Graphical convention formation during visual communication

Robert X. D. Hawkins*
Department of Psychology
Stanford University
rxdh@stanford.edu

Megumi Sano*
Department of Psychology
Stanford University
megsano@stanford.edu

Noah D. Goodman
Department of Psychology
Stanford University
ngoodman@stanford.edu

Judith E. Fan
Department of Psychology
UC San Diego
jefan@ucsd.edu

Abstract

Drawing is a versatile technique for visual communication, ranging from photorealistic renderings to schematic diagrams consisting entirely of symbols. How does a medium spanning such a broad range of appearances reliably convey meaning? A natural possibility is that drawings derive meaning from both their visual properties as well as shared knowledge between people who use them to communicate. Here we evaluate this possibility in a drawing-based reference game in which two participants repeatedly communicated about visual objects. Across a series of controlled experiments, we found that pairs of participants discover increasingly sparse yet effective ways of depicting objects. These gains were specific to those objects that were repeatedly referenced, and went beyond what could be explained by task practice or the visual properties of the drawings alone. We employed modern techniques from computer vision to characterize how the high-level visual features of drawings changed, finding that drawings of the same object became more consistent within a pair of participants and divergent across participants from different pairs. Taken together, these findings suggest that visual communication promotes the emergence of depictions whose meanings are increasingly determined by shared knowledge rather than their visual properties alone.

Keywords: drawing understanding; alignment; drawing; iconicity; visual abstraction

Introduction

From ancient etchings on cave walls to modern digital displays, visual communication lies at the heart of key human innovations (e.g., cartography, data visualization) and forms a durable foundation for the cultural transmission of knowledge and higher-level reasoning. Perhaps the most basic and versatile technique supporting visual communication is drawing, the earliest examples of which date to at least 40,000-60,000 years ago (Hoffmann et al., 2018). What began as simple mark making has since been adapted to a wide array of applications, ranging from photorealistic rendering to schematic diagrams consisting entirely of symbols.

Even in the relatively straightforward case of drawing from observation, there are countless ways of depicting the same object. How does a communication medium spanning such a broad range of appearances reliably convey meaning? On the one hand, prior work has found that semantic information in a figurative drawing, i.e., the object it represents, can be derived purely from its visual properties (Fan, Yamins, & Turk-Browne, 2018). On the other hand, other work has emphasized the role of socially-mediated information and context for making appropriate inferences about what even a figurative drawing represents (Goodman, 1976).

How can these two perspectives be reconciled? Our approach is to consider the joint contributions of visual information and social context in determining how drawings derive meaning (Abell, 2009), and to propose that a critical factor affecting the balance between the two may be the amount of shared knowledge between communicators. Specifically, we explore the hypothesis that accumulation of shared knowledge via extended visual communication may promote the development of increasingly schematic yet effective ways of depicting a physical object, even as these *ad hoc* graphical conventions may be less readily apprehended by others who lack this shared knowledge.

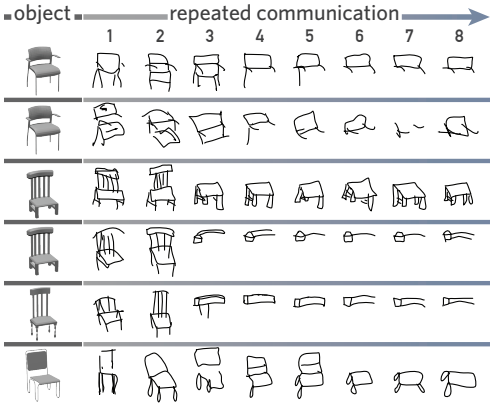

Figure 1: Repeated visual communication about the same object.

To investigate this, we used an interactive drawing-based reference game in which two players repeatedly communicate about visual objects, and examined both how their task performance and the drawings they produced changed over time (see Fig. 1). Our approach was inspired by a large literature that has explored how extended interaction influences communicative behavior in several modalities, including language (Krauss & Weinheimer, 1964; Clark & Wilkes-Gibbs, 1986; Hawkins, Frank, & Goodman, 2017), gesture (Goldin-Meadow, McNeill, & Singleton, 1996; Fay, Lister, Ellison, & Goldin-Meadow, 2014), and drawings (Garrod, Fay, Lee, Oberlander, & MacLeod, 2007).

There are three aspects of the current work that advance our prior understanding: *first*, we include a control set of objects that were not repeatedly drawn, allowing us to measure the specific contribution of repeated reference vs. general practice effects; *second*, we measure how strongly the visual properties of drawings drive recognition in the absence of interaction history for naive viewers, while equating other task variables; and *third*, we employ recent advances in

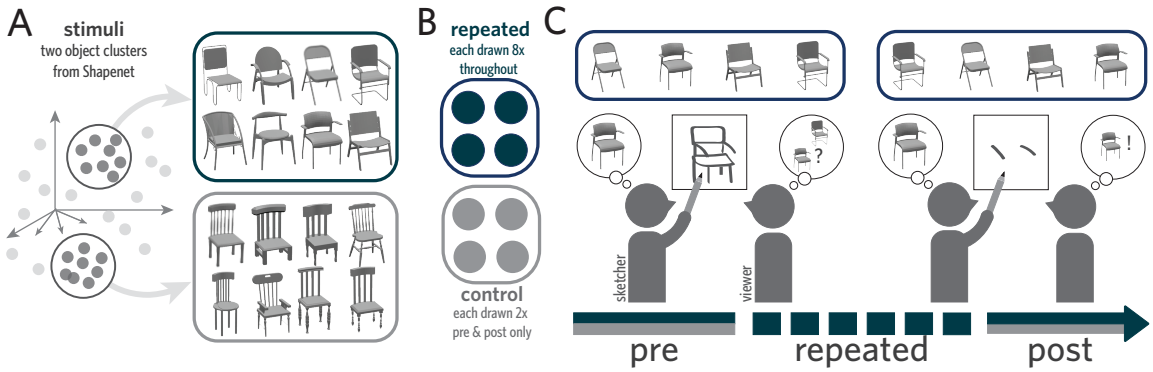

Figure 2: (A) Stimuli from ShapeNet. (B) Each pair of participants was randomly assigned two sets of four objects, each set from one of these categories. (C) Repeated objects were drawn eight times throughout; control objects drawn once at the beginning and end of each interaction.

computer vision to quantitatively characterize changes in the high-level visual properties of drawings across repetitions.

Part I: How does repeated reference support successful visual communication?

Our first goal was to understand how people learn to communicate about visual objects across repeated visual communication. To accomplish this, we developed a drawing-based reference game for two players. On each trial, both players were shown several images of objects, one of which was privately designated as the 'target' to the sketcher. The sketcher's goal was to draw the target so that the viewer could select it from the context as quickly and accurately as possible. We hypothesized that learning would be *object-specific*: that over repeated visual reference to a particular object, participants would discover ways of depicting that object more effectively relative to non-repeated control objects.

Methods: Visual communication experiment

Participants We recruited 138 participants from Amazon Mechanical Turk, who were automatically matched to form 69 pairs. Data from two pairs were excluded due to unusually low performance (i.e., accuracy $<$ 3 s.d. below the mean). In this and subsequent experiments, participants provided informed consent in accordance with the IRB[1].

Stimuli In order to make our task sufficiently challenging, we sought to construct contexts of objects whose members were both geometrically complex and visually similar. To accomplish this, we sampled objects from the ShapeNet database (Chang et al., 2015), which contains more than 51,300 3D mesh models of real-world objects. We restricted our search to 3096 objects belonging to the `chair` class, which is among the most diverse and abundant in ShapeNet. To identify groups of visually similar chairs, we first extracted high-level visual features from 2D renderings of each objects using a pre-trained deep convolutional neural network, VGG-19 (Simonyan & Zisserman, 2014). These 4096-dimensional feature vectors reflected VGG-19 activations to object renderings in the second fully-connected layer (i.e., `fc6`). We then applied dimensionality reduction (PCA) and k-means clustering on these feature vectors, yielding 70 clusters containing between 2 and 80 objects each. Based on these clusters, we selected two categories of visually similar objects containing eight exemplars each (Fig. 2A).

Task Procedure On each trial, both participants were shown the same set of four objects in randomized locations. One of the four objects was highlighted on the sketcher's screen to designate it as the target. Sketchers drew using their mouse cursor in black ink on a digital canvas embedded in their web browser (300×300 pixels; pen width of 5 pixels). Each stroke was transmitted to the viewer's screen in real-time and sketchers were not able to delete previous strokes. The viewer was allowed to guess the identity of the drawn object by clicking one of the four objects as soon as they were confident, and participants received immediate feedback: the sketcher learned when and which object the viewer had clicked, and the viewer learned the true identity of the target. Finally, participants were incentivized to perform both quickly and accurately. Both participants earned an accuracy bonus for each correct response, and the sketcher was instructed to take no longer than 30 seconds to produce their drawings. If the viewer responded under this time limit, participants also received a speed bonus inversely proportional to the time taken until the response.

Design For each pair, we randomly sampled two sets of four objects that served as contexts in the reference game: one was designated as a *repeated* set while the other was a *control* set (Fig. 2B)[2].The experiment consisted of three phases (Fig. 2C). During the central *repeated reference* phase, there were six blocks of trials, and each of the four *repeated* objects appeared as the target once in each block. In a pre-test at the beginning of the experiment, and a post-test at the end, both repeated and control objects appeared once as targets (in their respective contexts) in randomly interleaved order.

[1] All materials and data will be made available upon un-blinding at `https://github.com/XXX`.

[2] In half of the pairs, the four control objects were from the same stimulus cluster as repeated objects; in the other half, they were from different clusters. We collapse across these groups in our analyses.

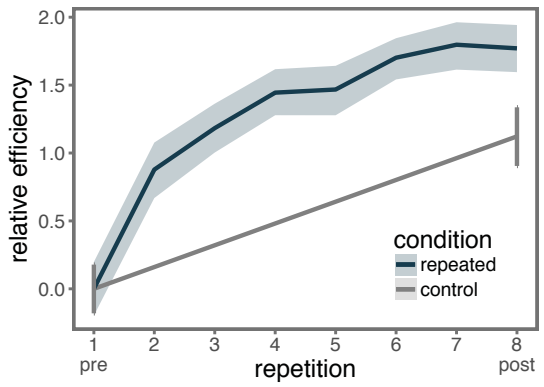

Figure 3: Communication efficiency for repeated and control objects across repetitions. Efficiency combines both speed and accuracy, and is plotted relative to the first repetition.

Results

Because objects were randomly assigned to repeated and control conditions, we expected no differences in task performance in the pre-test phase. We found that pairs identified the target at rates well above chance in this phase (75.7% repeated, 76.1% control, chance = 25%), suggesting that they were engaged with the task but not at ceiling performance. We found no difference in accuracy across conditions (mean difference: 0.3%, bootstrapped CI: $[-7\%, 7\%]$).

In order to measure how well pairs learned to communicate throughout the rest of their interaction, we used a measure of communicative efficiency (the *balanced integration score,* Liesefeld & Janczyk, 2018) that takes both accuracy (i.e., proportion of correct viewer responses) and response time (i.e., time taken to produce each drawing) into account. This efficiency score is computed by first z-scoring the accuracy and response time variables to map these values to the same scale and then subtracting the standardized response time from standardized accuracy. It is highest when pairs are both fast and accurate, and lowest when they make more errors and take longer, relative to their own performance on other trials[3].

To evaluate changes in communicative efficiency, we fit a linear mixed-effects model with maximal random effect structure, including random intercepts, slopes, and interactions for each pair of participants. We found a main effect of increasing communicative efficiency for all targets between the *pre* and *post* phases ($b = 1.45,\ t = 14.3,\ p < 0.001$), reflecting general improvements due to task practice. Critically, however, this analysis also revealed a reliable interaction between phase and condition: communicative efficiency improved to a greater extent for repeated objects than control objects ($b = 0.648,\ t = 3.09,\ p = 0.003$; see Fig. 3). Thus, there are benefits of repeatedly communicating about an object that accrue specifically to that object, suggesting the formation of object-specific graphical conventions.

[3]Results are similar for accuracy alone, but we adopted this integrated measure to better control for speed-accuracy tradeoffs.

Part II: What explains gains in efficiency?

Our visual communication experiment established that pairs of participants coordinate on more efficient and *object-specific* ways of depicting targets. This raises the question: to what extent do these gains in efficiency reflect the accumulation of *interaction-specific* shared knowledge between a sketcher and viewer, as opposed to a combination of task practice and the inherent visual properties of their drawings?

To measure the contribution of the latter, we conducted two control experiments that tested the recognizability of these drawings both inside and outside the social context in which they were drawn. One group of naive participants were shown a sequence of drawings constructed exclusively from a single interaction, thus closely matching the experience of a particular viewer in the visual communication experiment. For a second group, however, the sequence of drawings was pieced together from many different interactions. Under our *interaction-specificity* hypothesis, we predicted that the latter group would be impaired in their recognition performance compared to the former.

Methods: Recognition Control Experiments

Participants We recruited 245 participants via Amazon Mechanical Turk. We excluded data from 22 participants who did not meet our inclusion criterion for accurate and consistent response on attention-check trials (see below).

Task, Design, & Procedure On each trial, participants were presented with a drawing and the same set of four objects viewers originally saw accompanying that drawing. They also received the same accuracy and speed bonuses as viewers in the communication experiment. To ensure task engagement, we included five identical attention-check trials that appeared once every eight trials. Each attention-check trial presented the same set of objects and drawing, which we identified during piloting as the most consistently and accurately recognized by naive participants. Participants who responded incorrectly on at least four out of five of these trials were excluded from subsequent analyses.

Each participant was randomly assigned to one of two conditions: a *yoked* group and a *shuffled* group. Those in the yoked group were matched with one pair in the communication experiment and viewed 40 drawings in the same sequence the original viewer had. Those in the shuffled group were matched with a random sample of 10 distinct pairs from the communication experiment and viewed four drawings from each in turn, which appeared within the same repetition cycle as they had appeared originally. For example, if a drawing was produced in the fifth block of repetitions in the original experiment, then it also appeared in the fifth block here.

At the trial level, groups in both conditions thus received exactly the same visual information and performed the task under the same incentives to respond quickly and accurately. At the session level, both groups received exactly the same amount of practice recognizing drawings. Thus any dif-

ferences between these groups are attributable to whether drawings came from the same communicative interaction, which would support the accumulation of interaction-specific experience, or from several different interactions, where such accumulation would be minimal.

Results

Interaction-specific history enhances recognition by third-party observers We compared the yoked and shuffled groups by measuring changes in recognition performance across successive repetitions using the same efficiency metric we previously used. We estimated the magnitude of these changes by fitting a linear mixed-effects model that included group (yoked vs. shuffled), repetition number (i.e., first through eighth), and their interaction, as well as random intercepts and slopes for each participant. While we found a significant increase in recognition performance across both groups ($b = 0.18,\ t = 12.8,\ p < 0.001$), we also found a large and reliable interaction: yoked participants improved to a substantially greater degree than shuffled participants ($b = 0.10,\ t = 4.9,\ p < 0.001$; Fig. 4). Additionally, the yoked group was more accurate overall at identifying the target object (yoked: 75%, shuffled: 69%, $t = 3.6,\ p < 0.001$). Taken together, these results suggest that third-party observers in the yoked condition who could observe an entire interaction were able to take advantage of this continuity to more accurately understand which object each drawing represented. Observers in the shuffled condition who were deprived of this interaction continuity were significantly hindered in their ability to understand drawings even as the temporal order of the drawing was preserved.

Viewer feedback also contributes to gains in performance Unlike viewers in the interactive visual communication game, participants in the yoked condition made their decision based only on the final drawing and were unable to interrupt or await additional information if they were still uncertain. Sketchers could have used this feedback to modify their drawings on subsequent repetitions. As such, comparing the yoked and original communication groups provides an estimate of the contribution of these viewer feedback channels to gains in performance (Schober & Clark, 1989). In a mixed effects model with random intercepts, slopes, and interactions for each unique trial sequence, we found a strong main effect of repetition ($b = 0.23,\ t = 12.8,\ p < 0.001$) and a weaker but significant interaction with group membership ($b = -0.05,\ t = -2.2,\ p = 0.032$, Fig. 4), showing that the yoked group improved at a dampened rate as viewers in the original communication experiment. When considering pure recognition accuracy, however, the feedback gap was more substantial. Viewers in the original experiment had an overall success rate of 88% compared to 75% in the yoked condition, $t = 6.2,\ p < 0.001$[4].

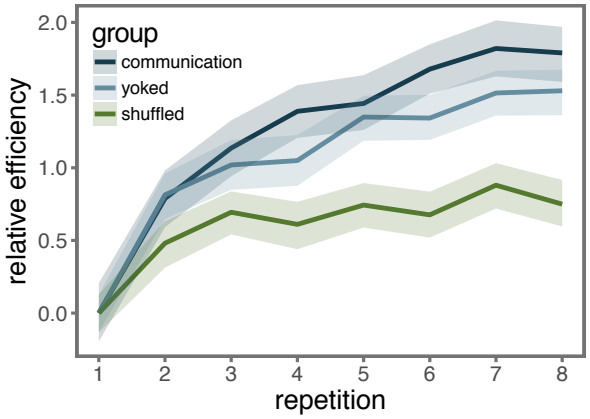

Figure 4: Comparing drawing recognition timecourse between viewers in communication experiment with those of yoked and shuffled control groups. Error ribbons represent 95% CI.

Part III: How do visual features of drawings change over the course of an interaction?

The results so far show that repeated visual communication establishes object-specific, interaction-specific ways of efficiently referring to objects. An intriguing implication is that interacting pairs achieved this by gradually forming *ad hoc* graphical conventions about what was relevant and sufficient to include in a drawing to support rapid identification of the target object. Here we explore this possibility by examining how the drawings themselves changed throughout an interaction. Concretely, we investigated four aspects that would reflect the increasing contribution of interaction-specific shared knowledge: *first*, decreasing number of strokes used (i.e., reducing motor cost of each drawing); *second*, increasing dissimilarity from the initial drawing produced (i.e., cumulative drift from the starting point); *third*, increasing similarity between successive drawings (i.e., convergence on internally consistent ways of depicting objects within an interaction); *fourth*, increasing dissimilarity between drawings of the same object produced in different interactions (i.e., discovery of multiple viable solutions to coordination problem).

Measuring visual similarity between drawings

Measuring visual similarity between drawings depends upon a principled approach for encoding their high-level visual properties. Here we capitalize on recent work validating the use of deep convolutional neural network models, pre-trained on challenging visual tasks, to encode such perceptual content in drawings (Fan et al., 2018). As when identifying clusters of similar object stimuli, we again used VGG-19 to extract 4096-dimensional feature vector representations for drawings of every object, in every repetition, from every interaction. Using this feature basis, we compute the similarity between any two drawings as the Pearson correlation between their feature vectors (i.e., $s_{ij} = {}^{cov(\vec{r}_i,\vec{r}_j)}/\sqrt{var(\vec{r}_i)\cdot var(\vec{r}_j)}$).

[4]Note that response times in the yoked condition were purely a function of viewer performance whereas in the original experiment they were a joint function of sketcher *and* listener behavior, so accuracy may be a purer measure for comparison.

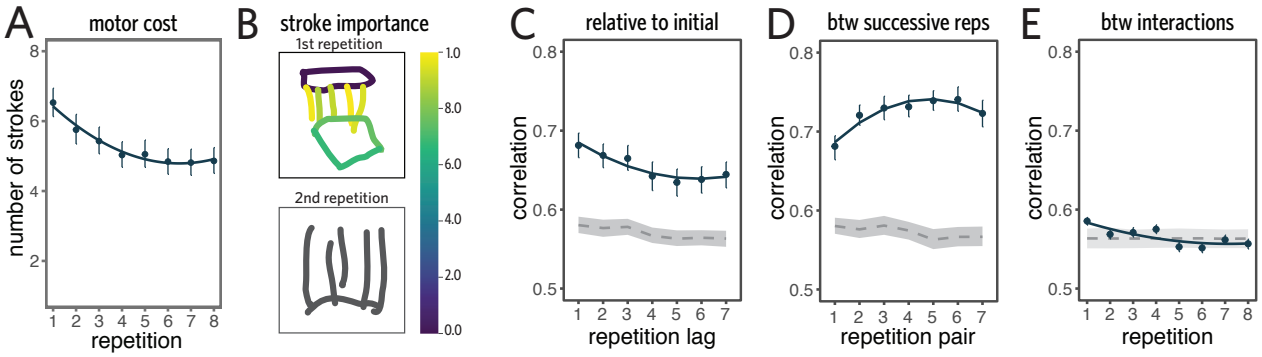

Figure 5: (A) Sketchers use fewer strokes over time. (B) Visualizing importance of individual strokes in successive drawings. (C) Drawings become increasingly dissimilar from initial drawing. (D) Drawings become more consistent from repetition to repetition. (E) The same object is drawn increasingly dissimilarly by different sketchers. Error ribbons represent 95% CI, dotted lines represent permuted baseline.

Results

Fewer strokes across repetitions A straightforward explanation for the gains in communication efficiency observed in Part I is that sketchers were able to use fewer strokes per drawing to achieve the same level of recognition accuracy by the viewer. Indeed, we found that the number of strokes in drawings of repeated objects decreased steadily as a function of repetition in a mixed-effects model ($b = -0.216$, $t = -6.00$, $p < .001$; Fig. 5A), suggesting that pairs were increasingly able to rely upon shared knowledge to communicate efficiently. This result raises a question about *which* strokes are preserved across successive repetitions during the formation of graphical conventions. In ongoing work, we are using a lesion method to investigate the "importance" of each stroke within a drawing for explaining similarity to the next repetition's drawing of that object. We re-render the drawing without each stroke and compute the similarity, yielding a heat map across strokes (see Fig. 5B for a preliminary visualization).

Increasing dissimilarity from initial drawing Mirroring the observed reduction in the number of strokes across repetitions, we hypothesized that there was also cumulative change in the visual content of drawings across repetitions. Concretely, we predicted that drawings would become increasingly dissimilar from the initial depiction. We tested this prediction in a mixed-effects regression model including linear and quadratic terms for repetition as well as intercepts for each target and pair. We found a significant decrease in similarity to the initial round across successive repetitions, ($b = -0.62$, $t = -5.59$; Fig. 5C), suggesting that later drawings had moved to a different region of visual feature space. However, since the entire distribution of drawings may have drifted to a different region of the visual feature space for generic reasons (i.e., because they were sparser overall), we conducted a stricter permutation test. We scrambled drawings across pairs but within each repetition and target and re-ran our mixed-effects model. The observed effect fell outside this null distribution ($CI = [-3.53 - 0.88]$, $p < .001$), showing that successive drawings by the same sketcher deviated from their own initial drawing to a greater degree than would be expected due to generic differences between drawings made at different timepoints in an interaction.

Increasing internal consistency within interaction As sketchers modified their drawings across successive repetitions, we additionally hypothesized that they would gradually converge on increasingly consistent ways of depicting each object. To test this prediction, we computed the similarity of successive drawings of the same object by the same sketcher (i.e. repetition k to $k+1$). A mixed-effects model with random intercepts for both object and pair of participants showed that similarity between successive drawings increased substantially throughout an interaction ($b = 0.53$, $t = 5.03$; Fig. 5). Again, we compared our empirical estimate of the magnitude of this trend to a null distribution of slope t values generated by scrambling drawings across pairs.The observed increase fell outside this null distribution, $CI = [-3.21, -0.60], p < .001$, providing evidence that increasingly consistent ways of drawing each object manifested only for series of drawings produced within the same interaction.

Inceasingly different drawings across interactions Our recognition control experiments suggested that the graphical conventions discovered by different pairs were increasingly opaque to outside observers. This effect could arise if early drawings were more strongly constrained by the visual properties of a shared target object, but later drawings diverged as different pairs discovered different equilibria in the space of viable graphical conventions. Under this account, drawings of the same object from different pairs would become increasingly dissimilar from each other across repetitions. We tested this prediction by computing the mean pairwise similarity between drawings of the same object from each repetition, but from different games. In a mixed-effects regression model including linear and quadratic terms, as well as random slopes and intercepts for object and pair, we found a small but reliable negative effect of repetition on between-game drawing similarity ($b = -1.4$, $t = -2.5$; Fig. 5E). We again conducted a permutation test to compare this t value with what would be expected from scrambling

sketches across repetitions for each sketcher and target object. We found that the observed *slope* was highly unlikely under this distribution ($CI = [-0.57, 0.60]$, $p < 0.001$), even if the similarity at each round was not so unlikely.

Discussion

In this paper, we investigated the joint contributions of visual information and social context for determining the meaning of drawings. We observed in an interactive Pictionary-style communication game that pairs discover increasingly sparse yet effective ways of depicting objects they repeatedly refer to. Through a series of control experiments, we demonstrated that these conventionalized representations were both object-specific and interaction-specific: drawings were harder for independent viewers to recognize without sharing the same history of interaction. Furthermore, by analyzing the high-level visual features of drawings, we found that they became increasingly consistent within an interaction, but that different pairs discovered different equilibria in the space of viable graphical conventions. Taken together, our findings suggest that repeated visual communication promotes the emergence of depictions whose meanings are increasingly determined by shared knowledge rather than their visual properties alone.

A key design choice in our visual communication paradigm was to use visual objects as the targets of reference, by contrast with the verbal cues (e.g. "art gallery") or audio clips used in prior work (Galantucci & Garrod, 2011). As such, pairs were presented with the same visual information about the shape and appearance of targets, encouraging the production of more 'iconic' initial drawings that more strongly resembled the target object. As their communication became increasingly efficient across repetitions, their drawings also became simpler and apparently more 'abstract'. An exciting direction for future work is to develop robust and principled measures of the degree of visual correspondence between any drawing and any target object, thereby shedding light on the nature of visual abstraction and iconicity.

A major open question raised by our work concerns how people use feedback and context when deciding which strokes to preserve or drop on each trial. According to one account, sketchers may be guided a *primacy* bias where they drop out later, detail-providing strokes to preserve their initial strokes. A second possibility is that they are guided by a *recency* bias: when a particular drawing is successful, they keep the final strokes they put down and drop out earlier ones. Finally, these decisions may be more strongly guided by the communicative informativity or diagnosticity of each stroke than by their temporal sequence. For example, sketchers may remove an entire semantically meaningful part (e.g. the backrest) if it does not help distinguish the target from distractors in context. Computational models of sketch production, or lesion analyses like those preliminarily reported above, are promising approaches toward distinguishing among these possibilities in future work.

Visual communication is a powerful vehicle for the cultural transmission of knowledge. Over time, advancing our knowledge of the cognitive mechanisms underlying the formation of graphical conventions may lead to a deeper understanding of the origins of modern symbolic systems for communication and the design of better visual communication tools.

References

Abell, C. (2009). Canny resemblance. *Philosophical Review*, *118*(2), 183–223.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., ... others (2015). Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*.

Clark, H. H., & Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, *22*(1), 1–39.

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science*.

Fay, N., Lister, C. J., Ellison, T. M., & Goldin-Meadow, S. (2014). Creating a communication system from scratch: gesture beats vocalization hands down. *Frontiers in Psychology*, *5*, 354.

Galantucci, B., & Garrod, S. (2011). Experimental semiotics: a review. *Frontiers in Human Neuroscience*, *5*, 11.

Garrod, S., Fay, N., Lee, J., Oberlander, J., & MacLeod, T. (2007). Foundations of representation: where might graphical symbol systems come from? *Cognitive Science*, *31*(6), 961–987.

Goldin-Meadow, S., McNeill, D., & Singleton, J. (1996). Silence is liberating: removing the handcuffs on grammatical expression in the manual modality. *Psychological Review*, *103*(1), 34.

Goodman, N. (1976). *Languages of art: An approach to a theory of symbols*. Hackett publishing.

Hawkins, R. X. D., Frank, M. C., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proc. of the 39th Annual Meeting of the Cognitive Science Society.*

Hoffmann, D., Standish, C., García-Diez, M., Pettitt, P., Milton, J., Zilhão, J., ... others (2018). U-th dating of carbonate crusts reveals neandertal origin of iberian cave art. *Science*, *359*(6378), 912–915.

Krauss, R. M., & Weinheimer, S. (1964). Changes in reference phrases as a function of frequency of usage in social interaction: A preliminary study. *Psychonomic Science*, *1*(1-12), 113–114.

Liesefeld, H. R., & Janczyk, M. (2018). Combining speed and accuracy to control for speed-accuracy trade-offs. *Behavior Research Methods*.

Schober, M. F., & Clark, H. H. (1989). Understanding by addressees and overhearers. *Cognitive Psychology*, *21*(2), 211–232.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.