Naïve Meanings of Force: Coherence vs. Fragmentation

Andrea A. diSessa (disessa@soe.berkeley.edu)
Nicole M. Gillespie (ngillesp@uclink4.berkeley.edu)
Graduate School of Education, UC Berkeley
4533 Tolman Hall #1670, Berkeley, CA 94720 USA

Jennifer Esterly (jesterly@csustan.edu)
Department of Psychology, CSU Stanislaus
C223C, 801 West Monte Vista Avenue, Turlock, CA 95382 USA

Abstract

This paper contributes to the literature on conceptual change by engaging in direct empirical comparison of contrasting views. We take up the question of whether naïve physics ideas are coherent or fragmented, building specifically on recent work supporting claims of coherence by Ioannides and Vosniadou (2002). A partial replication of the Ioannides and Vosniadou study resulted in radically different results. We analyze several possible reasons for the differences in our results, but find that none can plausibly account for the differences in our data. We argue that the results of our study undermine claims for coherence in naïve conceptualizations of force.

Conceptual Change Research: The State of the Art

Since the constructivist revolution, there has been a fairly wide agreement that the phenomenon of naïve or intuitive conceptions in the learning of science deserves consideration. The phenomenon of naïve ideas strongly suggests that a conceptual change approach should be helpful in understanding those ideas and their trajectories during instruction. However, beyond a superficial agreement that conceptual change is an important phenomenon to understand, a huge diversity of points of view remains concerning the processes of conceptual change.

Among the fault lines in conceptual change research, one of the most contentious and probably most consequential concerns the nature of uninstructed knowledge relevant to learning particular domains, such as physics. On the one hand, some researchers contend naïve knowledge is coherent, even theory-like. For example, students may have "the impetus theory" (McCloskey, 1983), or they have one of a few models of the earth consistent with a coherent "framework theory" (Vosniadou and Brewer, 1994). On the other hand, some researchers (diSessa, 1988; Minstrell and Stimpson, 1992) argue that naïve ideas are many, diverse, and not theoretical in any strong sense of the word. In this paper, our reference theoretical frame is the Knowledge in Pieces view espoused by diSessa (1988). In this view, naïve knowledge consists significantly (but not exclusively) of hundreds or thousands of intuitive elements, which are activated in specific contexts and, as a whole, exhibit some broad systematicity, but are not systematic enough to be productively described as "a theory" or any similar term.

Aside from the intractability of deep theoretical differences, the study of conceptual change has been limited by the fact that researchers have been somewhat "spread thin," looking at a wide range of domains and issues (e.g., the shape of the earth, the effects of forces, the meaning of "alive," the distinction between heat and temperature) and a wide range of ages, from pre-school to university students. In addition, the methodologies of various researchers have involved data collection as diverse as clinical interviews, performance on physical or computer-implemented setups, and answers to paper-and-pencil questions. For the lack of common ground, it is possible different results are more the result of asking different questions, in different ways, of different subjects.

This research aims to respond to the diversity of theoretical frames and contentions about conceptual change in three ways. First, we aim to find common empirical grounds with other researchers, both in terms of age level of subjects and in terms of conceptual focus. Furthermore, we deliberately seek to minimize differences in methods. Finally, we aim to engage specific other theoretical frames and their empirical results, rather than pursuing only paths of investigation natural only to our own theoretical and empirical tradition.

In this work, we capitalize on recent study conducted by Ioannides and Vosniadou (2002), which claims that a framework theory guides and constrains the meaning children give to the word "force." Ioannides and Vosniadou ("I&V" for brevity) further claim that instruction destabilizes students' ideas, which results in increasing fragmentation of knowledge as students develop. This hypothesis of increasing fragmentation accommodates data and analysis from the Knowledge in Pieces perspective that show fragmentation and contextual dependency in older students' (high school and university) reasoning.

I&V's study is fortuitous for our purposes for several reasons. First, it covers some of the same ideas—namely force and motion—that have been the staple of Knowledge in Pieces research. Second I&V's work stems from and apparently corroborates a strong theoretical position on conceptual change favoring coherence and limited diversity

of student ideas. Finally, we believe Vosniadou's work (e.g., Vosniadou and Brewer, 1994) has been exemplary and influential with respect to the "coherence" point of view in conceptual change. To make best contact with this work, the experiment reported in this paper constitutes an attempt to partially replicate I&V's study specifically with respect to conceptual content (indeed, sharing many questions), age ranges of subjects, and also, to the extent possible, with respect to empirical methodology and subsequent analysis.

Review of Ioannides and Vosniadou Study

I&V's study involved a 27-item questionnaire in which students were asked about the existence of forces on stationary objects, stationary objects pushed by a human agent, stationary objects on top of a hill (in stable and unstable configurations—unstable meaning "it could easily fall down"), objects in free fall and objects that have been thrown. Their subjects were 105 Greek school children from a single school: 15 kindergartners, mean age 5 years 5 months; 30 fourth graders, mean age 9 years 7 months; 30 sixth graders, mean age 11 years 7 months; and 30 ninth graders, mean age 14 years 8 months. For each question, students were shown a simple drawing of an object in various contexts and were asked, "Is there a force exerted on the x? Why?" The kindergartners, however, were asked the question in the colloquial form, "Is there a force on the x?" because they did not appear to understand the "exerted" form. The questions were arranged in sets so that students would often (but not always) answer questions about the object in two states (e.g., a stone sitting on the ground and a stone falling), or different objects in the same state (e.g., a large stone sitting on the ground and a small stone sitting on the ground) and then compare the two scenarios. The students' responses were then coded based on their responses (yes, no), comparisons, and explanations.

Based on an initial analysis of the students' responses, I&V hypothesized that there were four core interpretations of force that make up the explanatory structure underlying students' understanding of force:

- *Internal Force* – an internal property of stationary objects related to size or weight
- *Acquired Force* – an acquired property of inanimate objects that explains their motion and their potential to act on other objects
- *Force of Push/Pull* – the interaction between an agent (usually animate) and an object (usually inanimate)
- *Force of Gravity* – the interaction, at a distance, between physical objects and the earth. (p. 33)

I&V then generated a pattern of responses (in terms of question-response codes) that would result if students used one of the four expected meanings of force consistently. When they compared their actual results to these expected patterns of response, they found that many students used the core internal and acquired meanings of force, but that none of the students in their sample systematically used the core meanings of force of Push or Pull or force of Gravity. However, I&V "hypothesized that they [the students] had used several [composite] meanings of force, consisting of combinations of the above-mentioned core explanatory frameworks." (p. 37) The composite meanings of force hypothesized by I&V are as follows:

- *Internal Force/Affected by Movement* – Force is due to size/weight of object only, but moving objects and objects that are likely to fall have less internal force than do stationary objects.
- *Internal/Acquired* – There is a force on stationary objects due to size/weight, and these objects acquire an additional force when they are set in motion. I&V included students in this category who were ambivalent about unstable objects and interpreted unstable objects as either lacking internal force or likely to acquire additional force.
- *Acquired and force of push/pull* – This involves the criteria described above for the Acquired meaning of force, but includes also force on an object acted on by an agent, regardless of whether it moves.
- *Force of Gravity/Other* – This composite force meaning consists mainly of the addition of gravity on to the core Acquired force meaning.

The composite meanings described by I&V are still claimed to be "internally" consistent; that is, students use each of the core ideas making up the composite meaning consistently in appropriate (different) contexts. This is in contrast to what I&V call a "Mixed" meaning of force, which was used to describe students who did not use the core or composite meanings of force consistently across the set of questions asked.

I&V predicted that each student's set of responses could be characterized as being indicative of the student's having and using either a core meaning of force or a composite meaning of force consisting of a combination of core meanings. In order to test this prediction, they again established a mapping from students' response codes to the various meaning of force and then compared the actual responses to these mappings.

The results of I&V's study are presented in Table 1. I&V do not explicitly state their mapping criterion, but we believe they assigned students to a particular category only if *all* answers were consistent with those projected by a meaning. In this regard, their coding had a certain degree of "softness" built in. For example, students were assigned to the Internal meaning if they said the small object had either a small or no force on it. Students assigned to Internal/Affected by Movement and Internal/Acquired meanings were allowed different interpretations as to whether instability (in contrast to stability) should imply the existence of a force, or the opposite. Students assigned to the Gravity/Other meaning were allowed to not mention gravity in agentive pushing situations.

Table 1 - Summary of I&V data, frequencies of Meaning of Force as a function of grade

Force Meaning	K	4th	6th	9th	Tot
Internal	7	4			11
Internal/Affected by Movement	2	2			4
Internal/Acquired	4	10	9	1	24
Acquired		5	11	2	18
Acquired/Force of push/pull			5	10	15
Force of Push/Pull				1	1
Gravity/Other		3	1	16	20
Mixed	2	6	4		12

I&V's results are striking and apparently offer compelling support for their claims of the existence of framework theories that guide and constrain children's understanding of the concept of force into a small number of consistently used interpretations. They remark that almost 90% of subjects "made use of a small number of relatively well-defined and internally consistent interpretations of force" (p. 5). Furthermore, of the seven meanings of force that most subjects apparently used, all meanings were combinations or variations of four core meanings (two of which are uninstructed, and two of which emerge, it is claimed, in interaction with instruction). The seventh meaning of force specified by I&V "gravitational and other," contains components of those same four core meanings.

A second key result of the I&V study is the trend towards older students using composite meanings. I&V use this finding to support their claims that older students exhibit "increasing fragmentation" as they are exposed to instruction and allows the framework theory hypothesis to be more consistent with data from older students that supports the Knowledge in Pieces hypothesis. However, note that none of the ninth graders used internally inconsistent mixed models, a finding that appears to be somewhat at odds with I&V's hypothesis of increasing fragmentation.

Our Study

The rationale behind our study was to replicate I&V's study in substance in order to analyze and evaluate similar data. While I&V only asked their subjects about forces, we alternated asking about forces with asking about pushes and pulls (balanced across subjects) in order to include some linguistic diversity. In our view, it is almost certain that both of these linguistic formulations connect fairly directly to intuitive resources that are implicated in the conceptual development of the technical concept of force. Thus, it is more valid to explore both meanings for intuitive conceptualizations relevant to conceptual change than to consider only one. Additionally, we asked a set of 14 unique questions ("extension study") later in the interview that was designed to expand on the contexts investigated by I&V. We will not report on the results of the extension study here.

Methods

Our experiment consisted of a series of 10 question sets presented in a clinical-interview format to a total of 30 students: 9 pre-school (mean age 5 years 1 month), 9 elementary (mean age 7 years 8 months), 6 middle school (mean age 12 years six months) and 6 high school (mean age 15 years 11 months.) Rather than attempting to duplicate the ages or grades of I&V's subjects exactly, we chose to interview subjects from approximately the same age range and years of schooling as did I&V, as differences in Greek and U.S. schooling made close replication of the subject pool in terms of ages, grades and level of instruction problematic. None of the high school students we interviewed had taken a physics class at the time of the interview. None of the younger students indicated that they remembered learning about force in school.

The questions we asked were replicates of a representative subset of I&V's questions, spanning the range of I&V's categories (i.e., stationary objects, stationary objects pushed by a human agent, stationary objects on top of a hill objects in free fall, and objects that have been thrown) and duplicated (as closely as possible) the drawings used by I&V in their study for the same question sets.

The primary differences between our study and the I&V study are as follows:

- We eliminated a few of the less important dimensions of the study to keep size manageable (in view of our wanting to ask additional questions as part of an extension study).
- We asked students to compare amount of force in each of 10 sets of questions, whereas I&V asked only sporadically about comparisons.
- Our questions were in a slightly different order.
- We asked each subject alternately about force, or push/pull in each of the 10 question sets, and orders were balanced across subjects.
- Our interviews were conducted in English rather than Greek, and we used the phrasing "Is there a force/push/pull on…" for all subjects.

We developed a coding scheme for our data that included three primary aspects: 1) whether subjects indicated there was a force (or pushing or pulling) on the focal object, 2) what was the nature of the force on the focal object (e.g., inherent in the object, applied by another object or person, gravitational, etc.) and what other object was involved, and 3) judgment of comparative strength. We did not include consideration of explanations, as I&V did, as we felt the difficulty of interpreting these would leave us open to a critique of bias. As such, our coding was more forgiving ("softer") and less sensitive to fragmentation, which should have favored a result of "coherence" in our results.

To test intercoder reliability, two coders independently coded a representative set of 6 subjects (two each from elementary, middle and high school). We had 99.4%

intercoder agreement on codes relating to existence and comparison (158 out of 159) and 98% agreement (265 out of 270) agreement on any codes that were used in our mapping to I&V meanings. Disagreements were mainly cases where one coder agreed s/he had made a mistake, or attempts to code situations where one coder felt the existing codes were inadequate to capture the meaning expressed by the subject.

Since we did not use precisely the same codes as I&V (they coded explanations, and integrated the different questions in a set into an overall code, whereas we did not code explanations and independently coded each aspect), we developed our own mapping from our codes to I&V's meanings. However, we rigorously maintained every softness I&V incorporated into their coding (e.g., we allowed subjects to respond "no force" when the contrast attribute (size) was small). We also independently checked to make sure our mappings were consistent with the textual descriptions of the meanings, and with the specific codes I&V used for each set, omitting explanations.

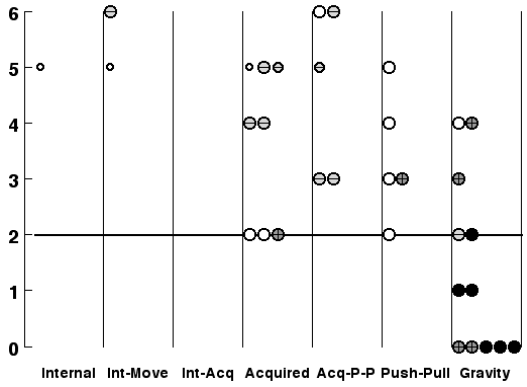
Results

After coding subjects, we computed a *meaning deviation score (*MDS), the number of sets (out of 10) on which each subject's responses did not match the allowed responses for each meaning. Then, we assigned each individual to the meaning on which they achieved lowest deviation. Figure 3 shows the results. The vertical bands, left to right, show the various meanings of I&V. Each subject shows up as one or more circles. Pre-school subject show as empty circles, elementary school subjects as light gray-filled circles (with a horizontal line), middle school subjects as dark gray-filled circles (with a cross), and high school student as black disks. If a subject matched two or more meanings to the same degree of deviation, we split the corresponding marker into smaller circles. At the level of 5 misses out of 10, one middle school subject equally matched Acquired and Acquired/Push-Pull, and one pre-school subject matched three meanings. For comparison purposes, note that all of I&V's subjects who were not assigned to the Mixed meaning could be assigned unambiguously to one meaning with a MDS of zero.

Only 5 of the 30 subjects fully matched the meaning specification of any meaning (compared to about 90% claimed by I&V), and all of these were on the Gravitaty/Other meaning. In general, we felt this meaning was too ambiguous to be diagnostic in terms of what subjects thought. In particular, because we left out explanations in our coding, matching Gravity/Other meant essentially only that they mentioned gravity as a force in every case. This issue is followed up on in the extension study, reported elsewhere.

The dark line drawn at the level of MDS=2 represents a somewhat arbitrary but also generous allowance for mistakes subjects might have made (i.e., a 20% error rate). In pursuing the question of whether I&V meanings really capture the reasoning of subjects, we will concentrate on subjects with 2 or fewer mismatches. Furthermore, we will also attend more to the elementary, middle and high school subjects for several reasons. In general, we found coding the preschool students' responses quite challenging, and we are much less confident that they are meaningful. To further emphasize the uncertainty of coding of pre-school subjects, we note that they are spread widely across the meanings (unlike I&V's results), while elementary, middle, and high school students show a fairly clear, if "smudged" developmental drift in the direction found by I&V. Finally, we note that the age of our pre-school students averaged about two years less than the youngest group of I&V, suggesting less comparability, and perhaps pushing the interview into a less sensible age range. For brevity of reference, we call the full set of subjects PEMH (pre-school, elementary, middle school, high school), and the reduced set, without pre-school subjects, EMH. Even with the generous allowance for mistakes, only 10 of 21 EMH subjects match any meaning, and 9 of these 10 match the problematic category, Gravity/Other.

Figure 3. Individual's meaning deviation score as a function of meaning.



Using a two-tailed z-test of the difference in population proportions, we find the probability that our and I&V's data could have come from the same population to be exceedingly small ($p<10^{-6}$). Given the dramatic difference in outcome of our study compared to I&V, we undertook several analyses to see if we could determine the source of the difference.

The Effect of Language: Force vs. Push/Pull

We looked at the question of whether the linguistic variety we introduced could have affected the outcome so substantially. Luckily, redundancy in our questions provided an opportunity to investigate this question. Three separate questions ask about the forces or push/pulls in one situation (a large rock on the ground) and two in another case (unstable rock on a hill). Because we alternated force and push/pull questions, we could compare responses. In the case of the rock on the ground we asked about forces in one place and push/pull in another, and on a third occasion we duplicated one of these two forms. Of the 21 EMH subjects, only 2 (9.5%) switched responses between the force and push/pull versions of the question. None of the elementary school

students and none of the high school students gave different responses to the different phrasings. The two middle school students who switched, switched oppositely, giving true and false, and false and true, respectively, to the force and push/pull versions of the question.

Even more telling, 4 out of 21 EMH subjects (19%) gave inconsistent answers, that is, they changed their answer to *identically phrased* questions asked at different time in the interview. Whether these were mistakes or involved unarticulated contextual factors, we are in no position to say. Among the pre-school subjects, 4 out of 9 (44%) gave contrasting answers to force or push/pull versions, while 3 out of 9 (33%) gave inconsistent answers to identical questions asked at different times.

The second question that was phrased in both force and push-pull forms yielded similar results. Three out of 21 (14%) EMH students gave different answers to different phrasing of the same question. Again, no high school students changed their answers. Five of 9 (55%) pre-school students switched answers.

Pooling these data suggests that only 10-15% of EMH subjects (skewed toward younger ages) switch between force and push-pull versions, and this is about the same as the number of switches that occurred between identically phrased questions at different times, either because of mistakes or unaccounted-for contextual factors. In addition 13 out of 18 (72%) EMH subjects who were explicitly queried about the difference between forces and push-pulls said either that they were identical, or that push-pulls were a kind of force (implying either that forces divided exclusively into pushes and pulls, or that there might be other kinds of forces).

To examine the possibility of a trend, we classified EMH subjects (mapped their responses to meanings) on the basis of their responses only to questions phrased in terms of forces. Three more subjects became classifiable as belonging to one of I&V's meanings within the 20% error range (i.e., a MDS≤1 out of 5 question sets). However, two other subjects fell out of the allowed error range, yielding a net change from 13 to 14 classifiable subjects. Furthermore, only 3 of 21 subjects changed which meaning they matched best.[1] Mapping subjects only on the basis of their answers to push-pull questions yielded similar results compared to restricting to force questions. Three more subjects became classifiable at the 20% error level, while one moved out of the allowed 20% range. No subjects changed the meaning they matched best compared to using the full data set.

To summarize this line of investigation, although we lack the statistical power to map in any detail the effect of asking about pushes and pulls, as opposed to forces, there is no indication of any substantial effect. On the contrary, it appears that any effect is roughly on the order of "noise" as measured by changes in response to identically phrased questions at different points in the interview.

The Effect of Asking Comparisons

One of the main differences between our question set and I&V's is that we asked more comparison questions. To test whether this difference could account for the disparity between our data and theirs, we repeated our analysis, eliminating the comparison questions that we asked and I&V did not. No substantial changes resulted. No subject changed best match meaning.[2] More telling, no subjects moved into the allowed 20% error range; no subject left the 20% range, and there was only one subject who changed deviation within the allowed 20% range, from MDS=2 to MDS=1.

Subject-to-Subject Correlation and "Hidden Meanings"

We performed one final analysis of our data to look at the amount of diversity in the data—how subjects correlated with each other—and to check for the possibility of "hidden meanings," that is, meanings other than I&V's that could potentially explain the difference between our data and theirs. For these purposes, we established a "meaning" ("model" is a better rendition in this case) according to the particular pattern of answers given by each subject. This eliminates the softness allowed by I&V mappings, for example, they allowed subjects to categorize a force as small or zero with no penalty. However, in the absence of starting with semantic models, we have no basis for allowing ambiguity. Finally, we matched each subject to the models produced by every other subject. We found only 11 matches of any subject to any other subject, and only one of these was better than the 2 allowed deviations. For calibration, assuming every subject matched one of seven models (following I&V's lead), the minimum number of matches would be 55. Even if we allow 10% "mixed" meanings, which would not match any model (again following I&V), one would have to have *at least* 39 matches among the remaining subjects to be consistent with there being only 7 meanings of force. By contrast, the number of matches in our data is more consistent with from 15 to 20 models. Given that the number of subjects is 30, an estimate of 15-20 meanings of force does not support the possibility that there are "hidden" meanings of force that could explain the difference between our data and I&V's. Still, our data are clearly not random. A Monte Carlo simulation involving random choices of response yielded an expectation of about .02 matches between 30 subjects per experiment.

There is no way definitively to rule out other meanings (including sensible ambiguities) that might make sense of this data. Indeed, with the degree of ambiguity in mappings allowed by I&V (re-inserting the kind of ambiguity we took out, above), it is likely, if not certain, that one could find patterns of answers that would match a data set of the size

[1] For these purposes, we ignore the fact that some meanings might move in or out of a tie for minimum deviation in case of multiple matches.

[2] Again, we ignored changes in the number of tied meanings.

they used. Unmotivated patterns of answers, however, make poor candidates for "meanings."

Discussion

The principal results of our study, shown in Figure 1, are strikingly at odds with I&V's results. Further analysis of our data suggested that obvious candidates for underlying causes of difference don't pan out. While these results are strongly in our favor, they surprised us. In reaction, we double-checked our codings, our mappings, and the program that provided the mapping for us. No problems were found. We do not understand the reasons for the discrepancy. The following are several conjectures we are working on:

Interviewing technique. It is possible that we gave students more time to answer or to rethink their original answers, or provided implicit guidance to the point that multiple ideas were evoked that were not evoked at I&V's pace. To some extent, with pacing at least, this is likely to be a consequential difference. However, it seems implausible that it could account for such dramatic differences in our data.

Coding. In contrast to our coding, I&V did not independently code different items in a set (e.g., forces on two objects plus comparison), but assigned codes based on a pattern of answers for the entire set. This might, in principle, have biased their analysis toward a "coherence" result. However, I&V coded explanations in addition to the features we coded (existence, relative size, and source), which should have inevitably resulted in students' appearing *less* coherent than using our methods.

Instructional differences. Few U.S. students receive any systematic instruction in force and motion before high school, and none of our subject reported any. In contrast, our understanding is that in Greece and in other European countries, force and motion is introduced much earlier. On the other hand, it is not at all clear this could have had a large effect, given how far from normatively correct all of our and I&V's subjects were.

Language. Perhaps the most intriguing possibility emerges from the observation that the Greek language is different than English. In particular, the Greek word for force, *dynamis*, connotes strength and power, in addition to being the technical term for physicists' force. The English word "force" has colloquial uses that include a "police force," "forces of nature", "forcing" someone to do something, the "force" of someone's anger, and so on. If differences in language account for differences in conceptualization enough to explain the divergence in our results, we might have established a striking effect of linguistic/cultural relativity.

Conclusions

Strictly speaking, our attempts at replicating the I&V study failed. Our results are statistically exceedingly far from theirs. Rather than 90% of subjects classifiable as having a consistent meaning of force, less than 50% of our elementary, middle and high school subjects were classifiable, even allowing an error rate of 2 questions out of 10.

Our results undermine I&V's claims that their 7 meanings constitute a full specification of the naïve concept of force. From a Knowledge in Pieces perspective, finding data that suggests I&V's experiment won't replicate, and that the systematic difficulty in the replication is inherent diversity in subjects' responses, is congenial. Nevertheless, the stark contrast between our data and theirs is unsettling and warrants further study.

While the empirical work here should give pause to "coherence" advocates, we caution against drawing too extreme conclusions from this study, primarily because we do not believe conceptual change is a homogeneous phenomenon. The Knowledge in Pieces perspective was developed specifically to deal with experientially rich domains, such as mechanics. Thus, results outside mechanics favoring the coherence view, including earlier work by Vosniadou, are not immediately threatened by the results of this work.

Acknowledgements

We are indebted to Stella Vosniadou and Christos Ioannides for putting their theories on the line with well-documented empirical data. We are grateful to Dave Kaufman, Joe Wagner for help in formulating our empirical plan and Barbara White, Orit Parnafes, Nathaniel Brown, Michael Ranney, Flavio Azevedo, and Ann Ryu for feedback on paper drafts. We would also like to thank the Berkeley Institute for Human Development, Marie-Eve Thomas of Hoover Elementary School and Suzy Loper of Longfellow Middle School for their assistance in arranging subjects for this study.

References

diSessa, A. A. (1988). Knowledge in pieces. *Constructivism in the computer age.* G. Forman et al Eds. Hillsdale, NJ: Lawrence Erlbaum Associates: 49-70.

Ioannides, C. and S. Vosniadou (2002). Exploring the Changing Meanings of Force: From Coherence to Fragmentation. *Cognitive Science Quarterly, 2*(1): 5-61.

McCloskey, M. (1983). Naive Theories of Motion. *Mental models*. D. Gentner and A. L. Stevens Eds. Hillsdale, N.J., Erlbaum: 299-324.

Minstrell, J. and V. Stimpson (1992). *Creating an Environment for Restructuring Understanding and Reasoning*. Paper presented at the Conference on Curriculum and Assessment Reform in Education (Boulder, CO, June 1992).

Vosniadou, S. and W. F. Brewer (1994). Mental models of the day/night cycle. *Cognitive Science 18*(1): 123-183.