Using Referential Communication to Study Mental Models

Julia Kalmanson (kalmanso@psych.columbia.edu)
Department of Psychology; Columbia University, 406 Schermerhorn Hall
New York, NY 10027

Arthur B. Markman (markman@psy.utexas.edu)
Department of Psychology, University of Texas, Mezes Hall 330
Austin, TX 78712

Abstract

In this paper we evaluate the utility of a referential communication paradigm to study the content and use of mental models. In this task pairs of people collaborate to determine which of a set of infra-red images depicts a physically possible situation. We demonstrate that the referential communication task provides insight into the interaction between the content and use of naive theories of physics in a problem-solving domain.

Introduction

Causal explanation is critical for our daily existence. Causal connections support our perception of the world as coherent, and they give us a sense of mechanism,- a sense of how things work. Our control of the physical world is to a great extent dependent on the accuracy of our understanding of the mechanics of the world.

A prominent suggestion is that causal knowledge is organized into theories that people use to reason about the world. The term “theory” has been used in a number of different ways by psychologists (see e.g., Gopnik & Meltzoff, 1997, for a discussion). One view assumes that theories are large bodies of knowledge that are coherently organized according to a few well-defined principles, so that all explanations can be deductively derived (e.g., Kuhn, 1991). A more local view of theories is presented by Murphy and Medin (1985), who define theories as "any host of mental explanations, rather than a complete organized scientific account" (p.290). Schemas and scripts contain implicit theories of causality that allow us to explicate the world, although may not possess properties of coherence and consistency.

One area where theories have received extensive attention is in naive physics, which is concerned with understanding how knowledge and experience are integrated to create an understanding of the mechanics of the world. Despite the fact that physical principles describe properties of objects with which we interact daily, people have serious difficulties understanding formal principles of physics. People's understanding of fundamental physical principles has been described as incoherent and full of misconceptions (diSessa, 1993; McCloskey et al., 1980; Clement, 1982; Cooke and Breedin, 1994).

Various theories have been proposed to account for these difficulties. Some researchers have suggested that these misconceptions arise from basic misunderstandings of physical systems that are formed prior to any formal training in physics (Caramazza and Green, 1980; McCloskey, 1983a, 1983b). The systematic nature of some errors has led researchers to suggests that certain misconceptions are not idiosyncratic, but instead are based on a more general system of beliefs or a naive theory (Clement 1982; McCloskey, 1983a, b). These theories are described as systematic, general, coherent, well-developed and well articulated conceptions that conflict with basic principles of physics, but that nonetheless adequately explain observed events in the world (McCloskey, 1983b).

An alternative view contends that people's understanding of physical phenomena is a collection of fragmented and loosely connected ideas about the world that can be used to generate situation specific explanations (diSessa, 1988, 1993). In his view, naive theories are nothing more than ad hoc explanations that are invented for particular situations. Ueno (1993) in his re-interpretation of diSessa's theory points out that these explanations are socially formed and shared. They are maintained through communication and are to a great extent guided by conversational pragmatics. For example it would be anomalous to cast a simple sentence like "Susie slapped Tom" in it's Newtonian physics equivalent of "Tom's face slaps the palm of Susie's hand, while the force of Susie's slapping is the same as the force of Tom's slapping". In everyday discourse the latter sentence would be judged nonsensical.

This reinterpretation is in line with current research on the role of communication in category acquisition. Category representations are structured in a manner that facilitates communication. People typically learn categories in the process of communicating with others. Further, people are constrained to form categories that are shared by other members of their culture if they are to use them effectively (Garrod & Doherty, 1994; Malt, 1995). In this view naive theories of physics are pragmatically motivated explanations of complex phenomena that are socially constructed to support our simplistic categorizations of the physical world.

Ultimately, of course, we would like to understand the structure of people’s naive theories. In the present paper,

we begin to address this issue by examining a novel method that can elicit people's naive theories of the physical world and to explore the causal relationships that make up those theories. This methodology draws on recent findings on the role of communication in category acquisition, and attempts to elicit and explore naive theories in a communicative setting.

A popular method of eliciting people's theories of physical phenomena is to ask people to explain their predictions or decisions in interviews or other verbal protocols. These verbal protocols permit a systematic examination of explanations people generate for their own errors in reasoning. Chi and her colleagues have successfully used this method to study learning in a variety of problem-solving tasks (Chi, 1989, 1983). The method we propose is novel in that it incorporates a referential communication design into the study of naive theories. In this task, pairs of people are presented with four infra-red pictures that show the heat emanating from objects. One of the pictures is an actual infra-red image, and the other three are doctored images that have been altered to contain systematic errors. Pairs of people are shown these images and are asked to determine which image is correct.

In order to perform this task, dyads must talk about the heat pictured in the image. In this way, they must use their naive understanding of thermodynamics. Because the task involves two people, many aspects of people's beliefs about heat are stated explicitly in the conversations. People have extensive experience with heat and have a naive theory of heat-flow.

Referential communication tasks are quite data-intensive, as full transcripts of conversations are developed and must be coded. In this paper, we limit our focus to three issues. First, does communicating about this task improve performance? This question is useful for understanding whether theory development occurs through communication. If the performance of dyads improves with communication, then it is plausible to think that theories develop when people communicate. Second, we are interested in whether people use correct theories when discussing thermodynamics. Finally, we will address the relationship between theories and topics of discussion, as well as the qualitative change in discussions over time.

Method

Participants: Participants were 70 members of the Columbia University community (50 in the dyad task and 20 in the control task). Six dyads involved two male participants, five involved two female participants, and the rest were mixed sex dyads. All participants were native speakers of English who did not know each other before the session. Data from one dyad had to be eliminated due to mechanical failure leaving a total of 24 dyads for analysis.

Materials and Procedure: The stimuli were 12 sets of false-color infra-red images of familiar objects and scenes such as plants, kitchen appliances, and street scenes. Each set consisted of one actual image and three variants of it. The actual image was a picture of the infra-red (i.e., thermal) energy at the surfaces of the image. The color scheme involved 10 colors that were each assigned to a range of temperatures. The resulting image is called a false-color image, because it appears in colors that differ from the colors of those objects in visible light. To complete each set three additional versions of each picture were created using Adobe Photoshop by changing some of the colors in the image to create thermal inconsistencies that are highly unlikely to occur naturally. For example, the nose of a dog might be made to appear cooler than the fur-covered skin, or the pattern of heat diffusion from a heat source might be changed so that temperature did not decrease monotonically with distance.

During the experiment the sets were presented in a random order. Each pair of participants was instructed to collaborate to figure out which image in the set was the actual thermal image. Both subjects had to agree on their response before the trial was completed. To encourage discussion rather than pointing, a divider was placed between the subjects, and each subject was given an identical set of 4 images. Thus, subjects were free to refer to pictures verbally but could not point to pictures or their elements to establish reference. The discussions were video-taped and later transcribed. All subjects were aware of being videotaped. The control group consisted of 20 subjects who performed the same picture selection task alone and without verbalizing their reasons.

General Coding: Each utterance from the transcript was coded along six dimensions. An utterance was defined as a turn each subject took when speaking. Thus, an utterance could contain as little as a sentence fragment or as much as a paragraph. Because of space limitations, we will focus on two codes: 1). the correctness of the theory and action taken by the dyad and 2). the topic of the discussion. To assess the reliability of the coding, ten of the transcripts were scored by both coders. Correlations ranging between .9 and .98 were obtained for all codes.

The correctness of the theory and action code focused on utterances where the dyad took an action (either selecting a particular picture as the correct one or rejecting a picture). First, the action was coded as correct or incorrect. A correct action was either rejecting a picture that was not an actual thermal image, or selecting the valid image. An incorrect action was either rejecting the correct image or accepting an incorrect one. Actions were typically justified in some way, and the theory part of the code assessed whether the justification was in accord with basic principles of physics. Thus, this code had four levels:

1. Correct action considered on the basis of incorrect theory
2. Correct action considered on the basis of correct theory.
3. Incorrect action considered on the basis of incorrect theory.
4. Incorrect action considered on the basis of correct theory.

The discussion topic code distinguished between five different topics including discussions of abstract physical principles, discussions of the thermal conductivity of materials, and discussions of the internal mechanics of an object depicted. Of these codes only two yielded enough observations to warrant further analysis: 1). discussion of temperature and 2). discussion of heat diffusion. Temperature referred to explicit discussions of the temperature at particular points in the image or to relative temperatures at neighboring points. Discussions of heat diffusion were cases in which people talked about the flow of heat from one location to another or to the dissipation of heat. Because naive theories often treat temperature as a physical quantity (rather than a measure of mean molecular kinetic energy), discussions of temperature are likely to be associated with poor reasoning about thermal images (Wiser & Carey, 1983). In contrast, discussions of heat flow and thermodynamics are more likely to be related to an accurate theory of thermodynamics, and so they should be associated with good reasoning about thermal images.

Predictions

In this paper, we focus on four aspects of the present task:

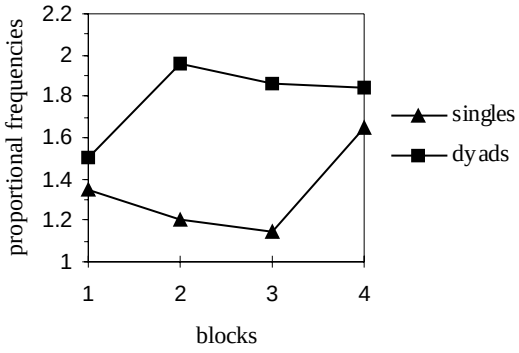
1. Communication: The first question to be explored is whether communication influences performance accuracy on this task. To address this issue, we test to see if dyads have higher accuracy than do people who perform the task alone. In addition to examining overall accuracy, we look at performance curves over the course of the experiment. Related to this issue, we can explore how the performance of dyads changes over time.
2. Correctness of theories: Expertise is typically characterized by the presence of a fully integrated representation of the domain of expertise. Experts in domains like physics reason better than novices, because they are able to focus on deep relational aspects of the situation rather than being derailed by surface aspects of the task (e.g., Chi, Feltovich, & Glaser, 1981). In the present task, we expect that subjects who exhibit evidence of having a correct theory of thermodynamics will perform the picture selection task more accurately than those with fragmentary knowledge of this domain.
3. Considered actions: When discussing an image, a dyad could decide to classify it as one of the transformed images or to retain the picture in the set that could be classified as the unaltered thermal image. Considering a particular action in relation to a particular picture singles that picture out and temporarily makes it more salient than others. It may be that mere consideration of an action, be it based on a correct or incorrect theory, affects the decision by anchoring people on a particular picture. That is, if a dyad starts out considering a particular picture and spends extensive time and energy discussing it, that investment alone may be sufficient to influence the final choice.
4. Topics of discussion: As discussed above, temperature and heat differ in their relationship to a correct theory of thermodynamics. We will explore the relationship between the topics discussed and people's accuracy in classifying the pictures to see whether talking about terms that are more relevant to correct theories is related to accurate performance.

Analysis and Results

Communication: A key part of the present task was that dyads worked together to find the correct thermal image. As a way of exploring accuracy on this task, we explore differences between the dyads and the control group on their accuracy in selecting the actual thermal image. This analysis addressed the following questions: 1. Are dyads performing better than singles? 2. Are there differences in performance during the experiment? and 3. How does performance of dyads change over time? An examination of overall accuracy revealed that dyads were significantly more accurate (M=7.3 (out of 12)) than were the people in the control group who worked alone (M=5.3), $t=3.1$, $p<.05$. Chance performance would be 3 correct out of 12. Both groups performed reliably above chance.

To explore how the dyads differed from the control group more carefully, we broke down the performance data into four blocks of three trials. The accuracy for each block for both conditions is shown in Figure 1.

Figure 1: Accuracy by Blocks of Three Trials



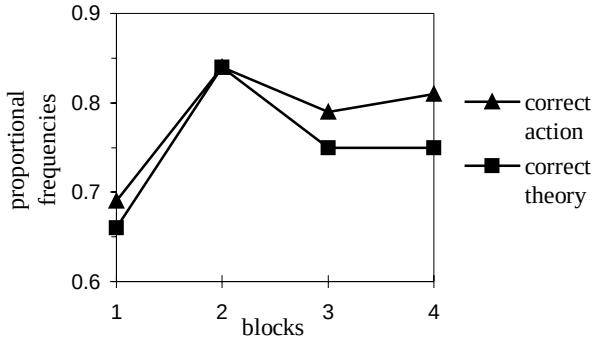
As we would expect if the people in each condition were approximately equivalent in their expertise, performance in the first block of trials is about the same in each group. The groups diverge after the first block. By the second block, the dyads are significantly more accurate (M=1.96) than the singles (M=1.2), $t= -2.66$, $p<.05$. The performance of singles does improve with practice, but this does not occur until after the last block (M=1.65).

Theories and Actions: We now turn to the relationship between type of action, correctness of theory and overall accuracy. For this analysis, we first converted the frequency of each combination of accuracy and theory correctness for each dyad to a proportion. This conversion allowed us to control for individual differences in the length and content of dyads' discussions. We expected that correctness of

theories would be the major factor that determined accuracy. However, this prediction was not borne out. As expected, consideration of a correct action was positively related to accuracy (r =. 58, p<.01). However, in contrast to our expectations, consideration of an incorrect action was negatively correlated with accuracy regardless of whether the statement was accompanied by a correct or an incorrect theory (r(incorrect action/incorrect theory, accuracy)=-.67; r(incorrect action/correct theory, accuracy)=-.47, both p<.05). If we examine correctness of action and correctness of theory separately, we also find that correct actions are a better predictor of accuracy than correct theories (r=.74, p<.001 and r=.48, p<.05 respectively). [1]

The changes in performance during the experiment may be related to the consideration of correct actions and theories over the course of the study. Figure 2 shows the frequency with which correct actions and theories are considered as a function of four performance blocks, each consisting of three trials.

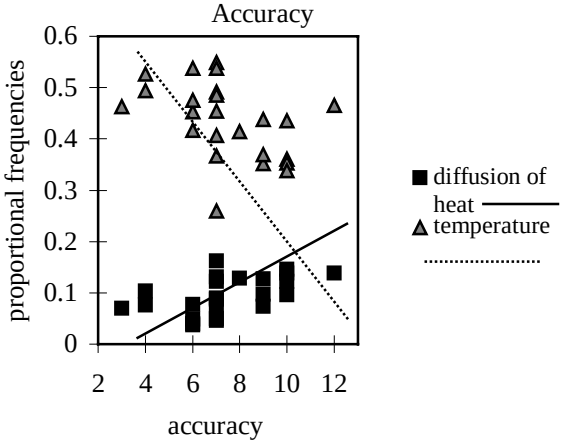
Figure 2 Correct Theories and Actions by Blocks of Three Trials



As shown in the graph, the pattern of change in the consideration of correct actions and theories closely follows that of the accuracy improvement pattern. The greatest increase in the frequency of correct actions and theories occurs between the first and the second quarters, followed by a plateau thereafter. Looking at the proportional frequency of correct actions considered on the basis of correct theories and incorrect actions on the basis of incorrect theories we see a similar separation.

Topics of Discussion: Finally, we can examine the relationship between accuracy and topics discussed. As described above, we are most interested in discussions of temperature and heat flow. The relationship between the use of these topics and accuracy is shown in Figure 3. As expected, discussion of temperature is negatively related to accuracy (r=-.43, p<.05) and discussion of diffusion of heat is positively related to accuracy (r=.68, p<.05). This finding reflects that temperature is a static quality of an image that may signal an incorrect theory, while heat flow is a property that is central to a correct theory of thermodynamics.

Figure 3: Topics of Discussion and Accuracy



We can also look at the relationship between discussion topic and the use of correct actions and correct theories. We expect the discussion of temperature to be negatively related to the use of the correct theory and to the selection of correct actions. Consistent with this prediction, there is a relationship between the frequency of discussion of temperature and the use of incorrect theories (r=.56, p<.05). There is also a positive relationship between the frequency of discussion of temperature and the consideration of incorrect actions (r=.59, p<.05). In contrast, the frequency of discussions of heat diffusion are related to the consideration of correct actions (r=.52, p<.01) Contrary to our expectation, no significant relationship was observed between the frequency of discussion of heat diffusion and the use of correct theories. This unexpected finding will be discussed in more detail below.

Finally, the use of these topics was explored across the four blocks of the experiment. While the frequency of discussion of temperatures did not change over the course of the experiment, there was an increase in the frequency of discussion of diffusion of heat during the experiment.

Discussion

This research is the beginning of a line of research aimed at exploring the development and use of mental models. In order to get a view of people's mental models (in this case mental models of heat flow), we examined their interactions in a task in which they selected an actual thermal image from a set of doctored images of the same object. The dyadic design allows us to get people to talk about their mental models without having to resort to a an unnatural task like a concurrent verbal protocol.

Any study involving transcripts of communication creates a large amount of data, and we have necessarily covered only a small fraction of what could be extracted from this work. We have focused on four main issues here: 1) the

[1] Correct action/incorrect theory code is not represented here because it yielded very low frequencies as compared to frequencies of other codes, and while the results are in the right direction, they were not significant.

influence of causal theories on performance accuracy in a problem solving task; 2) the relationship of considered actions to causal theories, and their influence on accuracy in a problem solving task; 3) the relationship between topics of conversation and causal theories; 4) the qualitative change in discussions and theories over time.

Theories and Actions: What sort of theories do people tend to have about thermodynamics? The mean accuracy in this task was 60.8%, which was well above chance. Thus, people who had probably never seen a thermal image before were pretty good at identifying actual images. While this performance must have been based on some knowledge of thermodynamics, it was not based on an accurate physical theory. The data suggest that correct performance in the task was influenced far more strongly by whether the dyad considered correct actions (properly accepting or rejecting a picture) than it was by the presence of accurate discussions of physical principles.

There are (at least) two important reasons for this finding. First, we defined the use of a correct theory as a discussion that was internally consistent and did not contradict basic principles of physics. It is quite likely, however, that this definition is too restrictive. Many people's naive physical models are successful at predicting performance in the world without necessarily embodying principles from the science of physics. McCloskey (1983b) points out that people's naive theories are well-developed conceptions that are useful for predicting the behavior of objects in the world. However, these theories often conflict with basic principles of physics. Further research must explore ways of characterizing people's naive theories of thermodynamics. One task that we have begun to use that has some promise is to give people a blank picture of a scene (such as an outdoor scene during the winter at night) and ask people to color the scene as if they were looking at the heat coming off surfaces. Pilot research with this technique suggests that it is capable of uncovering situations where people's mental model of heat differs from scientifically accepted principles.

A second reason why the presence of a correct model of heat did not always lead to correct performance is that people may have a correct model of thermodynamics, but may have some difficulty translating that model to thermal images. For example, people often correctly recognize that heat will escape from an open window if the room is warmer than the surrounding area. However, they may mistakenly expect the room to look cold in a thermal image, because it would feel cold to be standing in this room with the open window. In fact, such a room would look warm, as an infra-red camera would be seeing the heat energy escaping from the room. Thus, people may understand principles of thermodynamics but have difficulty transferring this knowledge to thermal images.

One factor that may account for the difference in the learning curves has to do with the learning benefit associated with constantly verbalizing one's thoughts in a collaborative process. Chi and her colleagues have successfully used talk change in the frequency of discussions of diffusion of heat followed the same pattern.

Topics of Discussion: Another way to explore people's mental models is to look at the topics that get considered for discussion. A key distinction involves differences between discussions of temperature and discussions of heat flow. Discussions of temperature were associated with lower accuracy and less use of correct theories in this study. There are two reasons why this relationship makes sense. First, to the extent that people are treating temperature as a property of objects rather than as a measure of heat energy, they are subscribing to a mental model that is not in accord with physical principles (Wiser & Carey, 1983). Second, even if they recognize that temperature is a measure of mean molecular kinetic energy, they are still focusing on an attribute of an object. Reasoning about physical principles also requires consideration of relational properties (e.g., Gentner, et al., 1997). Discussions of heat flow, in contrast, reflect a discussion of relational properties of the domain. An important aspect of heat is that it flows from high temperature regions to low temperature regions. Focusing on these relations is often useful for understanding how thermal images are in error. In many cases, errors in thermal images reflect situations in which heat is flowing in an impossible way. The relations between locations are critical for finding errors in images.

Heat flow should also generally be related to the use of a correct theory of thermodynamics. Contrary to this expectation, there was no significant relationship between discussions of heat flow in our data. This discrepancy probably reflects the same problem raised above that our coding scheme focused on theories that were both internally consistent and in accord with physical principles. It is possible that people's models are fragmentary, and thus prone to exhibit inconsistencies. Further work must address this issue.

Communication: Another striking aspect of the data was that dyads were significantly more accurate than were people who performed the task alone. This difference in accuracy manifested itself in a difference in performance across blocks. The dyads showed the greatest improvement in performance accuracy in the shift from the first block of three trials to the next. In contrast, singles did improve until the final block of three trials. The frequency of correct actions and theories for dyads closely followed the pattern of the performance curve. Similarly, the pattern of out-loud protocols to study problem solving strategies in a variety of tasks. One finding that emerged from this methodology is the learning benefit of self-generated explanations (Chi, 1989; 1993). Chi argues that learning requires integration of existing knowledge with new information and that the process of self-explanations

facilitates this integration. Self-explanations derived from talk out-loud protocols have been shown to improve understanding and to enhance learning (Chi, 1989, 1994; Webb, 1989). High self-explainers display deeper understanding and more complete mental models than low self-explainers as assessed by ability to answer complex questions. Chi argues that the beneficial effect of self-explanations is partly due to the fact that self-explaining is essentially a constructive activity. Self explanations provide an opportunity to construct new declarative knowledge and to generate new rules that can subsequently be used to solve complex problems. In our study, dyads are forced by the nature of the task to engage in explanatory activity from the very outset. Since the task itself is novel, there is a strong demand to integrate and adopt an existing knowledge and to construct new rules appropriate for the task at hand. To the extent that self-explanation is a constructive activity, the construction of the new knowledge structure needed to succeed on the task is started from the very onset of the task through self-explanations and explanations designed for the partner. No similar demand was placed on the singles performing the task. They were not required to verbalize their strategies, although it is interesting to note that a few subjects had spontaneously attempted to think out-loud in the course of the study, and had to be stopped by the experimenter. Thus, while the same mental process of self-explaining may be going on in the minds of the singles, there is no experimental constraint to facilitate it. This may account for the delay in improvement among singles. The learning benefit of self-explanations and explanations generated for the partner has not been explored in the context of referential communication design. We believe that it offers a potent medium to explore these issues.

Acknowledgments

This research was supported by AFOSR grant F49620-97-1-0155 given to the second author.

References

Chi, M., & Koeske, R. (1983). Network representation of a child's dinosaur knowledge. *Developmental Psychology*, 19, 29-39.

Chi, M., Hutchinson, J. E., & Robin A. F. (1989). How inferences about novel domain-related concepts can be constrained by structured knowledge. *Merrill-Palmer Quarterly*, 35(1), 27-62.

Chi, M. & Slotta, J. (1993). The ontological coherence of intuitive physics. *Cognition & Instruction*, 10, 249-260.

Chi, M., Slotta, J. D. & de Leeuw, N. (1994). From things to processes: A theory of conceptual change for learning science concepts. *Learning & Instruction*, 4(1), 27-43.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics*, 50, 66-70.

Cooke, N. J. & Breedin, S. D. (1994). Constructing naïve theories of motion on the fly. *Memory & Cognition*, 22(4), 474-493.

diSessa, A.A. (1988). Knowledge in pieces. In G.Forman & P. Pufall (Eds.), *Constructivism in computer age*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

diSessa, A. A. (1993) Toward an epistemology of physics. *Cognition & Instruction*, 10(2-3), 105-225.

Garrod, S., & Doherty, G. (1994). Conversation, co-ordination and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*. 53(3), 181-215.

Gentner, D., Brem, S., Ferguson, R., Markman, A.; Levidow, B. ; Wolff, P. Forbus, K. Analogical reasoning and conceptual change: A case study of Johannes Kepler. *Journal of the Learning Sciences*. 6(1), 1997, 3-40.

Gopnik, A., & Meltzoff, A.N. (1997). Words, thoughts, and theories. Cambridge, MA: The MIT Press.

Kuhn, D., (1991) *The skills of argument*. Cambridge University Press.

Malt, B.C. (1995). Category coherence in cross-cultural perspective. *Cognitive Psychology*. 29(2), 85-148.

McCloskey, M. (1983a). Intuitive physics. *Scientific American*, pp. 122-130.

McCloskey, M. (1983b).Naïve theories of motion. In D.Genter and A. L. Stevens (Eds.), *Mental Models*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

McCloskey, M., Caramazza, A., & Green, B. (1980). Curvilinear motion in the absence of external forces: Naïve beliefs about the motion of objects. *Science*, 210, 1139-1141.

Murphy, G. L. & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92(3), 289-316.

Ueno, N. (1993). Reconstructing P-prim theory from the viewpoint of situated cognition. *Cognition and Instruction*, 10(2-3), 239-248.

Webb, N.M. (1989). Peer interactions and learning in small groups. *International Journal of Education Research*, 13, 21-39.

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A.L. Stevens (Eds.) *Mental models*. Hillsdale, NJ: Lawrence Erlbaum Associates.