Kirsh, D. Situating Instructions, in
In: Kokinov, B., Karmiloff-Smith, A., Nersessian, N. J. (eds.) ***European Perspectives on Cognitive Science***.
 ISBN 978-954-535-660-5

Situating Instructions

David Kirsh (kirsh@ucsd.edu)
Dept of Cognitive Science
University of California, San Diego

Abstract

A videographic study of origami is presented in which subjects were observed making four different origami objects under five modes of instruction: photos + captions, illustrations-only, illustrations with small captions, illustrations with large captions, and text-only as control. The objective of the study was to explore the gestures and other actions that subjects produce as they try to follow instructions rather than to determine the most effective style of instruction per se. We found that the task of situating instructions to the context at hand is error prone and that to facilitate it subjects gesture, point, re-orient illustrations, and generally do things that have no function other than to change the epistemic and interactive landscape of activity so they can more easily understand what is to be done. These studies bear on the new questions designers are asking about the placement, timing, and pace of instructions that digital aids now provide and on the fundamental question of how humans embed themselves in an activity by framing their task in a situation specific manner.

Keywords: Situated cognition; instructions; registration; interactivity; framing; embedding; design.

Introduction

The study reported here focuses on the 'extra' actions that people perform to make sense of origami instructions. For simplicity, we can call these extra actions 'instruction-comprehending' actions or 'gestures'. They are not communicative gestures; they are gestures whose function is to facilitate correspondence, or alignment between the semantic elements in origami instructions and the elements and procedures involved in making origami structures. This process of moving from a shareable, public representation of an action to a clear idea of the personal, perspectival action to be performed in the here and now is at the heart of situating instructions. It is part of the story of how people frame an activity space and so tune or reshape their goals and expectations to get things done where they are and in the context they face.

The background for this inquiry is an observation that has never failed: the more closely we observe people the more evident it is that they do *other* things when performing a task than just task-advancing actions [Kirsh 09, Kirsh & Maglio 94]. Whenever people follow instructions they inevitably spend part of their time making sense of the instructions. This might be done silently and in the head. But more often, there is something external and physical that they do, something that involves reconfiguring the environment, or themselves, something that is meant to help them understand their task better.

For instance, when people are given directions in a mall, they do their best to locate landmarks and cues in real time. Gestures on the speaker's part help to direct the listener's attention to elements that can *anchor* terms used in discourse. As the speaker utters 'Turn right at Macy's, go past the cell phone kiosk, and when you see the William's and Sonoma…' she might point, or orient her body to help the listener identify cues to fixate on. Tying instructions to attributes, structures, and objects in the environment is part of the process of situating instructions. It doesn't bring an agent *physically* closer to the goal, but it is a necessary part. When the key information resource is a map, rather than an informant, the need to situate instructions often involves re-orienting the map, pointing to landmarks, moving one's finger over the map, or keeping place with one hand while searching for physical correspondences. What kind of thing are these actions? In a slightly different fashion, when a cook is given a novel recipe, he needs a moment to mentally elaborate the steps, to translate them into a set of *projections*, *intentions and expectations* – an activity plan – that makes sense in terms of the surfaces and layout of the kitchen, the tools at hand, the ingredients, and other elements in the kitchen. These crucial elaborations are assumed to be part of a cook's skill set. But sense making is not part of following the recipe; it is part of interpreting it. This act of interpretation often involves physically re-shaping the cognitive scaffolds in the kitchen; making the kitchen cognitively congenial to the task at hand.

In origami, as I will soon show, it's the same story. There is a gulf between what the instructions say to do, and what an agent must actually do to follow them. Resolving this gulf is what situating instructions is about. It is invariably non-trivial, and involves doing things that have little to do with the instructions themselves, but much to do with jigging the mindset of the subject and reorganizing the physical setup.

Improving interaction design is a second background motive of this study. Because situating instructions is effortful and error prone, any deeper understanding of how we do it may lead to improved designs of instruction-rich environments: kitchens, factories, laboratories and hospitals. Historically, outside the field of education, psychological studies of instruction have focused on the comparative value of media, for instance, determining when animations are better than illustrations [Mayer et al, 2005, Wong 2009]. Few studies have focused on how a subject interactively engages media while following directions, or the way the environment is altered by what that subject does to simplify instruction following. With

the pervasiveness of digital enhancement, however, we can now radically alter the way directions are given. This raises a new set of questions concerning the placement, timing, and pace of instructions that go beyond the classical concerns with form and content. For example, *where* should instructions, or instruction parts, be placed to help an origami player as he moves deeper into the folding process? *When* should the steps be shown or 'read out'? Timing and placement are resources to be optimized for cognitive ends. They can scaffold execution.

Origami Study

Our conjecture: People perform special instruction-comprehending actions (called gestures above) in order to facilitate their origami performance. These actions are not necessary for completing the assigned tasks, but occur naturally and are apparently of value to the subjects. The nature and frequency of these actions varies with the format and mode of presentation of the instructions.

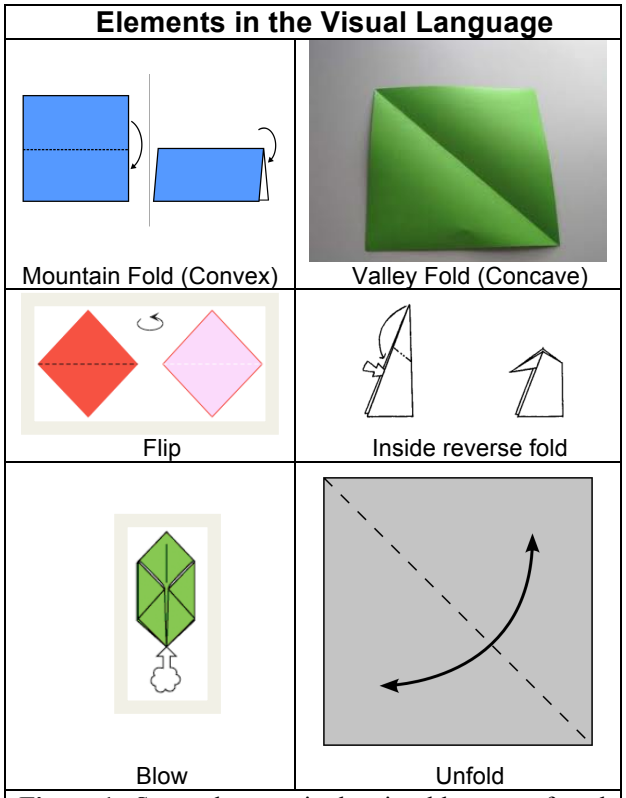


Figure 1. Some elements in the visual language found in Origami instructions. Note that not all actions are fold related. Each of these actions, as well as others not mentioned in instructions, were named and coded when we annotated the videos of origami activity.

Method: to demonstrate that most people perform instruction-comprehending actions, we must first distinguish task-necessary from task-unnecessary actions, and tie these to instructions. Instruction-comprehending actions are a special class of task-unnecessary actions – a type of epistemic action, rather than a pragmatic action [Kirsh & Maglio 94]. Anchoring, registration, and certain other gestures are the sort of task-unnecessary actions we want to identify. Muttering would be another example of a task-unnecessary action, but it was not studied here.

To operationalize the distinction, we collected several dozen origami instruction sets from the Internet and from the book, The Complete Origami Collection (Takahama, 1997). These were used to identify the basic elements of the visual language of origami instructions. See figure 1 for a subset of that system. Each element in this visual language was then matched with a behavior we observed in origami activity. Actions that origami players performed that were not specified in the instructions are prima facie task-unnecessary actions. Of course, whether an action is specified explicitly in an instruction is not always transparent. It depends on the granularity of observation and whether the observer treats proper sub-goals of a larger goal to be part of the instruction. Operating within these qualifications, if 'unnecessary' actions can be shown to be adaptive or helpful, especially to situate or make sense of an instruction, then we have found a physical, behavioral thing that subjects do to improve understanding, something that they do for reasons other than to bring them physically closer to achieving a goal or sub-goal, i.e. an epistemic action, a cognitively helpful but task-unnecessary action.

Stimuli: We created five types of stimuli or 'instruction styles' to reflect the different ways origami instructions are given: Text only (control), Illustrations + Short Caption, Images/Illustrations Only, Illustrations + Long Caption, Photos + caption. Jointly, the five instruction styles define all the elements in the visual language of origami, though

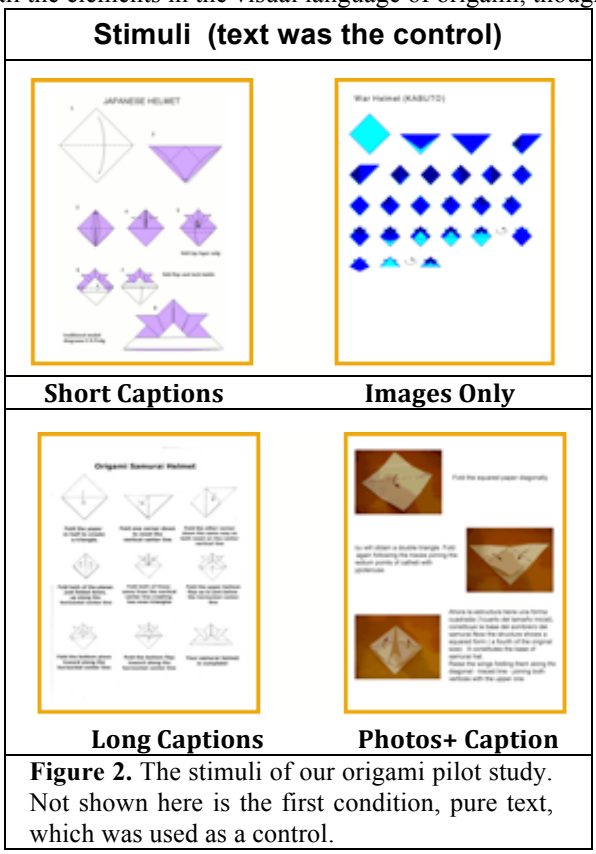


Figure 2. The stimuli of our origami pilot study. Not shown here is the first condition, pure text, which was used as a control.

it is possible that new illustrators might introduce new elements, or that players might invent new kinds of action. We tested subjects on four origami objects – a balloon, a cicada, a crane, and a helmet – in each instruction style.

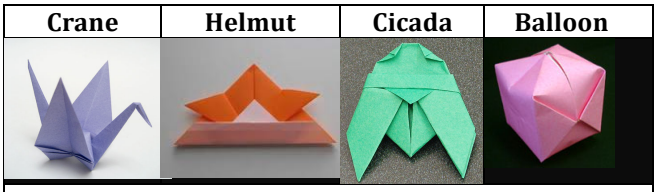


Figure 3. The four origami objects to be made.

Subjects: Twelve subjects (three females, nine males) were recruited from the UCSD undergraduate population.

Procedure: Each subject folded all four objects. Each object was specified in a different instruction style. This meant that each subject was run on four of the five instruction styles, a decision taken because of the time some subjects took to complete four objects. In order to control for differences in the experience and spatial ability of the subjects, as well as for learning effects brought on by the varying difficulty of the objects and types, a Latin square design was employed to assign experimental conditions to the subjects. Instruction types were assigned using a five by five Latin square (minus a column). A different four by four Latin square was used to assign the order each subject folded the objects.

Two video cameras recorded participants' actions from different angles. Camera One filmed a side shot of the participants' upper body and the folding surface (table), capturing the entire action-space of the scene. Camera Two filmed a close-up of the participants' hands.

Before the origami folding sessions began, participants filled out a questionnaire to determine their origami experience, how long it had been since their last origami activity, and whether they were right handed, left handed, or ambidextrous. They also performed a spatial task to estimate their spatial cognition skills.

During the session, subjects were instructed to vocalize their thought process as they worked with the instructions. After completing each object, they were asked to rate each step from one (easy) to five (difficult), and to explain their reasoning verbally. They were also invited to comment freely.

If participants did not follow an instruction correctly and improperly folded the object, they were given the time it took them to complete the next step to realize their error. If they did not recognize their mistake in that time, the experimenter would stop them and give them a hint. If they did not perform the step correctly in three minutes, the experimenter would perform the step for them, and then the participant would continue on to the next step.

Gross quantitative Results: Mean performance time across all conditions was 8:51 mins/object with a mean error rate of 2.58. A correlated-samples ANOVA could not be calculated because subjects were not run on all instruction styles. Nonetheless, there were clear general trends derivable from partial correlated samples Anova's (mean p value = .16): Short Captions were fastest and with fewest errors, Long Captions were slowest, and Photos + text, though fast, caused vastly more errors. See figure 4.

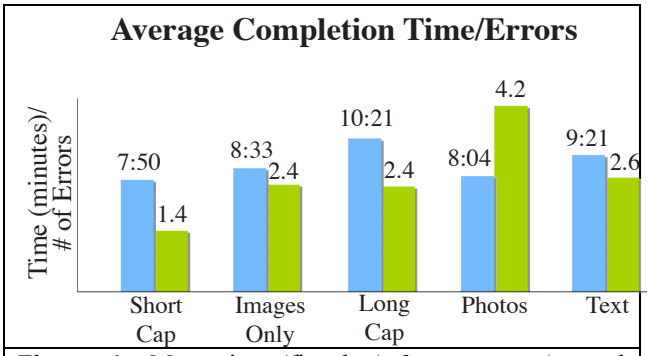


Figure 4. Mean time (first bar) & error rate (second bar) for the five different instruction styles.

It would be necessary to run more subjects to deepen these quantitative findings. But that was never our real goal. Our real interest lay in the fine grain of behavior that shows how subjects work with instructions. What contortions do they go through to make sense of instructions given their work context?

Qualitative Results. To study the behavior of subjects we coded our video, including talk-aloud commentary, and took informal notes from the video'ed debriefings. Coding is an incremental process: as one phenomenon comes into focus it is useful to return to the video and add more behaviors or attributes to code; hence the code is expanded.

On the basis of the debriefing and our desire to explain the error level of Photos, for instance, we reviewed the video and found that one reason that Photo instructions, cause an abundance of errors is that subjects each have their own *style* of working with origami paper and their instruction sheet. Some subjects make folds upside-down in relation to the shapes shown on the instruction sheet; others prefer to match the orientation of their paper with the diagrams; and still others reorient the paper each time they perform a fold. A Photo instruction set, even with captions, is of necessity based on a specific folding style – the author's – since it is his approach that is photographed. For subjects other than the lucky ones whose personal style matches the author's, the subjective complexity of instructions goes up as they have to mentally or physically adapt. This predicts increased errors for subjects with style mismatch.

Moreover, because the edges and creases of paper when photographed are not as clear-cut as when drawn subjects also can be expected to have a harder time comparing, registering, and verifying shapes. Even the extra detail that comes with photography is not a help since it means that what a subject sees in a photo may be different than what they have in their hands. Illustrations abstract from irrelevant detail, photos do not. If a photo contains

irrelevant or misleading detail it is the subject herself who must mentally compensate.

Some other qualitative causes of error can also be predicted in advance of observations. For instance, in the illustration-only condition, there are neither symbolic annotations nor captions to explain what is to be done. Between some illustrations there are arrows but there are no annotations on the illustrations themselves. Accordingly, actions must be inferred from before and after shots. This can be challenging. These factors likely contribute to the higher error rate we found in the photo + text and illustration-only conditions.

Coding: To code the videos we first enumerated the semantic elements in the collection of origami instructions we had. In table 1 column one contains all referential elements of the visual language found in the instructions we used. Less than 50% of these designate actions, as shown in column two. The rest denote structural elements, or they increase the specification of a form (e.g. they show object depth) or they help disambiguate sides (origami paper is colored differently on front and back). In the third column are the actions that are unnecessary for completing the object. We distinguished only five such gestures – registration, verification, pointing, shrugging, and gestural thought – because these seemed more basic than for instance, anchoring, which could be achieved by pointing or eye movements or registration. We chose these terms on the basis of what emerged in debrief and during talk-alouds, as subjects explained how they attempted to fix the referent of visual elements and what they did when trying to work out a procedure that would produce the state depicted in an illustration. It is challenging to justify our selection in a less subjective manner. Indeed, these 'gestures' might be thought to be epiphenomena, not actually part of the origami activity – a view partly justified by our observation that better players perform fewer of these actions than novices. But when better players are challenged or given complex instructions these 'superfluous' gestures recur, suggesting that they are part of the interpretation process that are omitted once practice leads to chunking and more automatic behavior.

Language	**Code for Actions**	
Visual Language Elements	**Necessary Actions**	**Unnecessary Actions**
Arrow for valley Fold	Valley Fold	**These have no counterpart in Visual Language**
Arrow for mountain fold	Mountain Fold	
Fold into flap	Flap Fold	
Unfold	Unfold	
Bring corners together	Bring corners together	
Blow into object	Blow/inflate	
Turn object over	Turn over	*Registration*
Pull object	Pull	*Verification*
Pre-fold state	**No counterpart**	*Pointing*
Post-fold state		*Shrug*
Labeled corners		*Gestural-Thought*
Labeled non-corners		
Mountain fold line		
Valley fold line		
Detail pop-up		
Object depth		
Side color		

Table 1. Column one contains all referential elements of the visual language found in the instructions we used. Column two lists the actions necessary to physically make the required origami pieces. Column three lists the actions that subjects regularly performed that were unnecessary to physically complete their pieces. These seem related to making sense of the instructions given the current state of the structure they were making.

Actions that help situate instructions

Our best evidence that at least some of these gestures serve a sense-making function is that gesturing predicts error. As can be seen in Table 2, gestures are disproportionately present when an error is committed. Unless a gesture is itself the cause of an error, the reason a subject gestures is most likely because he or she is having trouble understanding an instruction and is gesturing to somehow help or facilitate comprehension, though shrugging (see below) is likely different since it probably signals errors rather than helps comprehension.

Gesture	**% in Error Cells**	**Expected in Error Cells**
Registration	13.0	10
Verification	20.4	10
Gestural Thought	23.8	10
Attention Focusing (Pointing)	21	10
Trying it out	25	10
Shrugging (see above)	36.3	10

Table 2. Gestures were disproportionately present in error cells, suggesting they correlate with the instruction complexity and may serve a helping or related function.

In our video analysis we coded a total of 1179 action cells. An error was made in 125, or 10.6%, of these cells. Prima facie, gestures should be uniformly distributed over cells. There is no reason why a subject should gesture in one cell rather than another. Therefore, one would expect that only about 10% of any type of gesture will be in error cells. This was hardly the case. A verification gesture was performed 270 times with 55 of these appearing in error cells. Error verifications accounted for 20.4% of the total verifications, or about two times the expected amount. From talk-alouds it was apparent that when subjects made errors, they often had trouble understanding the instructions and tried to see if their objects matched the diagram depictions. Data from the gestural thought and attention focusing (pointing) support this finding, with errors accounting for 23.8% and 21.1%, respectively.

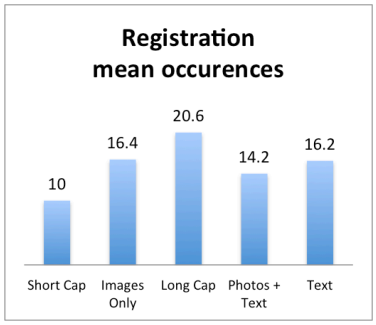
These gestures, too, indicate difficulties interpreting the instructions. Shrugs correlated most highly with error and were explained as signaling frustration or ignorance.

Overall, each of the gestures, besides registration, was overrepresented in the error cells by a factor of two. Error registrations only accounted for 13.0% of the total registration actions. This may reveal that registration is equally use*ful* all the time or reveal the exact opposite: it is use*less* all of the time, purely epiphenomenal.

I now turn to an explanation of these actions.

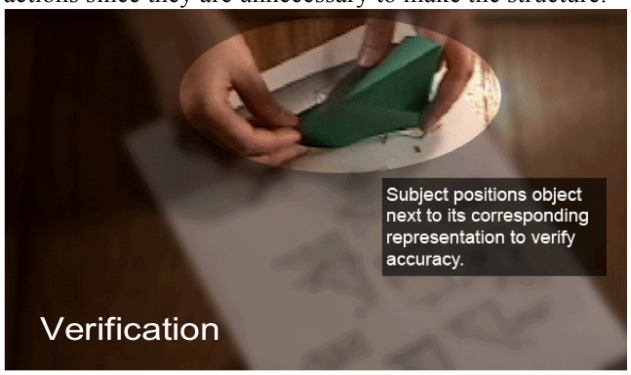
Registration: Registration refers to the process of aligning a *representation* with its *physical* reference. There is a surprising range of actions that people perform to a) bring an origami instruction or origami structure into mutual alignment, b) maintain that alignment, and c) test that they are correctly aligned**.**

A common method is to begin registration by identifying a few symbolic features – side of sheet, numbered corner, orientation of a fold – with visible features in the origami structure. But it is piecemeal. It must be repeated for other features. Accordingly, people will often use those easy correspondences as *anchors* and re-orient their map completely, thereby making it easier to maintain those correspondences, and easier to interpret new correspondences. To make sure the re-orientation is right – the testing phase – they typically check a few other symbolic features or shapes to see that the relational structure on the map mirrors the relational structure in the world.

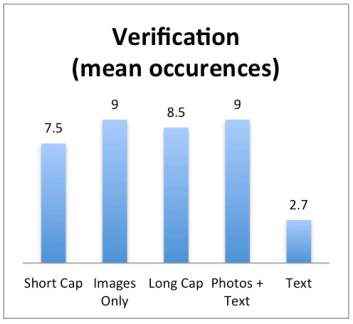


Physical realignment (registration) rather than mental realignment is powerful because it reduces the cost of translating between the outside world and the inside representation by bringing the two systems into alignment. In maps, it is easy to achieve; in instructions, far less so. It is a core sense making process, of situating an abstract representation in context; it deserves deeper study.

Verification gestures typically occur after a subject has come up with a candidate interpretation of an instruction and wants to check to see if it will hold water, given closer scrutiny of the image depicted or the specification given in words. At times, a subject will hold his current structure beside an instruction before beginning to execute the next step. We coded such actions as verifications only if it was clear the objective was to verify that no error had been made earlier. Given this post-facto role in construction, it is not likely that verification gestures are part of instruction sense making narrowly conceived. But they are epistemic actions since they are unnecessary to make the structure.

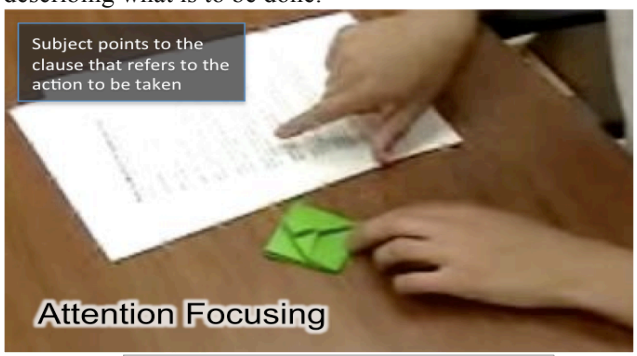


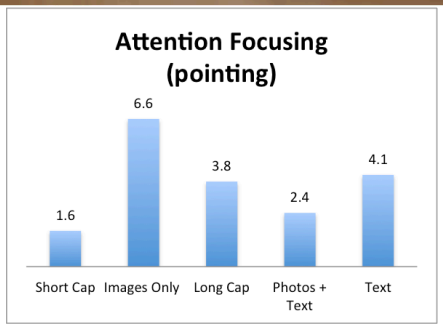
Verification is a high frequency gesture occurring in almost 24% of all action cells. In the Text condition the number of verification gestures (2.7) falls significantly below the others (ranging from 7.5 to 9). Why? Verification, as a gesture, and not simply a term to refer to any genuine mental recalculation, involves moving the current origami structure physically close to its depiction, and visually probing the two in a systematic manner to ascertain whether the current structure realizes the depiction. Since purely textual instructions do not lend themselves to this test the subject rarely bothered to bring instruction and paper structure close together and systematically review them.



Gestural thought occurs when an instruction is not clear enough for a subject to understand the action being depicted without some extra scaffolding. The gesture seems to facilitate spatial reasoning. For instance, in the photo above the subject is trying to work out the implications of unfolding and refolding before attempting to follow the instruction. The unfolding step requires undo'ing something already done and so, quite naturally, the subject wants to be clear about the point of the action before incurring the cost of taking something apart. This action is a good candidate for a gesture that aids in sense making because it helps the subject to see the point of the instruction. It reveals that the action is meant to be understood literally, or that it leads to a plausible sub-goal. Another function of these actions is that the gesture, while not quite an action of trying out, is nonetheless a gesture that embodies the main idea; it is a mock try out, a partial simulation. In dance, this kind of gestural thought is often called 'marking' [see Kirsh, 2010].

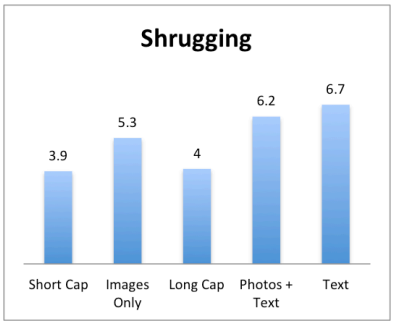
Attention focusing manifests gesturally as a type of pointing. Subjects performed this gesture when they have lost their place or when they are re-reading an instruction they do not immediately understand. The standard action is to keep a finger on the problematic clause. Sometimes they hold their finger on the noun phrase describing the part of their structure to be altered, or the verb phrase describing what is to be done.





The ***shrug gesture*** was performed when a subject was unsure of an instruction's meaning, but committed their guess to action. This gesture was usually accompanied by the subject vocalizing their uncertainty about a phrase, such as, "I guess". Error shrugs accounted for 36.3% of the total shrugs.

Shrugging is common. From discussion and close observation shrugging usually indicates a form of acceptance. Both the Text condition and the Photo + text condition have high shrug counts. Shrugging usually comes either after the study phase of an instruction and just before the assembly phase, or it comes at the end of an action. Shrugging rarely co-occurs with verification. We conjecture that in the text condition, where it is difficult to verify, shrugging shows that subjects are in a heightened state of uncertainty. They have found a solution but it is a solution that might likely be incorrect. Photos, similarly, are difficult to work with when folds and creases are hard to see in the photo. Subjects take decisions, but as is apparent from their very high error rate in the Photo condition, they cannot always tell if they are right.



Conclusion. In our 20 hours of origami observations, and coding of nearly 1200 events, we noted a variety of registration actions, partial try-outs, shrugging, muttering, asking advice or clarification, looking up terms, and more. These are not incidental elements of origami activity: they are part of the process people follow of embedding themselves in their activity space. Embedding facilitates sense making. To follow an instruction in origami a subject must know what the instruction implies doing to the at-hand materials. This requires interpreting how textual, figural or pictorial instructions relate to three-dimensional structures and inferring the actions that appropriately reshape those structures – a hard cognitive task that epistemic actions help to simplify.

Acknowledgments: I thank Monica Okubo for her careful and thoughtful work collecting, coding and initially analyzing the video data. Cody Frew was generous with his time helping to code.

References

Kirsh, D. Problem Solving and Situated Cognition. In, Philip Robbins & M. Aydede (Eds.) (pp. 264-306) The Cambridge Handbook of Situated Cognition. Cambridge: Cambridge University Press, 2009.

Kirsh, D. and P. Maglio. On Distinguishing Epistemic from Pragmatic Actions. Cognitive Science. Vol. 18, No. 4: pages 513-549. (1994).

Kirsh, D. Thinking With The Body. In (eds) S. Ohlsson, Catrambone, Proceedings of the 32nd Annual Meeting of the Cognitive Science Society. (2010).

Mayer, R.E, Hegarty M, Mayer S, Campbell J. When Static Media Promote Active Learning: Annotated Illustrations Versus Narrated Animations in Multimedia Instruction. Journal of Experimental Psychology: Applied. (2005), Vol. 11, No. 4, 256–265

Takahama, Toshie. The Complete Origami Collection. Japan Publications (USA). 1997.

Wong, Anna, t al. (2009). Instructional animations can be superior to statics when learning human motor skills, Computers in Human Behavior 25 (2009) 339–347.