meaning. *Cognitive Science*, *31*(4), 705–719. https://doi.org/10.1080/15326900701399954

Reisinger, D., Rudinger, R., Ferraro, F., Harman, C., Rawlins, K., & Van Durme, B. (2015). Semantic Proto-Roles. *Transactions of the Association for Computational Linguistics*, *3*, 475–488. https://doi.org/10.1162/tacl_a_00152

Rissman, L., & Majid, A. (2019). Thematic roles: Core knowledge or linguistic construct? *Psychonomic Bulletin & Review*, *26*(6), 1850–1869. https://doi.org/10.3758/s13423-019-01634-5

Rissman, L., & Lupyan, G. (2020). Agent-patient category structure. https://osf.io/a5cev/?view_only=446b016f85a14aaeb305514e1f1ecc28

Sandler, W., Meir, I., Padden, C., & Aronoff, M. (2005). The emergence of grammar: Systematic structure in a new language. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(7), 2661–2665. https://doi.org/10.1073/pnas.0405448102

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: Lexical and structural priming of syntactic constructions in young children. *Developmental Science*, *6*(5), 557–567. https://doi.org/10.1111/1467-7687.00312

Schlesinger, I. M. (1989). Instruments as agents: On the nature of semantic relations. *Journal of Linguistics*, *25*(1), 189–210. https://doi.org/10.1017/S0022226700012147

Seyfarth, R. M., Cheney, D. L., & Marler, P. (1980). Monkey responses to three different alarm calls: Evidence of predator classification and semantic communication. *Science*, *210*(4471), 801–803. https://doi.org/10.1126/science.7433999

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Strickland, B. (2017). Language reflects "core" cognition: A New theory about the origin of cross-linguistic regularities. *Cognitive Science*, *41*(1), 70–101. https://doi.org/10.1111/cogs.12332

Strickland, B., Fisher, M., Keil, F., & Knobe, J. (2014). Syntax and intentionality: An automatic link between language and theory-of-mind. *Cognition*, *133*(1), 249–261. https://doi.org/10.1016/j.cognition.2014.05.021

Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, *137*(0), 47–62. https://doi.org/10.1016/j.cognition.2014.12.007

Ünal, E., Ji, Y., & Papafragou, A. (2021). From event representation to linguistic meaning. *Topics in Cognitive Science*, *13*(1), 224–242. https://doi.org/10.1111/tops.12475

Ünal, E., Richards, C., Trueswell, J. C., & Papafragou, A. (2021). Representing agents, patients, goals and instruments in causative events: A cross-linguistic investigation of early language and cognition. *Developmental Science*. Advance online publication. https://doi.org/10.1111/desc.13116

Wasserman, E. A., & Castro, L. (2012). Categorical discrimination in humans and animals: All different and yet the same? *Psychology of Learning and Motivation*, *56*, 145–184. https://doi.org/10.1016/B978-0-12-394393-4.00005-4

White, A. S., Rawlins, K., & Van Durme, B. (2017). The Semantic Proto-Role linking model. *Proceedings of the 15th Conference of the European chap. of the Association for Computational Linguistics*, *2*, 92–98.

Wolff, P., Jeon, G., & Li, Y. (2009). Causers in English, Korean, and Chinese and the individuation of events. *Language and Cognition*, *1*(2), 167–196. https://doi.org/10.1515/LANGCOG.2009.009

Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science*, *26*(8), 1316–1324. https://doi.org/10.1177/0956797615588753

Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition*, *196*, 104135. https://doi.org/10.1016/j.cognition.2019.104135

Zheng, M., & Goldin-Meadow, S. (2002). Thought before language: How deaf and hearing children express motion events across cultures. *Cognition*, *85*(2), 145–175. https://doi.org/10.1016/S0010-0277(02)00105-1

Appendix A

Transitivity, Valence, and Description Match Values for Each Scene

Scene	Event	Transitivity	Valence	Description Match
27	One figure lifting another up through a crack in the floor	5.38	6.07	1.00
28	One figure kicking another	5.88	1.43	1.00
29	One figure punching another	5.80	1.36	1.00
30	One figure punching another	5.85	1.57	1.00
31	One figure kicking another	5.88	1.50	1.00
35	One figure holding another above its head	5.73	2.21	1.00
45	One figure performing CPR on another	5.30	5.57	1.00
47	One figure tripping another	5.73	1.64	1.00
58	One figure spanking a smaller figure	5.68	1.93	1.00
69	One figure strangling another	5.70	1.07	1.00
70	One figure painting another's chest blue	5.68	3.79	1.00
72	One figure shooting another with a gun	5.85	1.00	1.00
73	One figure chasing another with a hammer	5.10	1.50	1.00
94	One figure kicking another off a ledge	5.80	1.14	1.00
96	One figure pushing another	5.85	2.00	1.00
97	One figure tripping another	5.88	2.00	1.00
100	One figure jabbing another in the head	5.78	1.71	0.93
101	One figure dragging another	5.38	3.29	1.00
102	One figure shouting at another through a megaphone	5.50	2.93	1.00
200	One figure spanking a smaller figure with a stick	5.83	1.50	1.00
201	One figure kicking another	5.83	1.29	1.00
202	One figure massaging another	5.25	5.57	1.00
204	One figure pushing a smaller figure hanging off a grocery cart	5.35	4.79	1.00
205	One figure commanding another to stop moving	5.13	2.86	1.00
206	One figure yelling at another	3.15	2.71	0.87
208	One figure carrying a limp figure	4.85	4.21	0.87
209	One figure carrying another in their arms	4.38	6.50	1.00
210	One figure carrying another on their shoulders	4.10	6.43	0.87
211	One figure chasing another	3.23	4.00	0.87
212	One figure comforting another	2.75	5.36	0.87
213	One figure consoling another	3.18	5.50	1.00
214	One figure holding a small figure in their arms	3.73	6.71	1.00
215	One figure with a stethoscope listening to the lungs of another	4.13	5.71	1.00
216	One figure fighting another	4.83	2.43	0.73

(*Appendices continue*)

Appendix A (*Continued*)

Scene	Event	Transitivity	Valence	Description Match
217	One figure giving another a wrapped present	4.73	6.71	0.93
218	One figure giving another a key	3.38	5.50	0.93
219	One figure performing the Heimlich maneuver on another	4.98	5.79	1.00
220	One figure helping another who is falling down	4.98	5.00	0.87
221	One figure helping another up from a manhole	4.70	5.93	1.00
222	One figure helping another climb a wall	4.93	5.86	0.87
224	One figure leapfrogging over another	4.75	5.14	0.93
225	One figure lifting a smaller figure up in the air	4.58	6.14	1.00
226	One figure throwing a baseball to another figure with a bat	4.13	5.64	0.53
227	One figure blocking another figure who has a basketball	4.00	4.43	0.53
228	One figure swinging a smaller figure in the air	5.03	5.93	1.00
229	One figure poking another	2.98	4.50	0.93
230	One figure offering a heart to another	4.20	6.86	0.93
232	One figure pushing another on a stretcher	4.65	4.93	1.00
233	One figure pushing another in a wheelchair	4.85	6.07	1.00
234	One figure pushing another	4.90	3.57	1.00
235	One figure reading to another	3.38	6.79	1.00
236	One figure chasing another	4.30	2.64	0.79
237	One figure frightening another	4.23	2.36	0.80
238	One figure scolding another	4.60	2.43	0.93
239	One figure questioning another	3.70	3.07	0.93
240	One figure standing on another	5.08	1.57	1.00
241	One figure commanding another to stop	2.73	2.43	0.69
242	One figure strangling another	4.23	3.14	0.93
243	One figure taking a photo of another	3.63	4.86	0.93
244	One figure showing a laptop to another	3.05	5.43	0.87
245	One figure talking to another	3.40	3.64	0.93
246	One figure talking to another	2.03	6.07	0.73
247	One figure whispering to another	3.13	4.79	0.93
248	One figure yelling at another	4.15	2.71	1.00

Appendix B

Lists of Which Scenes Were Shown in Each Phase of Each Experiment

		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
Scene	Experiment 1	Training	Test	Training	Test	Training	Test	Training	Test
27	1	0	1	1	1	1	1	1	1
28	1	1	1	0	0	1	1	1	1
29	1	1	1	1	1	1	1	1	1
30	1	1	1	0	0	1	1	1	1
31	1	1	1	0	0	1	1	1	1
35	1	1	1	1	1	1	1	0	1
45	1	0	1	1	1	1	1	0	1
47	1	1	1	0	0	0	0	1	1
58	1	1	1	1	1	1	1	0	1
69	1	1	1	1	1	1	1	1	1
70	1	0	1	0	0	1	1	0	1
72	1	1	1	0	0	0	0	1	1
73	1	1	1	1	1	1	1	1	1
94	1	1	1	1	1	1	1	1	1
96	1	1	1	1	1	1	1	0	1
97	1	1	1	1	1	1	1	0	1
100	1	1	1	1	1	1	1	0	1
101	1	1	1	1	1	1	1	0	1
102	1	1	1	1	1	1	1	0	1

(*Appendices continue*)

Appendix B (*Continued*)

		Experiment 2		Experiment 3		Experiment 4		Experiment 5	
Scene	Experiment 1	Training	Test	Training	Test	Training	Test	Training	Test
200	1	1	1	0	0	1	1	1	1
201	1	1	1	1	1	1	1	1	1
202	1	0	1	1	1	1	1	0	1
204	1	0	1	1	1	1	1	0	1
205	1	1	1	1	1	1	1	0	1
206	0	0	1	0	1	1	1	0	1
208	0	0	1	1	1	1	1	0	1
209	0	0	1	0	1	1	1	1	1
210	0	0	1	0	1	1	1	1	1
211	0	0	1	0	1	1	1	0	1
212	0	0	1	0	1	1	1	0	1
213	0	0	1	0	1	1	1	0	1
214	0	0	1	0	1	1	1	1	1
215	0	0	1	0	1	1	1	0	1
216	0	0	1	0	1	1	1	0	1
217	0	0	1	1	1	1	1	1	1
218	0	0	1	0	1	1	1	0	1
219	0	0	1	1	1	1	1	0	1
220	0	0	1	1	1	1	1	0	1
221	0	0	1	1	1	1	1	1	1
222	0	0	1	0	0	0	0	0	1
224	0	0	1	1	1	1	1	0	1
225	0	0	1	1	1	1	1	1	1
226	0	0	1	0	1	1	1	0	1
227	0	0	1	0	1	1	1	0	1
228	0	0	1	1	1	1	1	1	1
229	0	0	1	0	1	1	1	0	1
230	0	0	1	0	1	1	1	1	1
232	0	0	1	1	1	1	1	0	1
233	0	0	1	1	1	1	1	1	1
234	0	0	1	0	1	1	1	0	1
235	0	0	1	0	1	1	1	1	1
236	0	0	1	0	1	0	0	0	1
237	0	0	1	0	1	1	1	0	1
238	0	0	1	0	1	1	1	0	1
239	0	0	1	0	1	1	1	0	1
240	0	1	1	1	1	1	1	1	1
241	0	0	1	0	1	1	1	0	1
242	0	0	1	0	1	1	1	0	1
243	0	0	1	0	1	1	1	0	1
244	0	0	1	0	1	1	1	0	1
245	0	0	1	0	1	1	1	0	1
246	0	0	1	0	1	1	1	1	1
247	0	0	1	0	1	1	1	0	1
248	0	0	1	0	1	1	1	0	1

Note. 1 = shown; 0 = not shown.

Received December 18, 2020
Revision received August 9, 2021
Accepted August 22, 2021 ■