Discourse Processes

# Unpacking the gestures of chemistry learners:

# What the hands tell us about correct and incorrect conceptions of stereochemistry

--Manuscript Draft--

| | |
|---|---|
| **Manuscript Number:** | DP-D-18-00124R1 |
| **Full Title:** | Unpacking the gestures of chemistry learners:<br><br>What the hands tell us about correct and incorrect conceptions of stereochemistry |
| **Article Type:** | Original Article |
| **Corresponding Author:** | Susan Goldin-Meadow<br>University of Chicago<br>UNITED STATES |
| **Corresponding Author Secondary Information:** | |
| **Corresponding Author's Institution:** | University of Chicago |
| **Corresponding Author's Secondary Institution:** | |
| **First Author:** | Raedy Ping, PhD |
| **First Author Secondary Information:** | |
| **Order of Authors:** | Raedy Ping, PhD |
| | Ruth B Church, PhD |
| | Mary-Anne Decatur, PhD |
| | Samuel Larson, PhD |
| | Elena Zinchenko, PhD |
| | Susan Goldin-Meadow, PhD |
| **Order of Authors Secondary Information:** | |
| **Abstract:** | When learners explain answers to tasks they have yet to master, they often gesture. In a pretest- posttest design, we show that these gestures provide insight into what learners know about a chemistry task, and whether they are ready to make gains on that task. Adults, naïve to organic chemistry, drew stereoisomers of molecules and explained their drawings. All participants gestured spontaneously during explanations, and often expressed strategies only in gesture (and not in speech). In some cases, these strategies conveyed information that was explanatorily relevant to the problem; in other cases, they conveyed information that was explanatorily irrelevant. Relevant strategies produced only in gesture on the pretest, and no other types of pretest strategies, predicted posttest performance. This finding supports the hypothesis that information conveyed uniquely in gesture and not in speech indicates readiness-to-learn, but advances the argument: Gesture-speech mismatch predicts learning not because it reflects general upheaval in learners, but because it reveals explanatorily relevant implicit knowledge that promotes change. |

Dear Dr. Richter:

We would like to resubmit our manuscript, *Unpacking the gestures of chemistry learners: What the hands tell us about correct and incorrect conceptions of stereochemistry* (Ref.: Ms. No. DP-D-18-00124). We found the reviews to be very helpful and have completely revamped our analyses along the lines suggested by you and the reviewers.

We have copied the action letter that you sent us below in italics, along with a description of how we have altered the revised manuscript in response to each point. We think the paper is substantially stronger than the original submission, and thank you and the reviewers for your suggestions.

We hope that the paper is now acceptable for publication in *Discourse Processes* and look forward to hearing from you.

**Responses to the Editor's Concerns:**

*The reviewers and I agree that the manuscript deals with an interesting and timely topic. The role of gesturing for learning is clearly relevant for readers of Discourse Processes. However, the reviewers and I also agree that the manuscript suffers from certain weaknesses regarding the methodology (design and data analysis) used in the study reported in the manuscript. These weaknesses considerably limit the contribution of the manuscript. Therefore, I cannot accept the manuscript for publication in its present form. However, it is possible (albeit not guaranteed) that the weaknesses can be fixed by a thorough revision of the manuscript.*

*Please let me briefly summarize the main concerns that emerged in my reading of your manuscript and the reviews.*

*1. Your study is based on a correlational design (a one-group pre-post test design), whose internal validity is low. A much better option would have been to manipulate gesturing experimentally (for example, by comparing one group that was allowed to gesture and the other one was not or to encourage certain types of gestures). This would still have allowed you to examine the question whether certain types of gestures are associated more closely to learning. As your design is purely correlational, you need to be much more careful with causal explanations ("relevant mismatches catalyze learning"...) and you need to utmost care to control potentially confounding variables.*

**Response:** We agree that any causal connection would have to be proven with an experiment. Our goals in this study were three-fold: (a) to identify whether gestures occur in adult learners of stereochemistry, (b) to carefully code the information that adults produce spontaneously in speech and gesture when explaining their responses to stereoisomer problems, and (c) to determine whether this information (its content and/or its form) predicts success on the problems after a brief lesson. Before attempting an experimental paradigm in which we manipulate gesture (which would address the causal question), we needed to examine naturalistic data to discover the types of gestures learners spontaneously produce in this context. Armed with this information, we can design the next study, which would address the causal question. We therefore agree with the reviewers that we need to avoid causal language and we have made the necessary wording changes throughout the manuscript.

We also wanted to fit our study into the literature on gesture-speech mismatch, which has been shown to predict learning in children. We know from previous research (described in the paper) that adults produce gesture-speech mismatches when talking about concrete and abstract ideas. Gesture-speech mismatch is therefore not just an outgrowth of being a developing child or a child learner. Our question is whether gesture-speech mismatch reflects readiness-to-learn in adult learners, as it does in child learners. We therefore replicated the child procedure—we examined the gestures and speech that adults spontaneously produced when asked how they solved stereoisomer problems; we gave them a brief lesson on how to solve the problem; and then we explored whether their pre-lesson gestures predicted their post-lesson performance.

*2. One potential confound that immediately comes to mind is with the total number of gestures. You need to control for matching gestures (i.e. include in your regression model as well). However, the matching gestures should not include trials with no gestures or uncodeable gestures. As Reviewer 2 notes, this does not make a lot of sense and compromises the validity of this category. And shouldn't pretest performance be controlled for as well?*

**Response**: We have totally revamped the way that we analyze gesture-speech mismatch in these data. The reviewers' concerns helped us realize that the adults' explanations cannot be coded in the same way as the children's simply because the adults frequently produce more than one strategy in the same modality within an explanation (typically, more than one strategy in speech). If, for example, an adult produces two strategies in speech, one strategy might match the accompanying gesture, but the other might not. We therefore cannot easily classify an explanation as either a match or a mismatch (this can easily be done in the child data as the children rarely produce two speech strategies or two gesture strategies within a single explanation).

As a result, we now analyze the data at the level of the strategy. We classified each strategy that an adult produced within a single explanation according to the modality (or modalities) in which that strategy was produced in the explanation. A strategy could either be produced (a) in both speech and gesture, (b) in gesture but not speech, or (c) in speech but not gesture. In our previous drafts (and elsewhere in the literature), a gesture-speech mismatch only refers to (b)—it is defined as instances where the gesture includes information that speech does not include. In our new analyses, prompted by the reviewers' questions, we explore the ability of relevant and irrelevant strategies in each of these three types of formats (both speech and gesture, gesture only, speech only) to predict posttest performance, and we find that it is relevant strategies produced in gesture only that are the best predictor. Our findings thus replicate the previous literature in that they implicate responses in which gesture conveys information that cannot be found in speech, and they also extend the phenomenon to include a focus not only on form, but also on content. This new way of analyzing the data eliminates the concerns that the reviewers raised about pretest performance and matching gestures (as they are both factors in the analysis), and total number of gestures should not be a potential confound in this new analysis (particularly since almost every explanation was accompanied by gesture, see below). We believe these changes make our manuscript easier to understand, and also allow us to take a more nuanced look at the relationship between speech and gesture. We thank the reviewers for the comments that led to this reorganization.

*3. Please include a table with the descriptive statistics and the correlations of all variables. If the hypothesized relationship between relevant mismatches and post-test performance holds, the bivariate correlation should be substantial and significant.*

**Response:** We have added Table 3 to the result section, which displays the number of each of the six types of strategies (irrelevant in gesture+speech, irrelevant in gesture only, irrelevant in speech only, relevant in gesture+speech, relevant in gesture only, relevant in speech only) produced on the pretest and the posttest. In addition, we have also included Table S.1 (in the supplementary materials), which displays correlations among these six strategy types on the pretest. We report on page 20 in the text that gesture-only relevant strategies were the only strategy type on the pretest that correlated significantly with posttest test performance ($r^2$ = .40, p<.; p's for the other 5 variables were > 0.50).

*4. The sample of participants is very small, which raises concerns about power and biased results (very few cases – e.g., those gesturing a lot – could have a very large influence. Please provide information about the distribution of your variables (count data tend to be skewed a lot, which might require a transformation before using them as predictors in a linear model) and provide a scatterplot of the focal relationship between relevant mismatches and post-test performance. Finally, please conduct regression diagnostics to make sure that the assumption underlying linear models are fully met and report the results of these diagnostics in the manuscript (very good guidelines are provided in Cohen, Cohen, West, & Aiken, 2015, Ch. 4).*

**Response:** Our participants gestured spontaneously and regularly during their explanations: 96% of pretest trials (*n* = 248/258 trials) and 94% of posttest trials (*n* = 246/258 trials) contained at least one gesture. Since almost all explanations contained gesture, strategies produced within an explanation had the potential to be produced in both gesture and speech, in gesture only, or in speech only, making it less likely that participants who gesture more than others are skewing the analysis. If sheer number of gestures were driving the effect, we would expect all of the factors that include gesture to be significant predictors. For example, in Table 4, four factors include gesture. Three—relevant strategies in both gesture and speech, irrelevant strategies in both gesture and speech, and irrelevant strategies in gesture only—were *not* significant predictors of performance. The only factor that predicted performance was relevant strategies in gesture only.

As noted earlier, we now provide information about how often each of the six strategy types was produced on the pretest and posttest (Table 3), as well as correlations among these six strategy types on the pretest in the supplementary materials (Table S1). We also include a scatterplot displaying the average number of gesture-only relevant strategies produced on the pretest in relation to posttest performance (Figure 4), which reveals the linear relation between the two.

The data meet the assumptions necessary to use multiple regression / correlation analyses, with the exception that the variables listed in Table 4 are slightly right-skewed. The data were normalized with a logarithmic transformation and modeled again—we found the same pattern of results with very similar statistical values. We report untransformed data in the paper because it is easier to understand. The outcome variable is bivariate, and so we have used logistic regression to model it (though our Figure 4 does show average performance as the DV).

*5. Please make your data available for inspection to reviewers and be prepared to make your data available in a public repository (such as OSF) should your manuscript be accepted.*

**Response:** We are prepared to share our data as requested by the editor.

*6. The regression model is reported in an odd way not consistent with APA style. Please follow APA style (there is a sample table for reporting regression results in the APA publication manual) and report effect sizes and the scaling of the predictor variables. You have a multilevel structure in your data (trials nested within participants) so it might be a good idea to use a mixed model/multilevel model for analyzing the data.*

**Response:** We now present the full summary of our models (Tables 4 and 5), and we have switched to a mixed model with participant as a random effect.

*7. The manuscript is relatively long, given that the results are a bit thin. Please try to shorten the manuscript by at least 5 pages. Relevant methodological details can be made available in an online supplement (for example, on Open Science Framework).*

**Response:** The manuscript is greatly simplified as a result of the revision, and the text of the manuscript is now 25 pages. We should be able to streamline the introduction a bit if you think the paper is still too long. If so, we would appreciate advice as to which parts of the introductory background we should stress (and not stress) for readers of *Discourse Processes.*

*In their very helpful and constructive reviews, the reviewers note a number of additional (mostly methodological) issues that need to be addressed by a successful revision. In addition, please make sure that your revised manuscript is consistent with APA style in each and every respect.*

**Response:** Please see responses to each of the reviewer's comments below

**Responses to Reviewer #2:**

*The authors address the interesting question of whether adults - similar as children - show by their gesture-speech mismatches, while explaining the solution of a chemistry problem, that they are "ready to learn". A pretest-instruction-posttest design was applied. In fact, mismatches between gesture and speech predicted the positive effect of the instruction, as revealed in the post-test. However, the effect emerged only for those participants who produced relevant information in their gestures in the pre-test, suggesting that gesture-speech mismatches do not always predict performance after learning but only when participants express by their gestures that they already possess some kind of implicit understanding of the problem - even though they are unable to express this explicitly.*

*Overall, the manuscript is well written, and the research question is original as it tells us something about the role of gestures in adults' learning. Therewith, it might be interesting for readers who work in both more foundational as well as applied fields considered with learning. However, before recommending the manuscript for publication, some issues have to be clarified. These include the notation and operationalization of "gesture-speech matches" and "gesture-speech mismatches". In addition, the analyses do not seem to be suited to check for effects in such small sub-samples that markedly differ in sample size. Also, a rational for computing the performance score in a binary manner*

*instead of as continuous variable should be provided. Also, the method section needs some clarification. Details can be found below.*

- ***Method:***

*o The term "gesture-speech matches" seems to be a bit confusing when it is also for trials in which "gesture was uncodeable; and when there was no gesture". How can there be a "match" when there is no second source of information? Relatedly, the term "mismatch" suggests a deviation in one of two directions. However, the authors use this term only for trials in which gesture outperforms speech (not in which speech outperforms gesture).*

*o The difference between gesture-speech matches and mismatches is not fully clear to me. As I understand it, trials were coded as matches when the strategy reflected in gestures was poorer or equal to the strategy reflected in speech. This might theoretically include the definition of the gesture-speech mismatch: "when gesture conveyed a strategy that was not conveyed in speech". I think the point here is that gesture surpasses speech, isn't it? But then, the irrelevant mismatches do not fit in this scheme ("if the strategies produced in gesture highlighted irrelevant components of the problem (only levels 0 or 1)". In addition, it seems to me that gestures were coded on a more fine-grained level than speech: A gesture could involve different levels of explanation (p. 16: "relevant/irrelevant mismatches if the gesture mismatches highlighted both irrelevant and relevant components at least one at levels 2, 3, or 4 combined with at least one at levels 0 or 1 within the same explanation"), while speech was coded only on one level. Similarly, on p. 18, the classification into relevant and irrelevant mismatches is not fully straightforward ("a relevant mismatch, in which the strategy added by gesture was at Levels 2, 3, or 4" - does the authors mean that the strategy expressed in speech was at least on one lower level?).*

**Response:** In response to these very helpful comments, we have completely changed how we approached our data. We think this change has greatly improved the quality of the manuscript and has resulted in a new way to think about the relationship between gesture and speech. As noted in the response to the editor, we now code at the level of the strategy. Every strategy is coded for explanatory relevance (relevant, irrelevant) and modality (produced in gesture+speech, gesture only, speech only). We now predict success after training using these variables in combination (i.e., relevance combined with modality, Table 4) and independently (i.e., relevance on its own, modality on its own, Table 5). We find that it is only the relevant strategies produced in gesture alone that predict success after training.

Following this procedure meant that we did not have to categorize explanations as matching or mismatching, and we did not have to bin individuals according to whether they had produced mismatches. Moreover, the new coding method includes the variable that the reviewer is looking for -- strategies produced in speech but not gesture, as well as strategies produced in gesture and not speech and strategies produced in both gesture and speech. Strategies, both gestured and spoken, that are not codable are not included in the current analyses. This new approach described in detail on page 15..

*o A summary of the design would be useful. It is not clear to me how the three visual representations of Fig. 1 are included in the procedure.*

**Response:** Figure 1 has been simplified. It now shows only the stimulus that was presented to participants.

*o Why was "The molecule with no stereoisomer … presented at the fourth trial at pretest and at the sixth trial at posttest" (and not at random)?*

**Response:** We randomized the order of trials, then used that same order across all participants. Since we have a single sample and a correlational design, we wanted to keep variables as similar as possible across participants.

*o It might be easier for the reader to have an idea of the task before the material is described.*

**Response:** We added an overview of the procedure to page 11.

*o Sample: Please report age. In addition - even though I know that it is usual to report the racial background, this does not seem to be very instructive to me. It might be more interesting to know whether the participants spoke English fluently.*

**Response:** We added on page 9 that the participants were all fluent speakers of English who were the average age of undergraduates.

*o P. 15: Please state how many trials could not be coded.*

**Response:** This information was added to page 14. Speech was codable on 90% of pretest trials (*n* = 231/258 trials) and 94% of posttest trials (*n* = 243/258 trials). Gesture was codable on 69% of the *pretest* explanations that included gesture (*n* = 171/248 trials) and on 72% of the *posttest* explanations that included gesture (*n* = 179/246 trials).

***Results:***

*o Why was the post-test performance scored as frequency of "level 4 explanation in speech, combined with a correct drawing"? As I understand it, performance could be an interval scaled variable.*

**Response:** Performance on each pretest and each posttest item was scored as 0 or 1. A response received a score of 1 when it had both: (a) a correct drawing and (b) a level 4 explanation in speech. We chose this criterion because it leaves no doubt that the individual explicitly understands the problem. However, we did reanalyze the data using the procedure suggested by the reviewer – we determined the highest strategy level expressed in speech in each response on the posttest and assigned the response that value. We then calculated the average score across the set of 6 problems for each participant and used that average as the participant's posttest score. Using this technique for calculating posttest success, we found the same pattern of results as we found with our original criterion. However, the problem with this approach, which seems on the face of it to be the better approach, is that it is not at all clear that the distance between level 1 and level 2 responses is the same as the distance between level 2 and level 3 responses, etc.—in other words, it's not clear that our levels can be treated as an interval (or ratio) scale, which we are implicitly doing when we average the level scores. As a result, we use our original criterion for posttest success in the analyses in the paper and indicate in footnotes 8 and 9 that the patterns hold when we attempt to analyze the data using a continuous scale.

*o P. 19: I do not believe that a parametric test is suited and powerful enough to explore differences between four groups, which differ in sample size between n = 4 and n = 17 (total sample size: N = 43)? The same applies to the analysis on p. 20/21. The*

*level of spoken explanations (p. 19) could perhaps be included as further predictor in the regression model. I would be interested in power analyses given the small sample size.*

**Response:** We agree and no longer divide our participants into groups and thus no longer explore differences between groups. Spoken explanations are included in the new analyses – strategies produced in speech are categorized according to relevance (relevant, irrelevant) and modality (in speech only, in speech and gesture).

*o Why were participants categorized into the four groups (only relevant / only irrelevant mismatches, a combination of both, and matchers), if the mean performance in the single trials served as predictor in the regression analysis?*

**Response:** We no longer categorize participants into groups in our new analyses.

*o Fig. 4 could be improved, it is confusing at the moment (and why is the mean performance so poor - what is the theoretical performance range)?*

**Response:** The old Figure 4 has been eliminated from the revision, as it is no longer relevant given our new analyses.

*o Were there also mismatches between drawings and verbal explanations (or gestures)?*

**Response:** The types of errors participants make in their drawings are not the same types of errors that they make in speech and gesture. As a result, the coding system for drawings does not map directly onto our speech and gesture coding systems, making it difficult to talk about matching and mismatching for drawings and explanations.

*o Did the number of relevant gestures also increase between pre- and posttest (p. 17)?*

**Response:** We now present this information in Table 3, which displays the number of relevant (and irrelevant) strategies produced on the pretest and posttest (the strategies are also categorized according to modality--gesture+speech, gesture only, speech only). The table indicates that, as expected, irrelevant strategies decrease after training and relevant strategies increase, particularly if those strategies are produced in speech.

*- Discussion:*

*o Please start with a short summary of the results before going into detail.*

**Response:** We have added a summary on page 21, and summarize our four central findings. First, our paper demonstrates that novices gesture at a very high rate when they talk about how they have transformed molecules in an attempt to create alternate spatial arrangements of organic compounds. The findings replicate studies showing that tasks involving mental rotation and other visuospatial skills tend to elicit gesture. Second, we developed a reliable coding system for identifying problem-solving strategies in speech and the accompanying spontaneous gestures, thus establishing that our paradigm can be successfully used to study the role of spontaneous gesture in chemistry learning. Third, we found that the window gesture offers onto an individual's understanding of the stereoisomer problem does not always provide the same view as the window offered by speech—gesture, particularly when it is analyzed in relation to speech, can tell us whether an individual is ready to profit from instruction on a stereoisomer task. Finally, we show that the content of a mismatch matters as much as its form in terms of predicting success after a lesson (i.e., relevant strategies predict

success on the posttest but only when those strategies are produced uniquely in gesture).

*o P. 26: What are "correct mismatches"?*
**Response:** We no longer use this terminology.

*- Minor issues:*
*o P. 3: "Traditionally, teachers have used spoken language, written text, and two-dimensional (2D) diagrams to teach chemistry phenomena" - according to my experience, they did use 3D models, too*
**Response:** We have added this point to page 3.

*o Serial order of references in parentheses is inconsistent (alphabetical vs. chronological); p. 19 and further: please report a reasonable number of decimal places*
**Response:** This problem has been fixed. We now use fewer decimal places in the text.

*o What is "an 8x11 piece of paper"?*
**Response:** Page 10 now clarifies that we mean an 8 inch x 11 inch piece of paper.

**Responses to Reviewer #3:**
*Are gestures a meaningful part of the learning process? And do they reveal implicit knowledge? In this study, organic chemistry students learn about stereoisomers - non-superimposable molecules with otherwise identical components and bonds - and then to fully describe whether a set of target molecules are indeed stereoisomers. Participants routinely gesture while describing the molecules and their gestures often reveal that participants understand aspects of the concept that they have not yet verbally articulated (mismatches). Moreover, participants who produce relevant mismatching gestures at pretest (e.g., rotating the hands in front of each molecule and pointing to the stereocenter) show improved accuracy after instruction. From this, the authors conclude that mismatching gestures during learning reveals implicit knowledge.*

*This is an ambitious study and the work is properly contextualized in educational research - mostly in children - demonstrating the pervasive role of gesture in learning, especially about science and mathematics. Moreover, this is not the first study to suggest that gesture-speech mismatches are particularly revealing about emerging knowledge. This is a captivating idea and stereoisomers are a clever way of extending this idea to mature learners. I also think the authors are right, but can we fairly draw these conclusions from this study?*

*How aware are participants that they are required to produce a level 4 explanation in speech in order to be scored as correct? Comparing some utterances between level 2, 3, and 4 looks more like participants have learned how to precisely articulate that two molecules are not superimposable, rather than gained some novel insight into the problem.*
**Response:** It is possible that participants have more knowledge about stereoisomers than we are giving them credit for. However, our goal was to assess when

participants have a full understanding of stereoisomers, reflected in our criterion that they complete a correct drawing and a level 4 explanation. Accordingly, as noted in the response to Reviewer 2, performance on each pretest and each posttest item was scored as 0 or 1 in our original analyses. A response received a score of 1 when it had both: (a) a correct drawing and (b) a level 4 explanation in speech. But we also loosened the criterion and accepted level 3 responses plus a correct drawing as success on the posttest, and found that the same patterns held (as now noted in footnotes 8 and 9). We also attempted to calculate posttest success using a continuous scale (as described in the response to reviewer #2). Using this method for calculating posttest success, we again found the same pattern of results. However, as mentioned earlier, the problem with this approach is that it is not clear that the distance between level 1 and level 2 responses is the same as the distance between level 2 and level 3 responses, etc.—in other words, it's not clear that our levels can be treated as an interval scale, which we are implicitly doing when we average the levels scores. We therefore report results in terms of our original criterion for posttest success in the text of the paper, and indicate in footnotes 8 and 9 that the patterns hold when we use additional criteria for posttest success.

*Further, for a study like this -especially one with such subtle differentiation between different classes of answers - at least two coders should judge the gestures and explanations produced by participants.*

**Response:** Reliability was assessed by having a second transcriber independently code a randomly selected 10% of the 516 pretest and posttest trials. Agreement between the two coders was 91% for speech strategies and 82% for gesture strategies.

*The descriptive statistics (and therefore the inferential statistic based on them) for participants who did not produce relevant mismatches at post-test do not appear to be useful (p. 21 - the last paragraph of the results section). How did these participants perform in line with their peers at pretest and then fail to answer a single question correctly at post-test? Were they confused by the instruction? Did they stop trying to do the task? Did they fail to grasp that only level 4 explanations would be accepted as correct? For a comparison that is so essential to the conclusion that we want to draw from this study, it raises too many questions to be sure of what's really going on.*

**Response:** As noted earlier, we no longer categorize participants into groups, but rather analyze responses at the level of the strategy. The descriptive statistics that the reviewer is describing are therefore no longer relevant. Using our new coding system (and the new statistical models that are needed to analyze them), we find that producing relevant strategies in gesture only is the only factor that reliably predicts success after training.

# Unpacking the gestures of chemistry learners: What the hands tell us about individuals' understanding of stereochemistry

[1]Raedy Ping, [2]R. B. Church, [1]Mary-Anne Decatur, [1]Samuel W. Larson,

[1]Elena Zinchenko & [1]Susan Goldin-Meadow

[1]University of Chicago
[2]Northeastern Illinois University

Address correspondence to:

Susan Goldin-Meadow
University of Chicago
Department of Psychology
5848 S University Avenue
Chicago, IL 60637
Telephone/Fax: 773-702-2585
Email: sgm@uchicago.edu

## Author Note

This work was supported by NICHD R01-HD47450 and NSF BCS-0925595 to SGM and NSF SBE-0541957 (Spatial Intelligence and Learning Center) to Nora Newcombe (PI) and SGM (co-PI). The authors are grateful to Mike Stieff for very helpful comments on a rough draft of the manuscript. We also thank Melissa Herrett, Theodora Koumoutsakis, and Marisha Kazi for help with the formatting of the document. RP is now at Northwestern University, MD at SOAS, University of London, and SWL at Seattle Children's Hospital.

**Abstract**

When learners explain answers to tasks they have yet to master, they often gesture. In a pretest-posttest design, we show that these gestures provide insight into what learners know about a chemistry task, and whether they are ready to make gains on that task. Adults, naïve to organic chemistry, drew stereoisomers of molecules and explained their drawings. All participants gestured spontaneously during explanations, and often expressed strategies only in gesture (and not in speech). In some cases, these strategies conveyed information that was explanatorily *relevant* to the problem; in other cases, they conveyed information that was explanatorily *irrelevant*. Relevant strategies produced *only in gesture* on the pretest, and no other types of pretest strategies, predicted posttest performance. This finding supports the hypothesis that information conveyed uniquely in gesture and not in speech indicates readiness-to-learn, but advances the argument: Gesture-speech mismatch predicts learning not because it reflects general upheaval in learners, but because it reveals explanatorily relevant implicit knowledge that promotes change.

**Keywords:** Gesture, gesture-speech mismatch, embodiment, chemistry, STEM, spatial thinking, individual differences, science education, STEM education, organic chemistry

For undergraduate students, Organic Chemistry (OChem) is the gatekeeper to post-graduate education—medical school, dental school, as well as the STEM disciplines (Science, Technology, Engineering, and Mathematics). Given issues of social inequity within these professional fields, it is critical to develop ways to help more students through the gate. Indeed, underrepresented minorities (UM) experience particular difficulty in OChem, largely because individuals with relatively few economic resources do not have the background needed for the course (Chen, 2013). OChem is difficult for all students in large part because solving chemistry problems relies heavily on complex spatial and dynamic thinking. Organic molecules are complex three-dimensional (3D) forms, intrinsically and extrinsically dynamic, and not visible to the naked eye. Our long-term goal is to create tools to help students develop accurate representations of the complex, three-dimensional and dynamic processes and phenomena common in science. Providing students with such tools has the potential to give them a helping hand with OChem and thus perhaps help to level the playing field.

Traditionally, teachers have used spoken language, written text, and two-dimensional (2D) diagrams to teach chemistry phenomena, but these instructional tools are not particularly well suited to teaching dynamic 3D processes. We argue that the manual gestures teachers and students often produce when discussing math and science are a particularly good vehicle for capturing dynamic spatial information (Alibali & Nathan, 2012; Alibali, Nathan, Fujimori, Stein, & Raudenbush, 2011; Hostetter & Alibali, 2008), in part because of the affordances of the hands and because gesture is not constrained by grammar, vocabulary, or the categorical nature of words (e.g., McNeill, 1992). Unlike 3 dimensional models, gestures are free of cost, are always available, and can easily be transferred from one learning situation to the next. From experts to novices, people spontaneously and effortlessly gesture when they talk about science (Atit,

Gagnier & Shipley, 2014; Crowder & Newman, 1993; Singer, Radinsky & Goldman, 2008). Compared to traditional tools like spoken language, written text and diagrams, gestures offer distinct advantages for the many students who struggle with OChem simply because talk about the spatial and dynamic processes involved in chemistry is likely to be accompanied by iconic gestures that convey 3-D, dynamic information. Examining the gestures that students produce when talking about chemistry could give instructors insight into the struggles their students are having with OChem. These insights could, in turn, help instructors devise lessons that can address their students' misunderstandings, particularly for UM students interested in pursuing medical and STEM careers (Rueckert, Church, Avila & Trejo, 2017).

We begin by reviewing learning sciences research on gesture in real-world scientific inquiry, and then turn to developmental psychology research on gesture, cognition, and learning. We draw upon both traditions to motivate our hypotheses about the role gesture plays in chemistry learning.

**Speech and Gesture Establish Common Ground in Collaborative Real-World Science Inquiry**

Learning sciences researchers have focused on gesture's role in communities where the practice is to represent and communicate about scientific objects and phenomena *in situ* (e.g., Becvar, Hollan, & Hutchins, 2005; Goodwin, 2000, 2007, 2010; Lave, 1991). Gestures are transcribed from videotapes of scientists and/or students interacting in the lab, field, or classroom, and are analyzed in the context of the accompanying dialogue. For example, Trafton and colleagues (Trafton, Trickett, Stitzlein, Saner, & Schunn, 2006) observed meteorologists and neuroscientists at work, and asked them periodically to talk out loud about what they were doing. The scientists produced more iconic gesture when they used spatial language than non-spatial

language, and more iconic gesture when they talked about dynamic processes than static objects. As another example, Becvar et al. (2005) traced the development of shared understanding about a particular protein (thrombin) over weeks. The lab leader produced an idiosyncratic spontaneous gesture for thrombin and other lab members adopted this gesture. Over time, this so-called "thrombin hand" became a conventionalized form that every lab member used to demonstrate processes involving thrombin. Gesture can be an integral part of coming to a shared understanding on the fly.

But gestures rarely become conventionalized forms, and are almost never verbally named ('thrombin hand' is an exception). Instead, they are produced anew by each speaker on each occasion. Radinsky, Ping, Hospelhorn, & Goldman (2012) found that even 6th grade students use one another's spontaneous gestures in science learning situations to improve their own understanding of tectonic plate movement. Although students did not realize that they were depending on one another's gestures, they did so rapidly, responding to others' gestures with their own gestures (and speech). Undergraduate OChem students and instructors may also be able to use gesture to come to a shared understanding, making the study of gesture potentially informative about chemistry learning. In this study, we focus on an individual speaker's gestures and ask what the information conveyed uniquely in gesture can tell us about a student's understanding of chemistry.

**Gesture-Speech Mismatches Reflect Knowledge in Transition in Child Learners**

Church and Goldin-Meadow (1986) asked 5- to 8-year-old children to explain how they solved Piagetian (1952) conservation problems. On each problem, the child was shown two identical tall, thin glasses and was asked to confirm that the amount of water in the two glasses was the same. While the child watched, the experimenter then poured the water from one of the

glasses into a short, wide bowl. Non-conserving children believe that the amount of water changed when it was poured, and typically justify this belief as follows: "The glass has more water because it's taller than the dish." The gesture equivalent of the height comparison strategy is a series of flat palm or point gestures that indicate the (higher) level of water in the glass and the (lower) level of water in the bowl.

Children usually express the same strategy in both speech and gesture, as in the height example just given—but not always. Consider a child whose speech conveys height information, but whose gesture conveys information about the widths of the containers—a C-shape mirroring the circumference of each container, combined with speech focusing on the heights of the containers. In this case, gesture is adding unique information to the spoken response (here, information about a second dimension, which is not mentioned in speech). Researchers have found that children who produce these so-called "gesture-speech mismatches" when explaining their solutions to the conservation task are particularly likely to profit from instruction in conservation (Church & Goldin-Meadow, 1986; Church et al., 2004). This phenomenon has been replicated in child learners tackling a variety of STEM-related tasks (Goldin-Meadow, Alibali, & Church, 1993; Perry et al., 1988; Pine, Lufkin, & Messer, 2004).

One reason that gesture-speech mismatches are related to learning may be because gesture indicates nascent understanding of how to solve a particular problem. For example, a child who talks about the height of the containers in speech while gesturing about their width is starting to attend to the width dimension—an important awareness underlying understanding that although the water changed in height, it also changed in width. Attending to both the height and width dimensions is one requirement for understanding conservation of liquid quantity after transformation. Gesture-speech mismatch can therefore be used to glean how a child mentally

represents a particular problem, and to provide a window onto thinking not readily afforded by the child's problem solutions or speech strategies for solving the problem.

**Do Gesture-Speech Mismatches Reflect Adult Learners' Knowledge?**

Adults also produce gesture-speech mismatches when talking about concrete and abstract ideas: reasoning about moral dilemmas (Church, Schonert-Reichl, Goodman, Kelly, & Ayman-Nolley, 1995); describing pictures of landscapes, abstract art, buildings, people, and machines (Morrel-Samuels & Krauss, 1992); narrating cartoon stories (Beattie & Shovelton, 1999; McNeill, 1992; Rauscher, Krauss, & Chen, 1996); explaining solutions to the Tower of Hanoi puzzle (Garber & Goldin-Meadow, 2002); explaining how gears work (Perry & Elder, 1997); and describing solutions to algebra problems involving continuous and discrete change (Alibali, Bassok, Olseth, Syc, & Goldin-Meadow, 1999). Gesture-speech mismatch is not just an outgrowth of being a developing child or a child learner. But does gesture play the same role in adult learners as it does in child learners? More specifically, is gesture-speech mismatch associated with readiness-to-learn in adult learners? We address this question by examining the gestures that adults produce when asked how they solved problems in stereochemistry, and exploring whether those gestures predict learning after a brief lesson.

Adult experts and novices produce iconic gestures during science problem solving, particularly when working on organic chemistry problems that rely heavily on spatial and dynamic reasoning (Stieff, 2007; Stieff, 2011; Stieff & Raje, 2010). One such concept is stereochemistry. Stereoisomers are chiral objects[1], not superimposable in real space because they are mirror images of one another across some plane. This geometric notion is often demonstrated by comparing the finger structure of the left and right hands—placing the left hand (palm down) directly on top of the right hand (also palm down), or vice versa, reveals that the two are *non-*

[1] Not all stereoisomers are chiral, but in our study we used chiral and chiral-looking molecules.

*superimposable* spatial arrangements of the same fingers. Structurally, the left hand is a mirror image of the right hand, and neither hand is symmetric (i.e., all 5 fingers are different and cannot be superimposed). Like the fingers of a hand, the substituents of stereoisomers lack an internal plane or point of symmetry—they contain a central atom called a stereocenter, from which extend four different substituent groups in a tetrahedron. Conversely, all parts of *superimposable* molecules can be perfectly lined up with one another—as if the fingers of the left hand were supplanted with those of the right hand. Since molecules are three-dimensional in a way that the hand analogy is not, if any two of the substituent groups in the tetrahedral configuration are identical, the molecule is symmetric and therefore would have a superimposable mate. In other words, it would *not* have a stereoisomer.

We asked adult participants to solve a set of stereoisomer problems and explain their answers. We found that all of our participants gestured when explaining their solutions. Moreover, overall, the gestures they produced conveyed the same types of problem-solving strategies that they conveyed in speech. We coded the problem-solving strategies participants produced in speech and in gesture and, for each explanation, determined whether the information conveyed in gesture was the same as, or different from, the information conveyed in speech. We replicated the logic and procedures of the child gesture-speech mismatch studies by then giving each adult a brief lesson about stereoisomers, and determining whether the relation between gesture and speech predicted post-lesson performance. We were therefore able to ask *whether* gesture-speech mismatch predicts readiness to learn in adult learners as it does in child learners.

The current work also allows us to explore *why* gesture-speech mismatch predicts openness to instruction. One possibility focuses on form—gesture-speech mismatch may index readiness-to-learn simply because gesture conveys a different strategy from speech; that is,

because information is conveyed across two modalities, regardless of the content of that information. But content may matter. A second possibility is that gesture-speech mismatch predicts learning not only because of its form, but also because of its content. The information that child learners convey in gesture and/or speech, although not always correct, has always been *explanatorily relevant* to the problem in the tasks studied thus far. In contrast, the adults in our study often conveyed information that was *not explanatorily relevant* to the stereoisomer problem, reflecting fundamental misconceptions of the task. Our adult data thus allow us to ask whether the explanatory relevance of the strategies produced in a gesture-speech mismatch prior to a lesson plays a role in predicting performance after the lesson; in other words, whether the content of a mismatch (its explanatory relevance to the problem), as well as its form (the modality in which this information is conveyed), both play a role in predicting learning.

## Method

### Participants

Fifty-two adults (54% female) participated in the study; 51% of participants were Caucasian, 17% Asian American, and 15% African American; 15% of participants identified as Latino/a, 6% identified as biracial, 6% responded other, and 4% opted not to report race. All participants were fluent speakers of English and were 18-22 years of age. Participants were compensated for their time with either course credit or a small monetary reimbursement. Participants were undergraduates recruited through a list-serve of psychology study volunteers. We prescreened participants for their level of chemistry education, and selected volunteers who had at least one year of formal chemistry education at the high school or undergraduate level, with no formal instruction in organic chemistry. Five participants (4 female) were excluded from

analyses because of experimenter error or non-compliance with instructions.[2] Four additional participants were screened out based on pretest performance (3 were unable to give any responses that followed the laws of chemistry and 1 mastered the task at pretest).[3] We present data from 43 participants.

## Materials

### Pretest and posttest molecules.

Figure 1 displays an example of a stimulus given to participants in our study. Participants were provided with two different visual representations of molecular structure and spatial arrangement on an 8 inch x11 inch piece of paper: a labeled color illustration of a 3-dimensional ball and stick representation of the molecule's structure (Figure 1, top), and a wedge and dash representation, which shows the molecule's skeletal formula in 2 dimensions (Figure 1, bottom). Participants were told that the dark colored triangles (wedges) indicate parts of the molecule coming out of the page in space towards the viewer, and that the light colored triangles (dashes) indicate parts going into the page in space away from the viewer. In this example, the chlorine (Cl) atom is coming out of the page towards the viewer, and the lone hydrogen (H) atom is going back into the page away from the viewer. The wedge and dash representation indicates single bonds as single lines between atoms and double bonds as double lines between atoms.[4]

Six unique molecules, one with no stereoisomers of any kind, were presented in a randomized order fixed across participants at pretest. Each time a molecule was presented counted as a new trial. A second set of six unique molecules, including one molecule with no

[2] For example, the experimenter forgot to turn on the video camera. In another case, the experimenter did not stop a participant from holding the marker during the explanations, which inhibited gesture.
[3] Some participants were not able to produce drawings with the same molecular formula and bonding order as given in the stimulus molecules.
[4] There are no double bonds in the molecule in the figure, but some molecules used in the study did contain double bonds.

stereoisomers, was presented to each participant in a posttest after a brief lesson. Following the random order used with all participants, the molecule with no stereoisomer was presented at the fourth trial at pretest and at the sixth trial at posttest.

**Training lesson.**

Scripted verbal training was given to each participant after the pretest trials. The experimenter laid out a 3-step procedure for determining whether each molecule has a stereoisomer for two example molecules, one with a stereoisomer and one without.[5] The experimenter first indicated the central carbon and the substituents. Participants were then shown how changing the spatial orientation of two of the substituents along the Z-axis could create a unique spatial representation of the original molecule. Participants were finally told to determine whether the new spatial arrangement actually created a non-superimposable stereoisomer of the original molecule by rotating the entire molecule in three dimensions and checking the non-manipulated substituents. If the non-manipulated substituents did not match, it was non-superimposable and therefore a stereoisomer of the original; if they did match, it was superimposable and therefore *not* a stereoisomer.

**Procedure**

Individuals participated in a single pretest-instruction-posttest session. Each hour-long session was videotaped with the participants' knowledge. When participants arrived, they were given a general description of the problems they would solve, and were shown the wedge and dash conventions for drawing molecules. They were then told about the importance of stereoisomers, using Thalidomide as an example, and stereoisomers were defined as "molecules with multiple non-superimposable spatial arrangements." The experimenter said that

[5] The molecule with a stereoisomer had an enantiomer, a diastereomer, and an enantiomer of the diastereomer.

stereoisomers are "molecules that have the same molecular formula, as well as the same bond order and connectivity, but different orientations in three-dimensional space." Participants were told, "no matter how stereoisomers are rotated in space, their parts are never perfectly superimposable on one another."[6]

The pretest consisted of six trials. On each, the participant was presented with a prompt (see Figure 1) and asked to determine whether the molecule has a stereoisomer and, if so, to draw a wedge and dash depiction of it on a white board.[6] Participants were then asked to explain how they created their drawing of the stereoisomer. For trials on which participants judged that the molecule does not have a stereoisomer, they were asked to justify why no possible non-superimposable spatial arrangement of the original molecule exists. Participants did not receive feedback on their drawings or their explanations.

After the pretest, each participant was given the brief training lesson described earlier. They then completed the posttest, following the same procedure as the pretest. Once the posttest was completed, the participant was debriefed and dismissed.

**Coding**

Participants' pretest and posttest drawings were coded in one pass through the data, and their explanations were coded in two additional independent passes.

---

[6] Although there are different types of stereoisomers, in the current study, we limited our prompts and instruction materials to molecules with mirror-image configurational enantiomers and, as a contrast, molecules with symmetry that do not exist in enantiomer form. We will use the terminology 'stereoisomer' and 'stereochemistry' throughout. Some of the molecules used as stimuli exist in diastereomer form as well. The same strategies can be used to create a stereoisomer for diastereomers.

[7] During each trial, participants were allowed to revise their drawing as many times as they desired, but were then asked to restart their explanation from the beginning each time they did. Only the final drawing and explanation were coded.

**Scoring drawings.**

A drawing was scored as either correct or incorrect. A correct drawing completely illustrated a possible stereoisomer of the molecule. For the two prompt molecules without stereoisomers, a response was considered correct if the participant did not produce a drawing and instead stated that the molecule lacked a stereoisomer. A second coder classified 10% of the 516 pretest and posttest drawings for correctness. Agreement between the two coders was 98%.

**Developing speech and gesture coding systems.**

Before beginning the study, we videotaped approximately a dozen individuals with varying levels of familiarity with stereochemistry, from psychology graduate students with little to no chemistry education to organic chemistry professors. We asked them to solve different types of stereoisomer problems and to explain their responses in depth. From this corpus, two coders (expert in coding gesture and speech and well-versed in stereochemistry) identified common problem-solving strategies. The coders then conducted a series of intensive working meetings with expert gesture coders and chemists to fine-tune the coding system.

Once this system was in place, one main coder took two separate passes through the pretest and posttest explanations. Each trial was counted as a single explanation, and the speech and gesture produced in that explanation were each coded for type of strategy. On the first pass, the coder transcribed and coded speech for type of strategy, without attending to gesture, that is, with the video turned off; the video was consulted only when a referent of the speech could not be interpreted without the pointing gesture (e.g., "I moved *this part*" or "I switched *these two groups*"). On the second pass, the coder transcribed and coded gesture for type of strategy, without attending to speech, that is, with the sound turned off. Reliability was assessed by having a second transcriber independently code a randomly selected 10% of the 516 pretest and posttest

trials. Agreement between the two coders was 91% for speech strategies and 82% for gesture strategies.

Speech could not be coded on trials where participants did not give a contentful spoken explanation (e.g., "I am not really sure how I got this but I think it's right"; "I just did what I did last time") and, more commonly, on trials where participants repeated the definition of a stereoisomer (e.g., "my drawing is not superimposable on the original molecule"). The experimenter gave a "why?" prompt in response to both kinds of answers; responses were classified as not codable if participants did not respond to the prompts. Speech was codable on 90% of pretest trials ($n$ = 231/258 trials) and 94% of posttest trials ($n$ = 243/258 trials).

Gesture could not be coded when participants produced: (1) only *beat* gestures (rhythmic, baton-like up and down movements of the hand and arm that sync with speech, but have no representational quality; McNeill, 1992); (2) when participants produced gestures that were too far away from the drawing to determine the appropriate strategy; or (3) when participants produced a series of individual disparate points. Gesture was codable on 69% of the pretest explanations that included gesture ($n$ = 171/248 trials) and on 72% of the posttest explanations that included gesture ($n$ = 179/246 trials).

**Coding speech and gesture for each stereoisomer problem.**

Our hypotheses concern both the content and the form of information conveyed in an explanation. First, we describe how we coded *content*. Table 1 presents the problem-solving strategies, with examples in speech and gesture. Some strategies highlight dimensions that are explanatorily *relevant* to the stereoisomer task (we call these *Relevant Dimension Strategies;* levels 2, 3 and 4; see Table 1b) and some highlight dimensions that are explanatorily *irrelevant* to the stereoisomer problem or ignore an important dimension altogether (we call these

*Irrelevant Dimension Strategies*; levels 0 and 1; note that these responses do pertain to the problem, that is, they are not off-topic responses, but they highlight dimensions that do not help to solve the problem (see Table 1a).

We next describe how we coded *form*. We were unable to code form in the adults' explanations in the same way that it has been coded in children's explanations simply because the adults frequently produced explanations that contained more than one strategy in speech. One of these spoken strategies might match the gestured strategy in the explanation, but the other might mismatch the gestured strategy. This explanation could therefore not be categorized as *either* a match *or* a mismatch. In contrast, children's explanations (at least those produced on the tasks that have been studied to date) tend to contain one strategy in speech and one in gesture. The two strategies either convey the same information or different information; each explanation is consequently either a match or a mismatch—it is never both.

Because many explanations in our adult data could not be classified as either a match or a mismatch, we analyzed the data at the level of the *strategy*. We classified each strategy that an adult produced within a single explanation according to the modality (or modalities) in which that strategy was produced in the explanation. A strategy could either be produced (a) in both speech and gesture, (b) in speech but not gesture, or (c) in gesture but not speech. For example, the explanation in Figure 2 contains 3 relevant strategies, and we made a form coding decision about each one: (1) Relevant Switch was produced in both speech and gesture; (2) Compare the Non-Manipulated Substituents was produced in both speech and gesture; and (3) Relevant Rotation was produced in both speech and gesture. As another example, the explanation in Figure 3 contains 2 relevant strategies: (1) Relevant Switch was produced in both speech and gesture; (2) Relevant Rotation was produced in gesture but not speech.

This coding procedure allowed us to investigate whether the *content* of the information, the *form* in which this information was expressed, or the relationship between these two factors, is the best predictor of posttest performance.

**Criteria for success on the pretest and posttest stereoisomer problems.**

Performance on each pretest and each posttest item was scored as 0 or 1. A response received a score of 1 when it had both: (a) a correct drawing and (b) a level 4 explanation in speech. We chose these criteria because they leave no doubt that the individual explicitly understands the problem.[7]

## Results

We first present descriptive information about how often each strategy was expressed in gesture and/or speech. We then ask whether posttest performance can be predicted by two factors: The *content* (i.e., explanatory relevance) of the strategies conveyed in an explanation (with two levels; relevant, irrelevant), and the *form* (i.e., modality) in which strategies were conveyed (with three levels; a strategy is conveyed in both speech and gesture, gesture only, speech only).

[7] It is possible that participants have more knowledge about stereoisomers than we give them credit for using this criterion. We therefore loosened our criterion and accepted level 3 responses plus a correct drawing as success on the posttest; we found that the patterns reported in the text hold for this relaxed criterion (see footnotes 8 and 9). We also attempted to use a more continuous variable to assess posttest success—we determined the highest strategy level expressed in speech on each posttest response and assigned the response that value. We then calculated the average score across the set of 6 posttest problems for each participant, and used that average as the participant's posttest score. Using this technique, we again found the same patterns (see footnotes 8 and 9). The problem with this approach, however, is that the distance between level 1 and level 2 responses may not be the same as the distance between level 2 and level 3 responses, etc.—in other words, it is not clear that our levels can be averaged as though they were on an interval scale. We therefore use the 'level 4 + correct drawing' criterion for posttest success when presenting data in the text.

**Which Stereochemistry Strategies Do Participants Express in Gesture and in Speech?**

All participants gestured when explaining at least one problem. They produced gesture on 96% of pretest trials ($n$ = 248/258 trials) and 94% of posttest trials ($n$ = 246/258 trials). They produced speech on all pretest and posttest trials. Moreover, participants often produced more than one strategy within a single explanation: 42 of the 43 participants produced at least one explanation at pretest that included more than one strategy (either in the same modality or in two different modalities), and produced an average of 1.71 strategies, SD=0.54, range: 0.50 to 2.67, per explanation. Dividing the strategies according to modality, we found that participants produced an average of 1.29 (SD = 0.43, range: 0.33 to 2.00) strategies in speech per pretest explanation, and 0.98 (SD=0.55, range: 0.17 to 2.33) strategies in gesture per pretest explanation.

Table 2 lists the 5 types of (2a) explanatorily irrelevant strategies and the 4 types of (2b) explanatorily relevant strategies, and indicates the number of explanations in which each strategy was conveyed in gesture and speech, in speech only, or in gesture only on the pretest. Note that both explanatorily irrelevant and relevant strategies not only were produced in gesture and speech, but were also produced uniquely in one of the two modalities.

Table 3 displays the mean number of relevant and irrelevant strategies in gesture+speech, speech only, and gesture only produced on the pretest and the posttest (see table S1 in supplementary materials for correlations among these 6 strategy types on the pretest). We used a within-subjects $t$-test to compare number of strategies at the two time points; the $t$-statistic and $p$-value are displayed in the table. The number of *irrelevant* strategies produced in gesture+speech and the number of irrelevant strategies produced only in speech only *decreased* significantly from pretest to posttest. Likewise, the number of *relevant* strategies produced in gesture+speech and the number of relevant strategies produced only in speech *increased* significantly. In

contrast, the number of strategies (either irrelevant or relevant) produced only in gesture did not change appreciably after the brief lesson.

**Do Strategy Relevance and Modality Predict Posttest Performance on the Stereoisomer Task?**

Only 4 participants produced a correct response (i.e., a correct drawing and a level 4 strategy in speech) on the pretest, and each of these participants produced only one correct response. Performance improved after the lesson, but the success rate was not high—9 participants produced at least one correct response on the posttest, and the mean number of correct responses these participants produced was 1.63 (SD = 0.74, range: 1 to 3). Controlling for pretest success, we asked whether we could use the explanations participants produced on the pretest to predict successful performance on the posttest.

In the child studies, when a participant produced a gesture-speech mismatch, the strategies conveyed in the speech and gestural components of that mismatch were always relevant to solving the problem (i.e., analogous to a level 2 or higher strategy in our data), although the strategies need not have led to a correct solution. In contrast, we found that the stereochemistry task elicited a number of strategies that were on-topic but were completely irrelevant to solving the problem (levels 0 and 1). The presence of explanatorily irrelevant responses in our participants' repertoires allows us to explore whether mismatch between gesture and speech *per se* predicts learning, or whether the explanatory relevance of the strategy matters.

We used a mixed logit model to predict posttest success on each stereoisomer problem using three factors: pretest score, strategy relevance on pretest explanations, and strategy modality on pretest explanations. For each explanation, we tallied the number of relevant strategies produced in both modalities, in gesture only, or in speech only; we also tallied the

number of irrelevant strategies produced in both modalities, in gesture only, or in speech only. We then summed those tallies across explanations for each participant and used those sums to predict posttest performance. Table 4 displays the model summary for this analysis. As expected, pretest score predicted posttest performance (*Beta*=18.69; *SE*=8.62; *t-value*=2.17; *p-value*= 0.03). The only other significant predictor of posttest performance was relevant strategies expressed in gesture only (*Beta*=4.38; *SE*=1.70; *t-value*=2.58; *p-value*= 0.01). None of the other strategy by modality predictors was significant.[8]

We repeated this analysis examining strategy relevance and modality as independent factors to determine whether either (or both) of these factors on their own would predict posttest performance. For each explanation, we tallied the number of relevant and irrelevant strategies produced, ignoring modality. We also tallied the number of strategies produced in both gesture and speech, in gesture only, or in speech only, ignoring relevance. We then summed those tallies across explanations for each participant and used those sums to predict posttest performance. Table 5 displays the model summary for this analysis. We found that producing relevant strategies in any modality (i.e., ignoring modality) did not predict posttest performance (*estimate*=0.92; *SE*=0.83; *t-value*=1.10; *p-value*= 0.27), nor did producing any strategy in gesture only (i.e., ignoring relevance) (*Beta*=2.01; *SE*=1.23; *t-value*=1.64; *p-value*= 0.10). This finding highlights the importance of taking both characteristics of the pretest explanations (both content and form) into account—the explanatory *relevance* of a strategy conveyed on the pretest

[8] When we loosen our criterion for success on the posttest to include level 3 responses + a correct drawing, we find that here too the only significant predictor of posttest performance (besides pretest score) was relevant strategies expressed in gesture only (*Beta*=4.45; *SE*=1.75; *t-value*=2.55; *p-value*= 0.01). Similarly, when we use our strategy levels to construct a continuous measure of posttest success, we again find that the only significant predictor of posttest performance (besides pretest score) was relevant strategies expressed in gesture only (*Beta*=0.89; *SE*=0.34; *t-value*=2.60; *p-value*= 0.01).

predicted posttest performance, but only when that relevant strategy was conveyed *uniquely in gesture*.

To better understand the relation between our two significant factors (pretest score and gesture-only relevant strategies), we conducted a stepwise analysis of the variables in Table 4. In the first step of both models, we made individual mixed logit models for each of the 6 variables (with participant as a random effect). The only significant predictor was the average number of relevant strategies in gesture only (*Beta* = 4.117, *SE* = 1.797, *z-value* = 2.291, *p-value* = 0.022). All of the other predictors had *p-values* > 0.50, except for pretest score (*Beta* = 11.334, *SE* = 7.800, *z-value* = 1.453, *p-value* = 0.146).

We then added each of the variables in Table 4 to the model containing gesture-only relevant strategies. The only factor that was even marginally significant was pretest score (*Beta* = 10.413, *SE* = 6.072, *z-value* = 1.715, *p-value* = 0.0864), and gesture-only relevant strategies remained a significant predictor (*Beta* = 3.823, *SE* = 1.578, *z-value* = 2.423, *p-value* = 0.0154). Although the fit of the model improved when we added pretest score, it did not reach a significant threshold (residual deviance = 95.392 for the model with gesture-only relevant strategies [AIC = 101.39] vs. residual deviance = 92.838 for the model with gesture-only relevant strategies *and* pretest score [AIC = 100.84], $\chi_2(1) = 2.554$, $p = 0.11$). In contrast, when we began with pretest score and added gesture-only relevant strategies to the model, we found a significant improvement in the fit of the model (residual deviance = 99.269 for the model with pretest score [AIC = 105.27] vs. residual deviance = 92.838 for the model with pretest score *and* gesture-only relevant strategies [AIC = 100.84], $\chi_2(1) = 6.4314$, $p = 0.0112$).

Consistent with this pattern, we found that gesture-only relevant strategies were the only strategy type on the pretest that correlated significantly with posttest test performance (Pearson r

= .40, $p<0.007$; p's for the other 5 variables were > 0.50).[9] Figure 4 displays a scatter plot of the number of gesture-only relevant strategies produced at pretest in relation to score at posttest, and illustrates the linear relation between these two factors. Producing relevant strategies uniquely in gesture prior to a brief lesson in stereoisomers thus appears to be the best predictor of the ability to take advantage of the lesson and improve on the posttest.

## Discussion

Our study has four central findings. First, we demonstrated that novices gestured at a very high rate when they talked about how they have transformed molecules in an attempt to create alternate spatial arrangements of organic compounds. Our findings thus replicate studies showing that tasks involving mental rotation and other visuospatial skills tend to elicit gesture (Chu & Kita, 2011; Ehrlich, Levine, & Goldin-Meadow, 2006; Stieff, 2007, 2011; Stieff & Raje, 2010; Trafton et al., 2006).

Second, we developed a reliable coding system for identifying problem-solving strategies in speech and the accompanying spontaneous gestures. We have thus established that our paradigm offers promise for studying the role of spontaneous gesture in chemistry learning.

Third, we found that the window gesture offers onto an individual's understanding of the stereoisomer problem does not always provide the same view as the window offered by speech—gesture can tell us more about what the individual may, or may not, know. Previous work has found that children on the brink of conceptual change often produce gestures-speech mismatches, defined as information conveyed in gesture that is not found in speech (Alibali &

[9] When we loosen our criterion for success on the posttest to include level 3 responses + a correct drawing, we find that here too gesture-only relevant strategies were the only strategy type on the pretest that correlated significantly with posttest test performance (Pearson r = .48, $p<0.001$; p's for the other 5 variables were > 0.50). Similarly, when we use our strategy levels to construct a continuous measure of posttest success, we again find that gesture-only relevant strategies were the only strategy type on the pretest that correlated significantly with posttest test performance (Pearson r = .40, $p<0.008$; p's for the other 5 variables were > 0.50).

Goldin-Meadow, 1993; Church & Goldin-Meadow, 1986; Perry et al., 1988; Pine et al., 2004; for review, see Goldin-Meadow, 2003)—in the terms we are using here, information conveyed in gesture only. Our study replicates this effect in adults.

Finally, our study allowed us to take one step beyond the child studies to investigate the aspects of gesture-speech mismatch that make it a good index of readiness to learn—in particular, whether the form of a mismatch (the fact that it contains two modalities) and/or the content of the mismatch (the fact that it conveys information relevant to solving the problem) predicts learning. Our findings make it clear that *both* form and content matter. We elaborate on this final point below.

In previous work, children dealt with relatively simple tasks (e.g., Piagetian conservation tasks in 5- to 8-year-olds, mathematical equivalence tasks in 9- to 10-year-olds). The incorrect information that the children conveyed in gesture only on these tasks, although leading to incorrect solutions, always contained information that was relevant to the task and was partially correct (i.e., information that was comparable to levels 2+ in our study). For example, on the mathematical equivalence problem 4+6+3=__+3 used by Perry and colleagues (1988), children often conveyed an *add-all-numbers* strategy in gesture (pointing at the 4, the 6, the left 3, and then the right 3), while producing an *add-to-equals-sign* strategy in speech (saying "I added the 4, the 6, and the 3"). The *add-all-numbers* strategy leads to an incorrect solution (i.e., 16 rather than 10) but note that, in order to solve the problem correctly, children need to notice the addend on the right side of the equation. Children who produce the *add-all-numbers* strategy have acquired this relevant knowledge and are displaying it uniquely in gesture. The bottom line is that children did not produce strategies in gesture as misguided as the irrelevant levels 0 and 1 strategies that the adults in our study produced.

Our findings can therefore shed light on *why* gesture-speech mismatch is an index of readiness to learn. Producing information in gesture that is not in speech could index who is ready to learn for at least two reasons. The first is that having several strategies activated at the same time, with some strategies expressed in speech and others in gesture, reflects a type of general cognitive instability in the learner. During such times of instability, learners may be more amenable to instruction in new strategies, possibly because they have become aware of inconsistencies in their thinking and are working to resolve those inconsistencies (Goldin-Meadow & Alibali, 2002; Goldin-Meadow et al., 1993). Variability in strategy choice, either within or across modalities and regardless of whether the strategies are correct or incorrect, has been shown to precede learning (see Siegler, 2007, for review; but see Church, 1999). Under this hypothesis, what is important in propelling change is only that multiple strategies be simultaneously activated, which puts the learner in an inconsistent and unstable state. In other words, the relevance of the strategies should not matter.

The second possibility is that gestured information represents implicit knowledge that the learner holds but is, as yet, unable to explicitly express in speech (Broaders, Cook, Mitchell, & Goldin-Meadow, 2007). This possibility suggests that the type of information embodied in the gestural component of a mismatch is an important index of whether the learner is open to change. When explanatorily irrelevant strategies are expressed in gesture only, gesture reflects implicit ideas that the learner holds, but those ideas are hindering the ability to improve on the task—or at least they are not helping. In this account, conveying information in gesture that is not found in speech is a good index of openness to instruction *only if* the information conveyed in gesture is explanatorily relevant to, and a step on the road to, solving the problem correctly.

Our data align with this second account, providing support for the view that the type of

information gesture adds to speech in a mismatch is telling. Gesture-speech mismatch likely is a signal of general cognitive instability and thus readiness for change. But our results suggest that the explanatory relevance of the information conveyed in gesture may indicate whether change for the better is likely to come easily to a given individual. Our findings suggest that mismatch on its own is not enough—the mismatches learners produce must convey information that is *explanatorily relevant* to solving the problem in order for them to profit from input and change for the better.

Of course, it is possible that findings from this work apply only to problems involving spatial arrangements of molecules, or to problems of molecule transformation, or even to the specific type of enantiomer task we use in this paradigm. Solving the task in our study seemed to rely on spatial processes—the participants produced iconic gesture at very high rates, replicating findings from the chemistry education literature showing that novices (and experts, when problems are difficult) approach problems of this sort spatially (Stieff, 2011; Stieff & Raje, 2010). There is a strong possibility that the effects of gesture on learning are more pronounced for problems that have a spatial, or specifically rotation, component (Chu & Kita, 2011; Cook & Tanenhaus, 2009; Hostetter & Alibali, 2010). However, our data also are consistent with a general gesture-speech mismatch phenomenon, established in the literature across a variety of content areas, including moral reasoning problems with no inherent spatial properties (Beaudoin-Ryan & Goldin-Meadow, 2014). The regularity of this effect across studies helps to strengthen the argument that gesture is a powerful lens through which to study learning more generally.

Our study also underscores the role that gesture can play in the classroom, offering a second window onto a student's understanding of stereoisomer problems. Importantly, extensive training is not necessary for teachers to understand their students' gestures. Ordinary listeners are

able to decode the information that speakers convey uniquely in gesture (Goldin-Meadow & Sandhofer, 1999), often without conscious awareness of gesture as the source of the information (McNeill, Cassell, & McCullough, 1994). Listeners decode iconic information in gesture online, as it occurs, and incorporate the information into their model of the spoken message, as evidenced by reaction time (Ping, Goldin-Meadow & Beilock, 2013). Teachers and undergraduate research participants are sensitive to gesture when identifying which problem-solving strategies are in a child's explanations of his answers to math and conservation problems after a brief primer on coding categories (Kelly, Singer, Hicks, & Goldin-Meadow, 2002) and even without the primer (Alibali, Flevares, & Goldin-Meadow, 1997; Goldin-Meadow, Wein, & Chang, 1992). Students' spontaneous gestures are thus accessible to teachers even if the teacher has not been trained in gesture coding.

One limitation of our study is that improvement from pretest to posttest was slight, which may not be surprising given that the task was complex and the participants received only a few minutes of training in which they simply compared images of molecules. Our intervention was not a comprehensive lesson in stereochemistry, but was instead brief training designed to be a small push in the right direction—a tool for studying whether gesture can provide insight into an adult learner's understanding of a scientific concept. In future work, our goal will be to develop a richer training program that will allow us to explore whether gesturing not only reflects adults' understanding of the stereoisomer problem (as we have shown here), but also plays a role in changing that understanding.

In sum, we have found that gesture can be used as a reliable index of an individual's understanding of stereoisomer problems—the gestures that people produce when they explain their responses to our task reflect the knowledge that they have about the task and, importantly,

reveal knowledge that often goes beyond the knowledge they convey in their speech. We have also found that the gestures adults spontaneously produce on this task are a good index of how well they will perform after new input, thus replicating previous work with children and demonstrating that, when considered in relation to speech, gesture is a good index of openness to change in all learners, young and old. But the mismatch between gesture and speech is only one component necessary to predict performance—the information conveyed in the mismatching gesture must also be explanatorily relevant to solving the problem. Overall, our findings make it clear that the gestures chemistry novices spontaneously produce can be used by researchers and teachers alike to gain insight into what learners know about chemistry and how ready they are to learn more.

## References

Alibali, M. W., Bassok, M., Olseth, K. L., Syc, S. E., & Goldin-Meadow, S. (1999). Illuminating mental representations through speech and gesture. *Psychological Sciences, 10,* 327-333.

Alibali, M. W., Flevares, L. M., & Goldin-Meadow, S. (1997). Assessing knowledge conveyed in gesture: Do teachers have the upper hand? *Journal of Educational Psychology*, *89*(1), 183-193.

Alibali, M. W., & Goldin-Meadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child′ s state of mind. *Cognitive psychology*, *25*(4), 468-523.

Alibali, M. W., & Nathan, M. J. (2012). Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the learning sciences*, *21*(2), 247-286.

Alibali, M. W., Nathan M. J., Fujimori Y., Stein N., Raudenbush S. (2011). Gestures in the mathematics classroom: What's the point? In: Stein N, Raudenbush S (Eds), *Developmental cognitive science goes to school* (pp. 219–234). New York, NY: Routledge, Taylor & Francis.

Atit, K., Gagnier, K. & Shipley, T.F. (2014). Student gestures aid penetrative thinking. *Journal of Geoscience Education, 63*(1), 66-72.

Beattie, G., & Shovelton, H. (1999). Do iconic hand gestures really contribute anything to the semantic information conveyed by speech? An experimental investigation. *Semiotica, 123*(1-2)*,* 1-30.

Beaudoin-Ryan, L., & Goldin-Meadow, S. (2014). Teaching moral reasoning through gesture. *Developmental Science*, *17*(6), 984-990.

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological science*, *21*(11), 1605-1610.

Becvar, L. A., Hollan, J., & Hutchins, E. (2005). Hands as molecules: Representational gestures used for developing theory in a scientific laboratory. *Semiotica, 2005*(156), 89-112.

Broaders, S. C., Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2007). Making children gesture brings out implicit knowledge and leads to learning. *Journal of Experimental Psychology: General, 136*(4), 539-550.

Chen, X. (2013). STEM Attrition: College Students' Paths Into and Out of STEM Fields (NCES 2014-001) (National Center for Education Statistics, Institute of Education Sciences, U.S). Washington, DC: Department of Education

Chu, M., & Kita, S. (2011). The nature of the beneficial role of spontaneous gesture in spatial problem solving. *Journal of Experimental Psychology: General, 140,* 102-116.

Church, R. B. (1999). Using gesture and speech to capture transitions in learning. *Cognitive Development*, *14*(2), 313-342. 10.1016/S0885-2014(99)00007-6.

Church, R. B., Ayman-Nolley, S., & Mahootian, S. (2004). The role of gesture in bilingual education: Does gesture enhance learning*? International Journal of Bilingual Education and Bilingualism*, *7*(4), 303-319.

Church, R. B., & Goldin-Meadow, S. (1986). The mismatch between gesture and speech as an index of transitional knowledge. *Cognition, 23*(1), 43-71.

Church, R. B., Schonert-Reichl, K., Goodman, N., Kelly, S. D., & Ayman-Nolley, S. (1995). The role of gesture and speech communication as reflections of cognitive understanding.

*Journal of Contemporary Legal Issues, 6*, 123-154.

Cook, S. W., Duffy, R. G., & Fenn, K. M. (2013). Consolidation and transfer of learning after observing hand gesture. *Child Development,* 84(6), 1863-1871. 10.1111/cdev.12097.

Cook, S. W., & Tanenhaus, M. K. (2009). Embodied understanding: Speakers' gestures affect listeners' actions. *Cognition, 113*, 98-104.

Crowder, E. M., & Newman, D. (1993). Telling what they know: The role of gesture and language in children's science explanations. *Pragmatics and Cognition, 1*(2)*,* 341-376.

Ehrlich, S. B., Levine, S. C., & Goldin-Meadow, S. (2006). The importance of gesture in children's spatial reasoning. *Developmental Psychology, 42*(6), 1259-1268.

Garber, P., & Goldin-Meadow, S. (2002). Gesture offers insight into problem-solving in children and adults. *Cognitive Science, 26*(6), 817-831.

Goldin-Meadow, S. (2003). Hearing gesture: How our hands help us think. Cambridge, MA: Harvard University Press.

Goldin-Meadow, S., & Alibali, M. W. (2002). Looking at the hands through time: A microgenetic perspective on learning and instruction. In: Granott N, Parziale J, editors. *Microdevelopment: Transition processes in development and learning* (pp. 80–105). New York, US: Cambridge University Press.

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept Acquisition: using the hand to read the mind. *Psychological Review, 100*(2), 279-297.

Goldin-Meadow, S., & Beilock, S. L. (2010). Action's influence on thought: The case of gesture. *Perspectives on Psychological Science, 5*(6), 664-674.

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesture gives children new ideas about math. *Psychological Science, 20*(3), 267–271. 10.1111/j.1467-

9280.2009.02297

Goldin-Meadow, S., & Sandhofer, C. M. (1999). Gesture convey substantive information about a child's thoughts to ordinary listeners. *Developmental Science*, *2*(1), 67-74.

Goldin-Meadow, S., Wein, D., & Chang, C. (1992). Assessing knowledge through gesture: Using children's hands to read their minds. *Cognition and Instruction*, *9*(3), 201-219.

Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics, 32*(10), 1489-1522.

Goodwin, C. (2007). Participation, stance and affect in the organization of activities. *Discourse & Society, 18*(1), 53-73.

Goodwin, C. (2010). Things and their embodied environments. In Malafouris, Lambros, & Renfrew (Ed.), *The cognitive life of things: Recasting the boundaries of the mind* (pp. 103-120). McDonald Institute for Anthropological Research.

Hostetter, A. B., & Alibali, M. W. (2008) Visible embodiment: Gestures as simulated action *Psychonomic Bulletin & Review, 15*(3), 495-514.

Kelly, S. D., Singer, M., Hicks, J., & Goldin-Meadow, S. (2002). A helping hand in assessing children's knowledge: Instructing adults to attend to gesture. *Cognition and Instruction*, *20*(1), 1-26.

Lave, J. (1991). Situating learning in communities of practice. In L. Resnick, J. Levine, and S. Teasley (Eds.), *Perspectives on Socially Shared Cognition* (pp. 63-82). Washington, DC: APA.

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

McNeill, D., Cassell, J., & McCullough, K. E. (1994). Communicative effects of speech-

mismatched gestures. *Research on Language & Social Interaction, 27*(3), 223-237.

Morrell-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(3), 615-622.

Perry, M., Church, R. B., & Goldin-Meadow, S. (1988). Transitional knowledge in the acquisition of concepts. *Cognitive Development, 3*(4), 359-400.

Perry, M., & Elder, A. D. (1997). Knowledge in transition: Adults' developing understanding of a principle of physical causality. *Cognitive Development, 12*(1), 131-157.

Piaget, J. (1952). The origins of intelligence in children. New York: International Universities Press.

Pine, K. J., Lufkin, N., & Messer, D. J. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology, 40*(6), 1059-1067.

Ping, R. M., Goldin-Meadow, S., & Beilock, S. L. (2014). Understanding gesture: Is the listener's motor system involved? *Journal of Experimental Psychology: General, 143*(1), 195-204. doi:10.1037/a0032246.

Radinsky, J., Ping, R., Hospelhorn, E., & Goldman, S. (2012). Making the absent present: Improvised representational fields in students' negotiation of meaning with GIS. In the Proceedings of the Annual Meeting of the American Educational Research Association.

Rauscher, F. H., Krauss, R. M., & Chen, Y. (1996). Gesture, speech, and lexical access: The role of lexical movements in speech production. *Psychological Science, 7*(4), 226-231.

Rueckert, L., Church, R. B., Avila, A., & Trejo, T. (2017). Gesture enhances learning of a complex statistical concept. *Cognitive Research: Principles and Implications*, *2*(1), 2. doi: 10.1186/s41235-016-0036-1

Siegler, R. S. (2007). Cognitive variability. *Developmental Science, 10*(1), 104-109.

Singer, M., Radinsky, J., & Goldman, S. R. (2008). The role of gesture in meaning construction. *Discourse Processes, 45*(4-5), 365-386. 10.1080/01638530802145601

Stieff, M. (2007). Mental rotation and diagrammatic reasoning in science. *Learning and Instruction, 17*(2), 219-234.

Stieff, M. (2011). When is a molecule three dimensional? A task-specific role for imagistic reasoning in advanced chemistry. *Science Education, 95*(2), 310-336.

Stieff, M., & Raje, S. (2010). Expertise algorithmic and imagistic problem solving strategies in advanced chemistry. *Spatial Cognition & Computation, 10*(1), 53-81.

Trafton, J. G., Trickett, S. B., Stitzlein, C. A., Saner, L., & Schunn, C. D. (2006). The relationship between spatial transformations and iconic gestures. *Spatial Cognition and Computation, 6*(1), 1-29.

Table 1

*Common problem-solving strategies with descriptions and examples from each modality*

**Table 1a**. Strategies highlighting dimensions that are explanatorily irrelevant to the stereoisomer problem (*Irrelevant Dimension Strategies*)

| Level | Strategy | Description | Speech Example | Gesture Example |
|---|---|---|---|---|
| 0 | Changed 2D Representation | By changing the length of the bonds between atoms or altering the angle between two bonds, they have created a stereoisomer. | "I made the bonds between the carbon and the hydrogen shorter." | Pinch two fingers together over the C and H. |
| | No Changes Possible | Regardless of how they change the molecule it can always be rotated back to the original form and thus cannot create a stereoisomer. Or, due to the laws of chemistry no changes are allowed to be made. | "Anyway I drew it, it would still be superimposable on the original molecule" | Point or sweep to every substituent and produce no other gestures. |
| 1 | Ignore Z Dimension | Indicate having changed the location of two substituents by altering their spatial relation on the 2-dimensional X-Y plane of the drawing in such a way that represents no 3-dimensional change. | "Originally the hydrogen was to the bottom left of the carbon, now it's to the bottom right. And the bromine was on the right, now it's on the left." | V-hand shape, pointing to each of the two substituents on the Z-axis, typically wiggling back and forth. |
| | Irrelevant Switch | By changing the 3-dimensional position of | "I pulled this part, attached over here on | Point to a substituent that is not attached to stereocenter, and then |

| | | | | |
|---|---|---|---|---|
| | | substituents that are NOT attached to the stereocenter they have created a stereoisomer. They have manipulated the spatial locations of irrelevant parts of the molecule. | the very very left, out from the board. So it was next to this, now it's coming out from it." | sweeps away from the board. |
| | Irrelevant Rotate | Participants indicate that rotating two or more substituents that are NOT attached to the stereocenter will produce a stereoisomer. | "I rotated these two $CH_3$s all around this carbon so now it is different." | Lift up one hand with fingers pointed upward and twist the hand either clockwise or counter-clockwise. The movement is produced in front of a substituent not attached to a stereocenter. |

**Table 1b.** Strategies highlighting dimensions that are explanatorily relevant to the stereoisomer problem (*Relevant Dimension Strategies*)

| Level | Strategy | Description | Speech Example | Gesture Example |
|---|---|---|---|---|
| 2 | Relevant Switch | By switching the positions of two substituent groups attached to the stereocenter, along anyone or combination of the X, Y, and Z axes, they have created a stereoisomer. | "It is not superimposable because I switched the orientation of the chlorine and the hydrogen. Now, Cl is coming out and H is going away." | Point with a V-shaped hand to two substituents attached to a stereocenter and flip hand. |
| | Relevant Rotate | Participants indicate that no matter how their drawn molecule is rotated, it cannot be superimposed on the original molecule. For items that do not have a stereoisomer, rotating the drawn molecule will return the arrangement to its | "Because regardless of how you rotate the molecule, the parts of this one are not going to line up with the parts of the original one." | Lift up one hand with fingers pointed upward (or both hands) and twist the hand(s) either clockwise or counter-clockwise. The movement is produced in front of the entire |

| | | | | |
|---|---|---|---|---|
| | | original form and thus it can be superimposed. | | molecule or stereocenter. |
| 3 | Mirror Image | Participants indicate that by creating the original molecule's mirror image they have created a stereoisomer. | "If you reflect the molecule over a mirror plane then you couldn't match it up with the original." | Place one hand over one drawing and then flip hand from palm down to palm up over the top of the original molecule |
| 4 | Check Non-Manipulated Substituents | Indicate that a molecule manipulated by either Mirror Image (Lv. 3) or Relevant Switch & Relevant Rotation (Lv. 2) must be compared to the original to check superimposability. Typically participants checked manipulated substituents against their original orientations, then checked the two non-manipulated substituents' locations to their original orientations. | "When you change which one of these two groups are going into the board and which one is coming out of the board, no matter how you rotate the molecule, it won't be the same. So I rotate this around, these will be back in their original spots, but the two things I didn't change are now in different spots. They don't match up so I can't superimpose this one on this one." | Point to one or both of the unchanged substituents attached to the stereocenter. |

Table 2

*Frequency of each strategy in each modality at pretest*

| **Table 2a.** Strategies highlighting explanatorily *irrelevant* dimensions | | In Both Speech and Gesture | In Speech Only | In Gesture Only |
|---|---|---|---|---|
| Level 0 | Changed 2D Representation | 1 | 2 | 14 |
| | No Changes Possible | 0 | 2 | 13 |
| Level 1 | Ignoring Z Dimension | 15 | 36 | 5 |
| | Irrelevant Switch | 21 | 54 | 6 |
| | Irrelevant Rotate | 8 | 9 | 12 |
| **Table 2b.** Strategies highlighting explanatorily *relevant* dimensions | | In Both Speech and Gesture | In Speech Only | In Gesture Only |
| Level 2 | Relevant Switch | 57 | 29 | 30 |
| | Relevant Rotate | 41 | 23 | 37 |
| Level 3 | Mirror Image | 2 | 18 | 0 |
| Level 4 | Compare Non-Manipulated Substituents | 1 | 4 | 2 |

*Note*: The number of times each strategy that highlighted (2a) an irrelevant dimension and each strategy that highlighted (2b) a relevant dimension was produced within a single explanation in both gesture and speech, in speech only, or in gesture only on the pretest.

Table 3

*Mean number of strategies produced per participant at pretest and posttest categorized according to relevance and modality*

| Type of Strategy | Mean Number of Strategies Produced (SD) | | *t*-value (*df*=42) | *p*-value |
|---|---|---|---|---|
| | At Pretest | At Posttest | | |
| Irrelevant Strategies | | | | |
| In Gesture+Speech | 1.04 (SD=1.37) | 0.51 (SD=0.86) | -2.78 | 0.008 |
| In Speech Only | 2.65 (SD=2.41) | 1.58 (SD=2.41) | -2.55 | 0.014 |
| In Gesture Only | 0.91 (SD=1.29) | 0.81 (SD=1.24) | -0.35 | 0.726 |
| Relevant Strategies | | | | |
| In Gesture+Speech | 2.35 (SD=1.97) | 3.67 (SD=2.45) | 3.38 | 0.001 |
| In Speech Only | 1.72 (SD=1.54) | 3.79 (SD=2.41) | 5.69 | <0.0001 |
| In Gesture Only | 1.60 (SD=1.45) | 1.16 (SD=1.13) | -1.63 | 0.11 |

Table 4

*Summary for Binary Mixed Model Predicting Success at Posttest*

| Predictor | Beta | Std Error | t value | p value |
|---|---|---|---|---|
| *(Intercept)* | -2.056 | 1.817 | -1.131 | 0.2580 |
| **Average Correct at Pretest** | **18.688** | **8.624** | **2.167** | **0.0302** |
| Relevant Strategies in Both Modalities | -1.812 | 1.741 | -1.041 | 0.2981 |
| Relevant Strategies in Speech Only | -3.701 | 2.752 | -1.345 | 0.1786 |
| **Relevant Strategies in Gesture Only** | **4.385** | **1.698** | **2.583** | **0.0098** |
| Irrelevant Strategies in Both Modalities | 3.565 | 2.522 | 1.414 | 0.1575 |
| Irrelevant Strategies in Speech Only | -4.623 | 3.000 | -1.541 | 0.1234 |
| Irrelevant Strategies in Gesture Only | -2.626 | 2.054 | -1.279 | 0.2010 |

*Note:* Summary for model predicting the average score at posttest from pretest score, strategy modality, and strategy relevance. The only significant predictor of success, besides pretest performance, is relevant strategies produced in gesture only.

Table 5

*Summary for models predicting success at posttest on the basis of relevance or modality considered independently*

| **Table 5a.** Strategy Relevance | Beta | Std Error | t value | p value |
|---|---|---|---|---|
| *(Intercept)* | -4.8439 | 1.2503 | -3.874 | 0.0001 |
| Average Correct at Pretest | 6.5227 | 5.2862 | 1.238 | 0.2157 |
| Relevant Strategies in Any Modality | 0.9205 | 0.8334 | 1.104 | 0.2694 |
| Irrelevant Strategies in Any Modality | 0.2913 | 0.7370 | 0.395 | 0.6926 |

| **Table 5b.** Strategy Modality | Beta | Std Error | t value | p value |
|---|---|---|---|---|
| *(Intercept)* | -4.1428 | 1.6722 | -2.477 | 0.0132 |
| Average Correct at Pretest | 11.9660 | 8.066 | 1.483 | 0.1380 |
| Any Strategy in Both Modalities | -0.5746 | 1.3234 | -0.434 | 0.6641 |
| Any Strategy in Speech Only | -0.8332 | 1.4702 | -0.567 | 0.5709 |
| Any Strategy in Gesture Only | 2.0072 | 1.2264 | 1.637 | 0.1017 |

*Note:* Summary for model predicting the average score at posttest based on strategy relevance (ignoring modality) at pretest (5a) and strategy modality (ignoring relevance) at pretest (5b). Neither factor is a significant predictor on its own, highlighting the importance of taking both factors into account.

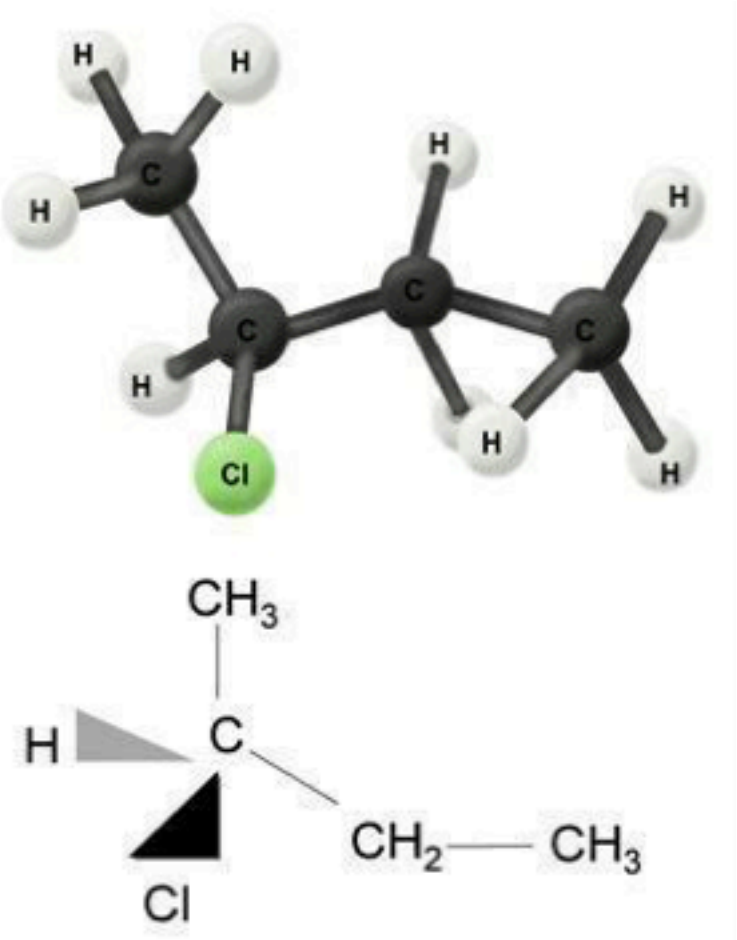
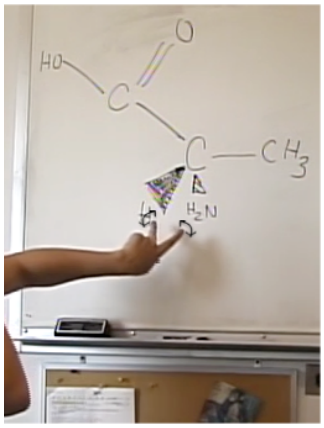


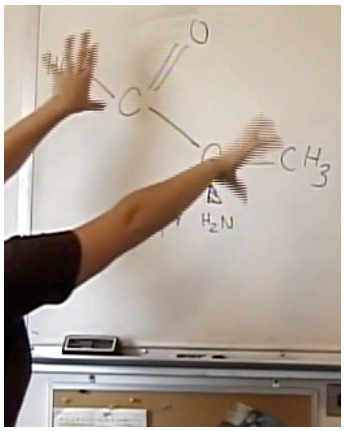

*Figure 1.* An example of a stimulus item given to participants ((S)-2-Chlorobutane).



*Relevant Switch*

**Speech**: "Well I *reversed the spatial arrangement of these two*, so now you can't like…

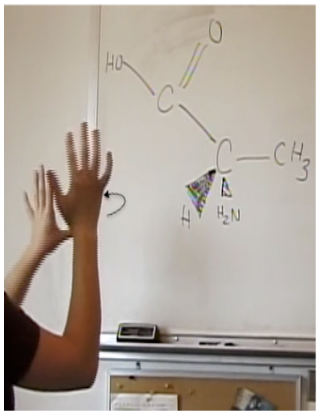
**Gesture**: Index and middle finger each point at one substituent lying at non-zero points on the Z-axis, and wiggle back and forth.



*Compare the Non-Manipulated Substituents*

**Speech**: …Since *these are not the same on either side* of the main carbon,

**Gesture**: Left hand flat—points at one substituent on the X-Y plane (where Z=0); right hand flat—points at the other substituent on the X-Y plane.



*Relevant Rotation*

**Speech**: then *you can't like flip it* at all."

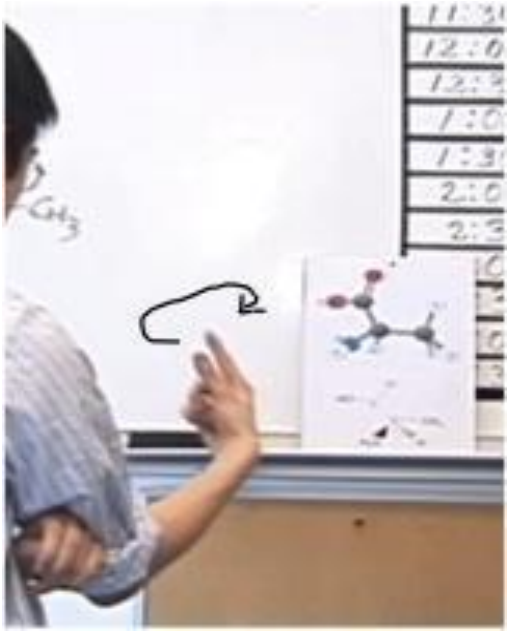
**Gesture**: Both hands mimic rotation in front of the molecule.

*Figure 2.* An example of three relevant strategies, each produced in both modalities within a single explanation.

*Relevant Switch*
(in gesture + speech)

**Speech**: "There's an $NH_2$ group and an H group *connected to this carbon atom. If you switch these two groups around* then

**Gesture**: Index and middle finger each point at one substituent lying at non-zero points on the Z-axis, and wiggle back and forth.



*Relevant Rotation*
(in gesture only)

**Speech**: according to my head, you can't superimpose this on top of that."

**Gesture**: Finger pointed upward and hand rotates counter clockwise twice. The movement is produced in front of the entire molecule generally (and not a specific substituent)

*Figure 3.* An example of Relevant Switch in both modalities (left) and Relevant Rotation in gesture only (right) produced within a single explanation.

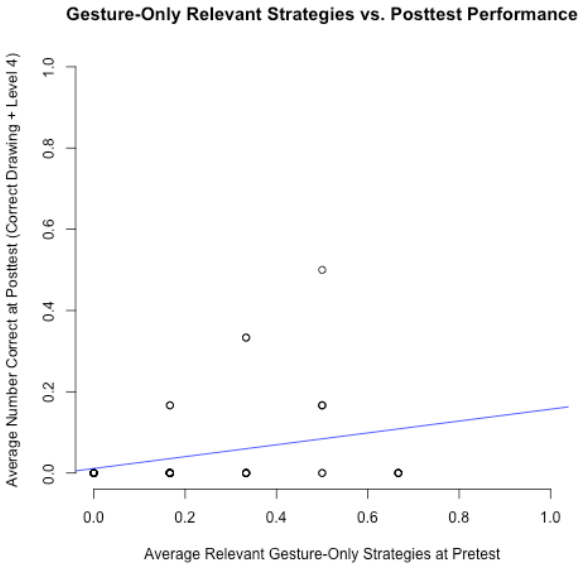



*Figure 4.* Scatter plot of posttest success as a function of the average number of gesture-only relevant responses produced on the pretest (a level 4 strategy in speech combined with a correct drawing was considered success on the posttest).

Table S1

*Correlations among the six strategy types at pretest*

| | | Irrelevant Strategies | | | Relevant Strategies | | |
|---|---|---|---|---|---|---|---|
| | | In Gesture + Speech | In Speech Only | In Gesture Only | In Gesture + Speech | In Speech Only | In Gesture Only |
| Irrelevant Strategies | In Gesture + Speech | x | .31 * | .31 * | -.11 | -.23 | -.04 |
| | In Speech Only | | x | -.19 | -.48 ** | -.38 * | .01 |
| | In Gesture Only | | | x | .19 | -.19 | .07 |
| Relevant Strategies | In Gesture + Speech | | | | x | -.11 | -.05 |
| | In Speech Only | | | | | x | -.09 |
| | In Gesture Only | | | | | | x |

*Note*: * $p < 0.05$; ** $p < 0.01$