Cognitive Science 28 (2004) 843–900

COGNITIVE SCIENCE

http://www.elsevier.com/locate/cogsci

Coherence versus fragmentation in the development of the concept of force

Andrea A. diSessa[a,*], Nicole M. Gillespie[a], Jennifer B. Esterly[b]

[a] *Graduate School of Education, University of California, 4647 Tolman Hall, Berkeley, CA 94720, USA*
[b] *Department of Psychology, California State University, Stanislaus, Turlock, CA 95382, USA*

Received 31 January 2003; received in revised form 30 March 2004; accepted 3 May 2004
Available online 27 September 2004

Abstract

This article aims to contribute to the literature on conceptual change by engaging in direct theoretical and empirical comparison of contrasting views. We take up the question of whether naïve physical ideas are coherent or fragmented, building specifically on recent work supporting claims of coherence with respect to the concept of force by Ioannides and Vosniadou [Ioannides, C., & Vosniadou, C. (2002). The changing meanings of force. *Cognitive Science Quarterly 2*, 5–61]. We first engage in a theoretical inquiry on the nature of coherence and fragmentation, concluding that these terms are not well-defined, and proposing a set of issues that may be better specified. The issues have to do with *contextuality*, which concerns the range of contexts in which a concept (meaning, model, theory) applies, and *relational structure*, which is how elements of a concept (meaning, model, or theory) relate to one another. We further propose an enhanced theoretical and empirical accountability for what and how much one needs to say in order to have specified a concept. Vague specification of the meaning of a concept can lead to many kinds of difficulties.

Empirically, we conducted two studies. A study patterned closely on Ioannides and Vosniadou's work (which we call a *quasi-replication*) failed to confirm their operationalizations of "coherent." An extension study, based on a more encompassing specification of the concept of force, showed three kinds of results: (1) Subjects attend to more features than mentioned by Ioannides and Vosniadou, and they changed answers systematically based on these features; (2) We found substantial differences in the way subjects thought about the new contexts we asked about, which undermined claims for homogeneity within even the category of subjects (having one particular meaning associated with "force") that best survived our quasi-replication; (3) We found much reasoning of subjects about forces that cannot be accounted for by the meanings specified by Ioannides and Vosniadou. All in all, we argue that, with a greater attention to contextuality and with an appropriately broad specification of the meaning of a

[*] Corresponding author.
E-mail address: disessa@soe.berkeley.edu (A.A. diSessa).

doi:10.1016/j.cogsci.2004.05.003

concept like force, Ioannides and Vosniadou's claims to have demonstrated coherence seem strongly undermined. Students' ideas are not random and chaotic; but neither are they simply described and strongly systematic.

Keywords: Conceptual change; Coherence; Naïve physics

1. Introduction

1.1. Problematic

Since the constructivist revolution in studies of learning, there has been wide agreement that the phenomenon of naïve or intuitive conceptions in science learning deserves consideration. Researchers have documented apparently uninstructed—or at least non-normative—student ideas and the persistence and untoward effects of such ideas long into the instructional process. "Misconceptions" or alternative conceptions have been well-documented in fields such as physics, biology, programming, astronomy, and statistical reasoning (Carey, 1985; Confrey, 1990; Kahneman, Slovic, & Tversky, 1982).

The influence of naïve ideas in learning 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 views reigns concerning the entities and processes of conceptual change. Some researchers liken students' conceptual change to revolutions in professional science (Carey & Smith, 1995). Others deny or downplay this contention (diSessa, 1988). For some, the central theoretically explanatory idea is "concepts" (Carey, 1999), while others locate the persistence of naïve conceptions at different levels, in "theories" (Gopnik & Wellman, 1994; Wellman & Gelman, 1992) or "ontologies" (Chi, 1992). Even when there is agreement that, for example, concepts constitute the right level of explanation, a wide range of assumptions underlies exactly what constitutes a concept and how one should view conceptual change (diSessa & Sherin, 1998).

To compound the difficulty, there has been limited crosstalk among the different camps. With few exceptions, focused competitive argumentation in conceptual change research is limited.

Aside from the intractability of deep theoretical differences, the study of conceptual change has been limited by the fact that research has been spread somewhat thin across a wide-ranging set of domains (e.g., the shape of the earth, the effects of forces, the meaning of "alive," the distinction between heat and temperature) and across a wide range of ages, from preschool to university students. In addition, the methodologies of various researchers have involved data collection as diverse as clinical interviews, performance in physical or computer-implemented setups, and answers to paper-and-pencil questions. Analysis of data ranges from "code and count" to qualitative theory building. For lack of common ground, it is possible that different results are more the result of asking different questions, in different ways, of different subjects. Indeed, there is precious little argument, let alone convincing data, that conceptual change is a homogeneous phenomenon.

This research aims to respond to the diversity of theoretical frames, conceptual foci, and methodologies in several ways.

1. We aim explicitly to articulate and explore an important and broad division among conceptual change theorists. We wish to contribute to the relatively sparse body of literature that self-consciously contrasts different views, and pursues an avenue intended to bring the debate to conclusion.
2. We aim to find common empirical grounds with other researchers, both in terms of age level of subjects and in terms of conceptual focus.
3. We deliberately seek to minimize differences in methods, rather than pursuing paths of investigation natural only to our own theoretical and empirical tradition.

1.2. *Coherence versus fragmentation*

Among the fault lines in conceptual change research, one of the most contentious and probably among the most consequential concerns the nature of uninstructed knowledge. On the one hand, some researchers contend naïve knowledge is coherent—even theory-like—and is *compactly characterizable*. For example, students may have "the impetus theory" (McCloskey, 1983) that competes essentially head-to-head with Newtonian theory in learning elementary physics. "Compactly characterizable" means that relatively little needs to be said to describe a naïve theory or concept. To illustrate, consider McCloskey's definition of the naïve impetus theory:

> First, the theory asserts that the act of setting an object in motion imparts to the object an internal force or "impetus" that serves to maintain the motion. Second, the theory assumes that a moving object's impetus gradually dissipates (either spontaneously or as a result of external influences), and as a consequence the object gradually slows down and comes to a stop. (p. 306)

Carey (1999) and Gopnik and Wellman (1994) are among the most visible and articulate advocates of the "Theory Theory," the claim that naïve ideas are insightfully characterized as theories that are similar to scientists' theories in important ways. We note, however, that there is a diversity of opinions about the nature of naïve theories. On the conservative side, Vosniadou (2002) lists dimensions along which expert theories and naïve theories differ. For example, Vosniadou contends that the theories of scientists differ from naïve theories in terms of articulateness, external representation, meta-conceptual awareness (scientists know they have theories), and so on. Other researchers, while acknowledging some difference between professional and naïve theories, stress their deep similarity not only with respect to form and function, but with respect to development as well (Gopnik & Wellman, 1994). Despite differences among Theory Theorists, we believe they are as a group committed to a fairly high degree of coherence and consistency in naïve knowledge. In this article, we do not inquire exactly what the meaning of "theory" might be, but we fall back on and investigate the weaker presumption of "coherence."

On the other side of the coherence versus fragmentation debate, a complementary set of researchers argue that naïve ideas are many, diverse, and not theoretical in any deep sense. Minstrell speaks of *facets* (Minstrell & Stimpson, 1996), which are relatively independent explanatory facts, such as "heavier things fall faster (because of their weight)." Minstrell and colleagues track hundreds of facets in elementary physics, for example (Hunt & Minstrell, 1994). diSessa (1988, 1993), speaking of "Knowledge in Pieces," argues that intuitive physics consists substantially of hundreds or thousands of inarticulate explanatory primitives, which are activated in specific contexts and, as a whole, exhibit some broad systematicity, but are not deeply systematic enough to be productively described individually or collectively as a theory. In neither case, facets nor p-prims, is the intuitive knowledge system compactly describable if for no other reason[1] than the number of relatively independent elements involved.

Attending to the coherence versus fragmentation dispute is advantageous for a number of reasons. First, these views differ starkly in their characterization of intuitive knowledge. It is almost difficult to believe they both have survived empirically as long as they have. Furthermore, this dispute is relatively independent of many of the details (such as what, exactly, changes; how change is modeled) that cloud direct comparison of specific competitor theories of intuitive knowledge and conceptual change. It therefore makes a good avenue to pursue in sorting out the larger field. Finally—although we do not argue it extensively—these different characterizations of the naïve state would seem to have substantially different implications for educational strategies. For example, the "fragmented" view suggests extended collection and organization of elements along the path to expertise, and it has much more room for individual variation. In contrast, the "coherent" view suggests that debate and rational argument among a few alternatives might be effective.[2]

Although, as we suggested, direct competitive argument between these positions is somewhat limited, there have been prior attempts to look at relevant issues. O'Malley and Draper (1992)—in a book that systematically highlights the fragmented versus coherent issue—consider the area of users' conceptions of complex artifacts, such as computer systems. On the whole they side with a fragmented view, largely because they believe interaction with a device obviates the need for complete, consistent models. However, this work provides no focused empirical support, it does not specifically track the nature of naïve knowledge, and it has the disadvantage that there seem to be few who strongly adhere to the "coherent and consistent" view in this conceptual domain.

Clark (2000, 2003) did a longitudinal study of students learning basic elements of heat and temperature by tracking facet-like statements in interviews involving problem solving and explanation. His conclusions stand firmly on the "fragmented" side of the debate. Facets waxed and waned in students' reasoning, showed tentative connections to other elements, which connections grew only gradually and somewhat erratically. This contrasts markedly with prior claims about the development of heat and temperature knowledge (e.g., Wiser, 1987, 1995). These opposing views of the development of heat and temperature knowledge, however, involve contrasting methods of investigation and analysis, which confound direct comparison.

Closer to the focus of the present work, Anderson, Tolmie, Howe, Mayes, and Mackenzie (1992) looked at the issue of coherence by systematically varying parameters, such as mass and velocity, across tasks that instigated McCloskey's Theory Theory claims. Their results were somewhat mixed. With respect to predictions, students' responses showed significant context sensitivity. With respect to explanations, however, Anderson et al. found that students showed relative coherence and consistency. Following this line of work with similar questions, Cooke and Breedin (1994) claimed to show fairly dramatic inconsistency in student responses, at least with respect to impetus "theory" predictions. Ranney (1994), however, offered a critique of Cooke and Breedin's methodology. While contesting their methods, Ranney does not particularly question the conclusions. In fact, his own work in this area plays out mainly in the direction of "fragmentation." For example, Ranney (1987) [reported in Ranney, 1996] investigated subjects' predictions across isomorphic situations. In one case, comparing a pendulum bob release versus the release of a human (trapeze artist) in mid-swing, only 20% of subjects gave matched predictions to the two situations. Even when subjects professed isomorphism (pronounced the situations "fundamentally similar"), only 31% of their drawn predictions of trajectory were the same for both cases. The subject ages and conceptual focus of the references in this paragraph, by and large, coincide. However, methodology is evidently in dispute.

Thagard (1992, 2000) is responsible for an extended line of work concerning coherence in conceptual change. In short, he argues that explanatory coherence is an important driving force in such processes as theory change. However, there has been relatively little empirical work following his framework in conceptual change during school learning (as opposed to conceptual change in the history of science), and his framework also does not supply ready non-empirical answers to the question of relative coherence of naïve conceptions in school science.

A final reference is particularly relevant to the coherence versus fragmentation debate as approached here. Samarapungavan and Wiers (1997) substantially follow the methodology and theoretical frameworks of Vosniadou, whose work will be reviewed in detail shortly. Although in the domain of evolutionary biology rather than physics, their results follow the pattern established earlier in other domains by Vosniadou (Vosniadou & Brewer, 1992, 1994). Roughly 80% of their subjects gave answers to a semi-structured interview that implicated a coherent, consistently applied, and compactly describable "explanatory framework." Furthermore, there were only a few such explanatory frameworks. Samarapungavan and Wiers specifically contrast their results with diSessa's Knowledge in Pieces point of view, although they also comment on the difficulty of comparison due to different conceptual domains and methodologies.

1.3. Focus on elementary mechanics—force and motion

The work in this article attends to naïve conceptions in elementary mechanics for multiple reasons. First, mechanics might be the single most researched topic of naïve conceptions. A firmer research base means more likely progress with respect to fundamental issues. In addition, there are strong advocates of both coherence and fragmentation in this domain. On the whole, adherents of coherence have had the upper hand in the literature, beginning relatively early in the constructivist revolution. In a review of naïve conceptions in science learning, an important contributor to the study of science education, Rosalind Driver (1989), reported a near consensus favoring a coherent, theory-like view of naïve conceptions. However, in a short section on "alternative perspectives," she briefly mentions Knowledge in Pieces and situated cognition. One of the most influential papers in the field of science education in the 1980s and early 1990s (Posner, Strike, Hewson, & Gertzog, 1982) drew heavily on a theory-change view that emphasizes parallels to the history of science. Certainly McCloskey's very widely cited work (McCloskey, 1983) enhanced the popularity of the Theory Theory view, and it has not shown a significant waning since (e.g., Carey, 1999). While Thagard (1992) mentions Knowledge in Pieces concerning childhood conceptions, he acquiesces to arguments by Carey, Wellman, and others that children's ideas can be profitably viewed as theories. In view of this leaning in the literature, a critical study of Theory Theory claims from a "fragmented" point of view is in little danger of appearing to tackle an easy target.

In this work, we capitalize on a recent study conducted by Ioannides and Vosniadou (2002), which for brevity we will refer to as "I&V." The 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 our own research. Recall that the first aim of this work was to join the coherence versus fragmentation debate on common conceptual ground. Second, I&V's work stems from and apparently corroborates an elaborated theoretical position on conceptual change favoring coherence of student ideas. In this regard, the results of the study are striking and challenging. In a cross-sectional study of 105 subjects from kindergarten to high school, they found only a small universe of meanings for "force": four primary meanings, and a smaller number of coherent but composite meanings. Almost 90% of subjects were unambiguously classifiable as consistently using exactly one of these meanings across a battery of 27 questions about basic situations of force and motion.

Vosniadou's work has been widely cited and influential with respect to the "coherence" point of view in conceptual change. The work has a relatively long line in multiple domains, such as the day/night cycle (Vosniadou & Brewer, 1994), children's models of the earth (Vosniadou & Brewer, 1992), and, now, elementary mechanics. As mentioned earlier, Vosniadou is not an adherent of an extreme Theory Theory position. Rather, with respect to claims that students' ideas are theoretical, she maintains that one should admit important differences between naïve theories and professional ones. However, she holds to the idea that student conceptions are coherent, limited in number,[3] and compactly characterizable.[4] Furthermore, her theoretical position is well-developed, having grown through years of study. Finally, Vosniadou (with Brewer) developed and consistently applied a well-structured methodology for tracking students' ways of conceiving of various domains. We explain Vosniadou's method of "model mapping" (our term) later in this article.

To satisfy the second and third aims listed earlier—namely to retain common focus and to avoid methodological disparity—the experiments reported in this article constitute an attempt to follow closely in Ioannides' and Vosniadou's footsteps, and then to extend their work slightly, specifically preserving conceptual content (indeed, sharing many questions), age ranges of subjects, and also, in the quasi-replication, empirical methodology and subsequent analysis.

1.4. Plan for the work and paper

We begin the body of this paper with a broader theoretical framing of the issues. While "coherence versus fragmentation" names a salient dispute in the literature, a significant difficulty lies in lack of a common language and common assumptions about the larger field of possibilities within which this dispute lies. Obviously, agreeing on common meanings is important, and those meanings need to be cogent and consequential. Furthermore, without attention to the larger field, researchers may inadvertently attend to limited aspects of conceptual performance and draw conclusions that do not extend to the full range of competence. A prototypical issue of this sort is how diverse and extensive should experimental probes be to establish the character of the implied knowledge (e.g., as fragmentary or coherent)?

Thus, the first aim of our theoretical section is to do some groundwork with respect to this larger field—that is, with respect to a cogent set of descriptors of knowledge systems. In particular, we develop two issues of concern. The issues are related to one another, but are theoretically distinct. The first issue, *contextuality*, concerns how students reason in different contexts. Within the larger framing of contextuality, fragmented versus coherent only roughly and partially specifies possible regimes of conceptual knowledge. That is, two conceptual systems may have consequentially different properties with respect to contextuality, and yet both may still be aptly characterized as either "fragmented" or "coherent." Furthermore, it is not entirely clear which conditions of contextuality should count as fragmented or coherent.

The second issue is *specification*. How much and what kind of information should a researcher present in order to characterize naïve knowledge adequately? The turning point here is that, unless we say enough about the knowledge that we attribute to people, we are in danger of misidentifying performance as belonging to one particular point of view (e.g., that pertaining to a naïve theory) simply because we have suppressed important details that implicate the use of multiple alternative conceptualizations.

After elaborating and justifying these general theoretical considerations, we go on to describe the two particular theoretical standpoints implicated in this study: the Knowledge in Pieces perspective of diSessa, and the Framework Theories view of Vosniadou. Neither view is well-characterized only by a position on the fragmentation versus coherence debate, and it is important not to reduce the diversity in the larger debate to a single dimension. Furthermore, the implications of either view with respect to fragmentation versus coherence are somewhat subtle, and stronger with respect to some issues of contextuality and specification than with respect to others. In looking beyond the narrower aims of this study, obviously, we want to test theories on their core entailments, not on marginal or only heuristically induced expectations. This discussion also opens up some uncertain aspects of the empirical study, specifically with respect to the role of language as reflecting conceptualization.

After theory (Section 2), we review the logic, methods, and results of Ioannides and Vosniadou's study (Section 3). This is important to give readers a basis for comparing the empirical work we do here, which is described in some detail in the following Section 4. As mentioned, our empirical work is divided into a "quasi-replication," which is intended to match Ioannides and Vosniadou's experiment closely, and an "extension study." The latter is intended to explicate issues of contextuality and specification that are not adequately handled in the original I&V study or in our quasi-replication.

The following two Sections 5 and 6 individually treat the results of the quasi-replication and of the extension study. The remaining Sections 7 and 8 review the logic and results of this work and suggest what seem to be profitable future avenues of study.

2. Theoretical issues: contextuality and specification

2.1. Contextuality

We begin with a crude first pass. Context is a central concept in the debate between advocates of coherence and advocates of fragmentation. Broadly speaking, views that advocate understanding naïve ideas as theory-like expect that they cover a relatively wide scope, and that they will be few in number, perhaps only one per individual per domain. These expectations mirror the usage of "theory" as it applies to professional science: Generalization is at the heart of theorizing, and if a "theory" covers only a tiny range of specific contexts, one would scarcely use the word. One simply does not have a scientific theory that treats only what happened to THAT glass, as it slipped off THAT table. Similarly, if one discovers that individuals have, say, 30 distinct ways of thinking about a domain, even if those ideas have distinct contexts of use (so that they do not conflict with one another in any specific application), one would scarcely say individuals have a theory of the domain.

We now proceed by degrees to elaborate this first pass. However, our intention is far short of producing a model or theory of contextuality. Instead, we aim only at a first-level refinement of the brief exposition above to show what kind of issues may arise and to prepare for some that will arise. In all this, we assume that we have both intuitive and related professional or instructed knowledge systems to compare.

Fig. 1a depicts what may be the ideal case for Theory Theory views of naïve conceptions. The straight-edged region marks "the context of application," in real-world configurations, of a conventional scientific idea, say, Newtonian mechanics. (However, we may use the same kind of diagram to depict contexts of application of any part of a scientific theory, not just "the whole theory." We use the generic term "element" to denote any of many possible parts of a conceptualization, including concepts or other such knowledge classes.) The rounded edge marks the range of application of some intuitive or naïve way of thinking about the mechanical world. It might correspond to "the intuitive theory of force and motion," if such exists. Both views cover substantially the same range of contexts, although it would be absurd to think the ranges exactly coincide.

Fig. 1b depicts a parody of a fragmented view, where the scientific domain is "covered" by many intuitive elements that are much smaller in their range of application. These elements might represent Minstrell's facets or diSessa's p-prims. We say this is a parody because it suggests things that we do not expect to be uniformly true of "intuitive fragmentation," and

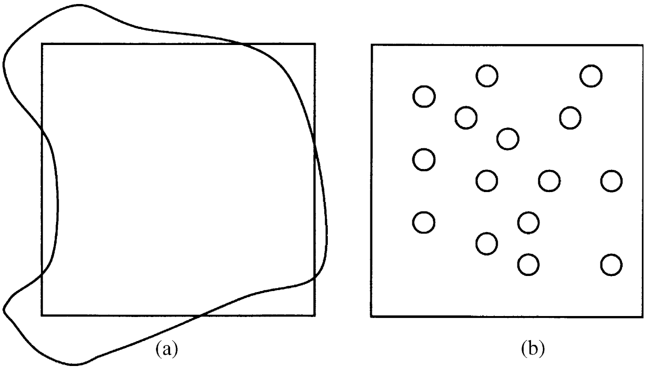


Fig. 1. (a) The contextual boundaries of an intuitive theory or other element of intuitive conceptualization (depicted as rounded) roughly matches the boundaries of the instructed theory (depicted as rectangular). (b) The range of an instructed theory is "covered" by a large number of "smaller" intuitive elements.

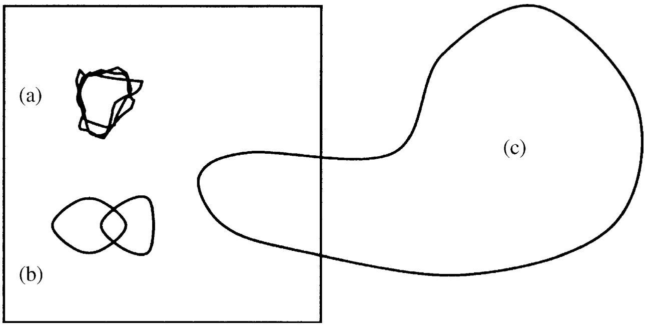


Fig. 2. Issues of contextuality: (a) fuzzy boundaries, (b) overlap, and (c) intuitive elements of wide scope.

leaves out some important aspects that we do believe to be both important and true. Subtleties notwithstanding, the principal phenomenology of this view is robust and consequential, that subjects will switch the way they think about phenomena that are considered homogeneously from the scientific point of view.

Fig. 2 depicts three thorny issues concerning contextuality, invisible in Fig. 1b. First, (a) the fuzzy boundary of one intuitive domain suggests that boundaries might be in principle uncertain and, indeed, may vary from one consideration of a context to the later such consideration (which we would describe, generically, as a *dynamic* aspect of contextuality). For example, contexts may be unstable under continued consideration; a student might see an element in a context, but then, later, it might escape him or her.[5] This possibility becomes important educationally if instructors can manipulate the visibility of an intuitive scheme that happens to produce normative answers, but is not normally seen. Brown and Clement (1989), in fact, show that it is possible to get students to see "springiness" in situations in which they do not usually see it. This case is particularly important since the manipulation is brief (around 20 min) and stable; subjects continue to see "springiness" in the target situations after the intervention.[6]

Versions of "fuzzy or easily shiftable boundary" contextuality less dramatic than Brown and Clement's instructionally important form are more likely to be commonplace. Students may well show several ways of thinking about a situation without any provocation. Indeed, we saw significant dynamic instability of this sort in our data. For example, students may be primed by the way they happen initially to approach the problem or by the way questions are sequenced. Further, the work by Anderson et al. (1992) suggests that whether one approaches a problem from the point of view of explanation or prediction may yield different results. Effects have been noted in the literature on the basis of presentation [e.g., static or dynamic images (Kaiser, Proffitt, Wheelan, & Hecht, 1992); textbook-standard language or informal language (Viennot, 1979), and modality of subjects' expression (language, drawing, or gesture; Church & Goldin-Meadow, 1986)]. In general, we would use the term *perspective* to describe the vast range of ways one may approach "the same physical context," for example, in terms of kind of question, setting, modalities, and even social function.

Fig. 2b shows the possibility of overlapping contexts, even if boundaries are stable. Several technical constructs from physics may apply to one situation (e.g., energy and force), yet the interactions between these are well-specified. In particular, relations are often explicitly prescribed by theory (*energy* change = *force* times distance), and, in any case, use of different technical constructs cannot lead to different predictions. In contrast, it is quite natural to imagine that, in the case of fragmented systems, there might be overlaps with no principled relationship between predictions, and no principled relations among the relevant elements. In situations of overlapping contexts of application, subjects' first reactions are obviously insufficient to probe contextuality thoroughly.

Inquiring why contexts might be unstable or overlapping raises a potentially more profound issue. What principles (presumably intrinsic to the view or theory) define boundaries of contexts? Technical constructs in well-developed fields can be expected to have explicit applicability conditions, often in virtue of explicit and operationalized definitions of constructs. Elements of intuitive knowledge may well have independent learning trajectories, and there is no particular reason to assume applicability will be well sorted out. Overlaps might be an intrinsic aspect of independent learning trajectories. Unprincipled contexts of application might well mark a deeper difference in contextuality between intuitive and professional conceptualization than fragmentation, per se.

Finally, Fig. 2c also shows the possibility that intuitive elements might have wide scope. This possibility is particularly likely to escape investigators to the extent that they have a narrow focus of attention, for example, assuming that the range of contexts covered by intuitive physics more or less coincides with—or sits within—that of professional physics. Instructionally-oriented empirical studies are liable to err precisely here, since they will preferentially attend to contexts that arise in instruction, hence invoke contexts that fit easily within the relevant instructed domain. The existence of elements of wide scope does not threaten the core of a fragmented view of intuitive knowledge. To get any significant coverage of the instructed domain, one might still require many elements—even if some have wide scope—and their contexts of applicability might be dynamically unstable, unprincipled, and/or overlapping.

Embedded in the figures, and probably in much discourse about fragmentation versus coherence, is a hidden assumption that is most likely false. Figs. 1 and 2 suggest the existence of a "neutral" metric on "context size." An "element of wide scope" covers many contexts (much area in the figures), and an "element of narrow scope" covers fewer. But, how do we count or measure contexts? When is a difference a real difference (a new context), and when is it a trivial difference? In our data, some preschool subjects claimed that the color of a block made a difference in questions of force. For them, contexts we perceived (and intended!) as identical were different.

Situations in the real-world do not come with markers of intrinsic identity or similarity. Instead, considerations that split or unify "domains" or "contexts" depend critically on point of view, how the relevant configurations are conceptualized. Where we can identify a common range of situations governed by two different conceptualizations, we can certainly ask whether a smaller or larger number of elements are involved. On the other hand, it is easy to imagine that such judgments might be prejudicial, choosing for reference exactly the range of circumstances covered by one (almost certainly the instructed) conceptual frame. Especially when domains do not coincide from different points of view, there are no obvious ways of computing "degree of fragmentation." For example, the extent of application of an element beyond the "physics" context shown in Fig. 2, is unmeasurable in physics terms. Physics has nothing to say about the fragmentation or unity of "a context" outside of its own range. Other approaches to measuring scope might be valuable, for example, the relative frequency of application in “daily life,” or in “professional practice.” But relative fragmentation—some technically adequate version of “elements per context”—by these considerations, is unlikely to be established without selecting a privileged view. It follows, for example, that we may never be able to decide whether intuitive physics or intuitive psychology is more fragmented.[7]

This discussion of contextuality bears directly on the prominent issue of “differentiation” (e.g., Wiser & Carey, 1983). Naïve conceptions are frequently described as undifferentiated. Yet, it may be that they simply do not make the same distinctions as the instructed/scientific version. Furthermore, scientific concepts are notably undifferentiated in the simple sense that they refuse “obvious” distinctions made by ordinary folks. The important point is that one must carefully develop theoretically cogent versions of “differentiated” (and other such terms), which do not accidentally select a privileged frame of reference, and then one must employ a sufficient empirical range in order to substantiate claims.

To summarize this discussion of relative fragmentation and “size of contexts,” the proper null hypothesis with respect to different conceptual views (say, instructed physics and intuitive physics) would seem to be that each view is fragmented with respect to the other. Physical considerations split apart contexts (“they are governed by different principles”) that are intuitively the same. For example, intuitively, balance scales and people’s karma may both be in or out of balance. But balance, as understood by a physicist, does not apply to karma. On the flip side, physicists see situations as isomorphic that are obviously different to beginners. It is unclear how one can say more than “each view fragments the others’ contexts” (or, “each view ‘unifies’ in a different way”) unless one chooses a frame of reference, or a natural one happens to exist, such as the case where domains coincide (and one can then sensibly, perhaps, just count elements). Perhaps the worst mistake is to take the professional domain as reference, without noticing that this is by definition precisely the set of contexts that professionals “unify”—that they treat homogeneously.

Contextuality represents one possible refinement of the issue of fragmentation versus coherence. However, considerations of contextuality are answered complexly, rather than in terms of a simple “more or less fragmented.” We need to specify, for example:

- whether the intuitive and instructed domains coincide, or specify something about their respective coverage;
- whether it is sensible to privilege one view over the other in tracking “contexts” (continuing to recognize the limits inherent in choosing a privileged conceptual frame of reference);
- whether contexts (the range of application of different elements) show overlap, have unprincipled boundaries, or have unstable boundaries in view of extended consideration, changes in perspective (such as mode of presentation or mode of expression).

In this framing, an obvious prototype for “fragmented” emerges. We call it *crowded-irregular* for reference. Crowded-irregular knowledge systems entail: (1) many elements (where a reference ground exists to allow counting), (2) overlapping contexts that are (3) ad hoc in their specification, and (4) perhaps even dynamic instabilities of a number of types. Recognizing this prototype, still, a more careful theoretical or empirical treatment should separately consider these attributes as they do not necessarily go together.

Contextuality is motivated mainly by the "fragmented" side of the coherence versus fragmented issue. Let us now consider coherence. Crowded-irregular knowledge systems may be roughly characterized as incoherent. Certainly individuals possessing such a knowledge system will frequently change the way they think about things, the reasons for change may be very hard to characterize, and those reasons might be complex enough to make changes appear random. However, coherent is not a clean opposite to fragmented. In particular, even if an individual shows none of the problems of conceptualization mentioned above—that is, has stable, principled, non-overlapping context boundaries—we can still ask about the relational structure[8] of the set of ideas he or she applies. Is one idea *deduced from* others? Are some ideas merely *similar to* others?[9] If the ideas in a knowledge system have no clean relationship with one another (once again, the paradigmatic case would be where elements develop largely independently from one another), one may still cogently argue that the word "incoherent" is appropriate.[10] As with contextuality, questions about relational structure must be answered complexly. Certainly one should specify the nature of relations that are implicated. In this regard, we feel it is absurdly optimistic to expect that intuitive conceptualizations would have a prototype of "logical" relations, that of an axiomatic system. Short of that, terms like "coherent" beg technical definition, which, as a rule, is not forthcoming in conceptual change research. Cognitive models or simulations, of course, display some types of relational structure explicitly. However, they also may hide important general issues in the technicalities of the modeling language, and may not explicate knowledge-specific relational structures, say, critical domain-specific characteristics that do not show up at the architectural level.

We can put a hard edge on this discussion of coherence and fragmentation with a particular hypothetical example. Suppose we have one knowledge system that has a large number of elements and a second that has a much smaller number (assuming a common range of application). Suppose, however, that the former has a very strong relational structure, something like that of an axiomatic deployment, and the latter has a very loose relational structure, say, merely analogical. How can one decide whether one is more coherent or fragmented than the other? The bigger has, in some sense, more fragments, but they go together tightly. The smaller system may be smaller and simpler, "governed by fewer principles"; but elements are weakly connected. We believe that the answer is that the meanings of coherent and fragmented must be better specified and operationalized to make progress, and we offer contextuality, relational structure, and specification (below) as directions for better definition. In this article, we will minimize discussion of relational structure, as the empirical base we use is almost devoid of leverage for answering such questions.

2.2. Specification

How much and what kind of things should a theorist of conceptual change say in order to specify a particular "concept" or "theory"? Will a paragraph of natural language prose, as suggested by the example given above (McCloskey's compact description of "the intuitive impetus theory") suffice? Under other circumstances, one might demand a process model of the operation of the knowledge system. Here, our aims are more modest and more specific. We wish to outline a plausible set of requirements for specifying important aspects of the content of a concept that is a physical quantity, such as force. This specification does not cover all concepts, it does not deal with form (as a process model might), and it does not obviously take on the issue of macro-structure—the nature of the embedding theory or model in the case that one considers concepts embedded in theories.[11]

The following set of principles for specification of conceptual content is simply a list of aspects of the concept that ought to be specified. Each aspect has a straightforward and very specific projection into expert competence. That is, we can fairly easily describe the nature of an expert's concept corresponding to each aspect. Furthermore, each aspect provides, at least potentially, a locus of contrast for how novices or naïve individuals think about force. These two characteristics mean that the list has strong face value in considering conceptual change: The relevant characterizations of an expert's concept constitutes a list of goals for instruction, and the differences in each aspect constitute dimensions of the transformation from naïve to expert thinking. We phrase these aspects in terms of the concept "force," although, as mentioned, they apply to any physical quantity.

2.2.1. Existential aspect

To what situations will a subject attribute the existence of a force? Physics textbooks provide a short list of kinds of forces—including electro-magnetic, gravitational, strong (nuclear), and weak forces. The existential basis for each of the basic types is remarkably simple. Gravitational forces exist between any two pieces of matter under any conditions at all. Electrical forces exist between any two charged pieces of matter under any conditions at all. In practice, only gravitational and electrical forces are relevant to mechanics, although "practical" manifestations of such forces are given special treatment. For example, ordinary contact forces of one body on another are a manifestation of underlying electrical forces in the constituent atoms and molecules as they come into proximity with each other. With respect to the existential aspect of force, one telling and well-documented difference between naïve subjects and physicists is that physicists assert the existence of a force pushing up on a book lying on a table, which most naïve and novice subjects reject (Minstrell, 1982).

2.2.2. Coarse quantitative aspect

Physics, of course, has stringent quantitative specifications for forces, which we cannot expect will be matched in intuitive physics. However, non-physicists often exhibit a fairly strong, if approximate, sense of the magnitude of influences (like forces) and effects. In particular, non-physicists often specify when a force is zero, or very small (negligible) with respect to other forces, and they sometimes can make benchmark specifications, such as one force being exactly the same as another (e.g., as they must "cancel out"). We use coarse quantitative, rather than quantitative, as a requirement: (1) to put experts and novices on more equal footing, (2) because some characterizations of naïve thinking highlight this kind of reasoning (e.g., Forbus, 1985), and (3) because we have in other places conjectured a special role for coarse quantitative reasoning in the process of conceptual change (diSessa & Sherin, 1998). In as much as zero or non-zero constitute specification of existence or non-existence, the coarse quantitative aspect of a quantity is simply an extension of the existential aspect. However, diSessa and Sherin (1998) document a case where a student regularly seems to have a normative existential conception of force, but whose coarse quantitative judgments are systematically at variance with those of physics. This motivates considering coarse quantitative aspects at least somewhat independently of existential ones.

2.2.3. Ontological aspect

The existence or non-existence of something tells little about its nature. What is the nature of a force, its ontology? In this aspect, we tread a middle ground between highly abstract ontological categories such as matter or process (Chi, 1992; Keil, 1989) and some full specification of "the nature of force." In particular, we include spatial–temporal issues, and issues concerning the relation of forces to objects. Where, exactly, is a force? Can it move? Is it conserved, or, alternatively, what is the typical period of decay or growth? With respect to relations to objects, is a force a property of an object as a whole, or of each piece of an object? Can it exist outside objects? To physicists, a force is not an object, it cannot be said to be "in" an object, nor does it "belong to" an object. It is not conserved, and it is spatially located quite precisely in the interface between two objects (or between specified parts of a single object). An important part of the spatial nature of a force is specified by saying it is "a vector quantity"; it has a direction. All these (professional) aspects of force's ontology can be put together compactly (diSessa, 1980) in saying that a force is "the flow of a conserved vector quantity." Specifying force as a flow of a vector quantity provides a highly restrictive spatial–temporal specification: It implicitly requires a boundary over which to specify the flow. This is a refinement of the more often mentioned but rougher characterization that force is "relational," that it requires specification of two objects.

In contrast, novices give the impression that forces, themselves, are at least partially conserved but may have induced or natural time-scales of decay (McCloskey, 1983). Forces can move between objects or even within objects (as we shall see later). Furthermore, young subjects sometime incorrectly assign the temporal/causal characteristics of "strength" to force (Ioannides & Vosniadou, 2002): that force may inhere in some object without any effect, only to be "called to life" by some circumstances (such as an individual's exertion of effort, or in reaction to the imposition of some external influence).

The final two aspects, below, can be grouped together as inferential aspects. That is, they relate to inferences one may draw from joint existence of multiple forces, or across the gap to other ontologies, such as motion.

2.2.4. Compositional aspect

Multiple forces, from an intuitive or professional perspective, combine when co-present in some situation on an object. What are the principles of composition? The principles are simple and mathematical for physicists: vector addition. For novices, forces may act to cancel or to "overcome" each other. Indeed, forces can act on each other, rather than combine (examples to come). We are not aware of prior research on compositional aspects of the concept of force. However, empirical work reported in this paper provides a starting point.

2.2.5. Causal aspect

What are the consequences of the existence of a force? The core of Newtonian mechanics, Newton's second law ($F = ma$), specifies one very important causal–inferential aspect of the professional concept of force: that one may determine a very particular aspect of motion, acceleration, from the existence of a force. Naïve and novice students more typically infer a different aspect of motion, speed, from a force (Viennot, 1979).

None of these aspects seem to us plausibly trivial or ignorable. Each one (with the exception of composition) are associated with well-known, well-documented, and claimed-to-be important "misconceptions" that students exhibit during instruction. In fact, we feel that the best argument for the inclusion of an aspect in a list such as this is that description at that level and in those terms turns out to be consequential for the learning of students. In contrast, we do not claim that these aspects are orthogonal or complete.[12]

In the following two sections, we briefly sketch the two reference models of conceptual change that underlie our own research and that of Ioannides and Vosniadou, and situate them with respect to the elaborated theoretical frame for discussing "coherence versus fragmentation," developed above.

2.3. Reference models

2.3.1. Knowledge in Pieces

We call our framework for considering conceptual change Knowledge in Pieces (diSessa, 1988, 1993). This view locates intuitive physics largely (but not exclusively) in hundreds or thousands of self-explanatory schemata, typically abstracted from common situations. We call these schemata *phenomenological primitives*, p-prims for short. "Phenomenological" connotes the fact that p-prims are often relatively apparent to people in real-world contexts. Everyone just knows that "more effort begets more result," and we see it in our everyday interactions with the world. "Primitive" involves two senses. First, p-prims are primitive in the sense that they are typically evoked as a whole. In addition, p-prims are explanatorily primitive in that one's attitude toward them is they are what happens naturally in the world. No further explanation is necessary or, typically, possible. "That's just the way things are." Thus, p-prims provide a sense of satisfactory understanding of situations in which they are evoked, and surprise or puzzlement when none are available, or if the world's behavior is inconsistent with the entailments of p-prims that are evoked. If an object is moving faster (more result), apparently without cause, one puzzles but achieves resolution by finding that someone or some thing is "working harder" on the object.

Heuristically, we can characterize p-prims as subconceptual entities, below the conceptual (word) level, perhaps close to the level of senses of individual words, such as various senses of balance, equilibration, and so on. A number of "misconceptions" noted in the literature have been re-explained in terms of p-prims. For example, McCloskey's "impetus theory" appears to be a local confluence of about a half-dozen p-prims adapted to a particular class of situations (diSessa, 1993, 1996). However, impetus p-prims do not always, or even mostly, work together. Refined studies of impetus-like expectations (e.g., Ranney, 1988) empirically confirm at least some degree of contextual divergence among different aspects of impetus. Recent work (diSessa & Sherin, 1998) has developed a model of how p-prims participate in the development of technical concepts, like force. Although highly relevant to a full theory of conceptual change in science learning, models of expertise are not relevant here since no subjects come close to technical competence with the concept of force.

The development of the system of p-prims involves a complex sorting of connections to various contexts and sorting of relative priorities (diSessa, 1993). However, the ultimate system is far from uniformly systematic, owing to the richness of physical experience and the intrinsic difficulty of developing integrated views of all of it. This roughly bounds the relational structure of the system of p-prims: There are, for example, common attributes among several p-prims, more and less important p-prims, and some weakly inferential relations. Yet, in terms of number, intrinsic diversity and relatively independent developmental paths, the system is primarily fragmented.

Contextuality with respect to p-prims is intrinsically complex for a number of reasons. First, as mentioned above, mere number and relatively independent developmental trajectories (leading to the need for relatively independent specifications) mean, for example, that compact specification of the set is impossible. Furthermore, p-prims are likely encoded preferentially in kinesthetic and visual-dynamic terms, making natural language description difficult and suspect.[13] This problematic relation of p-prims to language introduces some added complexity in interpreting Ioannides and Vosniadou's results in terms of p-prims, about which we say more in following sections.

Specification is, in view of Knowledge in Pieces, also complex for many of the same reasons as contextuality. Because no p-prim or any small collection constitutes "the naïve view of force and motion," it does not make sense to hold a p-prim individually accountable to all aspects of a concept like force in the way that it does for Ioannides and Vosniadou's meanings. In different situations different p-prims will account for predictions and explanations, which would all presumably be covered by Ioannides and Vosniadou's "theories of force." Tracking each p-prim's contributions to aspects of the concept of force, as it develops, is a worthy, but theoretically and empirically complex task, only part of which has been accomplished.[14]

While Knowledge in Pieces can be described roughly as a commitment to fragmentation, we also believe that at least some elements of intuitive physics are of wide scope (e.g., apply well-beyond "physics" contexts, and they are very frequently used in conceptualizing the world—see diSessa, 2000), and that "fuzzy," overlapping boundaries and dynamic instability may be as or more important than "large numbers" per se.

2.3.2. Framework Theory model

Ioannides and Vosniadou's work is cast in terms of Vosniadou's model of conceptual development, which has matured over a series of studies of conceptual change (e.g., Vosniadou & Brewer, 1992, 1994). Earlier work concludes that children's intuitive ideas about the physical world (e.g., the shape of the earth and the reasons for the day/night cycle), rather than being inconsistent and fragmented, represent a few "coherent models" that are constrained by presuppositions or entrenched beliefs that grow out of experience but are largely unavailable to conscious awareness. According to Vosniadou, children's presuppositions—for example, the belief that space is intrinsically organized in terms of up and down, or that unsupported objects fall—constitute a relatively "well-established and coherent explanatory system" of the physical world. These latent presuppositions about the physical world form what Vosniadou calls a *framework theory*, which constrains and shapes the ways children think about and understand many particular issues.

In addition to the framework theory, Vosniadou posits a set of second-order conceptualizations, or "specific theories" (in earlier work called "specific models"), which are embedded within and constrained by the framework theory. For example, a child's explanation that people live on top of a pancake-shaped planet is, according to Vosniadou, indicative of a specific theory (model) of the shape of the earth (pancake-shaped) that is shaped by the presuppositions of the framework theory, namely unsupported objects fall and space is organized in terms of up and down. Vosniadou claims that specific theories are easier to change through instruction than are framework theories, but even the core presuppositions of the framework theory are subject to elaboration and even radical revision. The main difference between the theories Vosniadou posits and theories from professional science is that the framework and specific theories lack the systematicity, abstractness and metaconceptual nature of scientific theories. Nonetheless, Vosniadou (Ioannides & Vosniadou, 2002; Vosniadou, 2002) consistently and specifically in opposition to Knowledge in Pieces characterizes these intuitive theories as constituting coherent and internally consistent explanatory systems.

In I&V's work on the concept of force, framework theories are said to contain ontological presuppositions such as "objects exist," "objects are intrinsically animate or inanimate," and "states (e.g., rest) are distinct from processes (e.g., motion)." They also contain epistemological assumptions such as "processes need explanation; states do not." The framework theories combine with cultural and contextual information to produce one of a small number of specific theories, which are the actual experimental locus of this work.[15]

In terms of relational structure, some aspects of framework theories seem reasonably well laid out. For example, assumptions are categorized into ontological and epistemological classes, and the classes are presumably independent. The relations of constraint on specific theories, however, are not explicitly spelled out, nor is there any specification of the relational structure of specific theories, aside from what can be inferred from the provided compact natural language descriptions of their content. Contextuality plays no explicit role in Ioannides and Vosniadou's description. However, they believe that framework theory constraints are sufficient to eliminate all but a few specific theories, which contrasts starkly with Knowledge in Pieces. Across Vosniadou's studies, it turns out generally that there are roughly a half-dozen specific theories per domain over the age ranges investigated, and almost no subject gives signs of holding more than one. These theories are described as "narrow but coherent" (Vosniadou, 2002). However, there is no specification of what "narrow" means, and no operationalization of "coherent," except that the empirical work presumes students will answer consistently over the range of questions asked. Rationale for the range of questions asked is limited to a list of presumed-to-be relevant attributes (size, weight, motion) and "extensive pilot work." Given our concern for contextuality, the exclusion of many possible attributes is likely highly consequential, which our data will show.

As far as a theoretical specification of coherence, we could find only two examples of specific judgments in I&V (2002), and neither involved explicit characterization of the nature of coherence, per se. Ioannides and Vosniadou defend meanings that are hybridized from two different other meanings as coherent because subjects apply those meanings in different contexts (which we would capture under contextuality—non-overlap—rather than under the more stringent meaning of "coherent": relational coherence of some sort). On another occasion, they claim that two such hybridized concepts (Internal Force and Acquired Force, described below) are inconsistent, which drives conceptual change forward. This inconsistency surely is not of a strictly logical nature. In fact, the nature of this inconsistency is not explicitly delineated by Ioannides and Vosniadou.[16] Furthermore, if the relation of hybridized components is truly an inconsistency, it would seem that the hybrid meaning ("Internal and Acquired") should be removed from the list of coherent meanings, which I&V do not do.

Ioannides and Vosniadou provide no explicit treatment of their principles of specification. From our perspective, their data deal only with existential and, to a lesser extent, coarse quantitative aspects. Several statements throughout the paper suggest that they are aware that their data do not bear much on ontological aspects: They comment several times with respect specifically to gravitation and at least once with respect to pushing and pulling that their data do not make clear what students mean by "gravity" or "pushing."

2.3.3. Bridging to empirical work

Ioannides and Vosniadou's empirical work deals specifically with the meaning of the word force. They ask if subjects see force in various situations (i.e., if they would use the word to describe what is happening in particular situations). This constitutes a small difficulty in connecting the experiment to their theory because force is, at best, a part of subjects' theories of force and motion. However, Ioannides and Vosniadou believe the constraints between word meaning for force and subjects' theories are tight enough that the former strongly reflects in the latter. One should see a "small number of relatively well-defined and internally consistent interpretations of force" (p. 1). I&V take a word to be a "model that consists of an interconnected set of presuppositions and beliefs that has a causal explanatory structure" (p. 4). We feel this is a flawed theory of word meaning. Specifically, words are typically polysemous and, at best, *fit into* causal models, instead of *constituting* them. Nonetheless, it seems at least plausible that causal theories constrain word meaning and, in any case, this is the empirical trace to which I&V have committed themselves.

As mentioned, the relationship of words to p-prims is not straightforward. We simply would not propose to expose individual p-prims cleanly in word meaning tasks. Nonetheless, for the purposes of keeping methodology comparable, we accept that asking subjects about the existence of forces and their relative magnitudes will reflect enough of their overall conceptualization of force and motion that the complex contextuality of p-prims should be evident in responses. We are more certain that what students say reflects their conceptual resources in the extension study, since tasks for the extension were constructed with specific p-prim reactions in mind.

In planning our quasi-replication and extension, we decided to use both the word "force" and the words "push or pull." 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. Indeed, physics books often define force in a preliminary manner as a "push or pull," and one of the meanings I&V use to model their data is Push/Pull.

Finally, I&V incorrectly project on Knowledge in Pieces the expectation that young children should be more fragmented than older subjects. They say, "According to the fragmentation hypothesis, children's initial meanings of force should be unsystematic and fragmented and we should see increasing systematicity and coherence in these meanings with development and instruction" (p. 5). We do believe relevant conceptualizations become more systematic and consistent *approaching physics competence*, but have no particular projections about spontaneous development before instruction, nor during early phases of instruction. No subjects in I&V's study nor in ours came close to competence with the Newtonian concept of force, so consideration of the ultimate change toward systematicity that Knowledge in Pieces projects is irrelevant.

3. Ioannides and Vosniadou's experiment

3.1. Method

I&V's basic methodology, which we call *model mapping*, is straightforward. Ask subjects a range of questions about a domain, and compare their answers to the profile of answers an idealized subject would give if he/she had one particular conceptualization of the domain. In this case, an initial set of models (meanings) was hypothesized, although they had to be modified and extended to cover the range of answers actually obtained.

In more detail, 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 represented as an object on a pointy hill and explained as "it could easily fall down"), objects in free fall, and objects that had 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 were asked the question in the colloquial form "Is there a force on the x?" because they did not appear to understand the form "Is there a force exerted on the x?" Some of the questions were compound, involving 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 asking for a comparison of the two situations. Each question and compound question was scored as a whole, using codes that entailed a specific pattern of existence/non-existence, classes of explanations, and comparisons.

Ioannides and Vosniadou hypothesized that it would be "possible to assign the majority of children in . . . [their] sample to the consistent use of a small number of meanings of force" (p. 28). Based on an initial analysis of the students' responses, they hypothesized that there were four core interpretations of force that make up the explanatory structure underlying students' understanding of force:

1. Internal Force: an internal property of stationary objects related to size or weight.
2. Acquired Force: an acquired property of inanimate objects that explains their motion and their potential to act on other objects.
3. Force of Push or Pull: the interaction between an agent (usually animate) and an (usually non-animate) object.
4. Force of Gravity: the interaction at a distance between physical objects and the earth (p. 28).

Ioannides and Vosniadou then generated a pattern of responses (in terms of question set 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 consistently used the core meanings of Force of Push or Pull or Force of Gravity. However, Ioannides and Vosniadou "hypothesized that they [the students] had used several [composite] meanings of force, consisting of combinations of the above-mentioned core explanatory frameworks" (p. 32). Again, the composite models described by Ioannides and Vosniadou are still claimed to be "internally consistent"; that is (operationally), students use each of the core ideas making up the composite model consistently in appropriate (different) contexts. This is in contrast to what Ioannides and Vosniadou 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 responses could be characterized as being indicative of the student's having and using either a core meaning of force, or a meaning of force consisting of a combination of core meanings. [In addition to the distinction between core and other models, I&V distinguish between *initial* models (unaffected by instruction) and *synthetic* models (affected by instruction).] In order to test this prediction, they again established a mapping from students' response codes to the various meanings of force and then compared the actual responses to these mappings. The criteria used by Ioannides and Vosniadou for assigning students to each of the meanings of force are as follows:

1. Internal Force (core, initial): Students were assigned to this meaning of force if they gave answers indicating that there is a force on all objects, or only on big/heavy object because they have weight or are big/heavy and do not refer to gravity, the object's motion or an agent.
2. Internal Force Affected by Movement (initial): Students were assigned to this meaning of force if they gave answers indicating the force is due to size/weight of object only, but also indicate that moving objects and objects that are likely to fall have less[17] Internal Force than do stationary objects.
3. Internal and Acquired (initial): Students were assigned to this meaning of force if they indicated that there is a force on stationary objects due to size/weight, and that these objects acquire an additional force when they are set in motion. Ioannides and Vosniadou 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.
4. Acquired (core, initial): Students who indicated that force is a property of objects that explains motion and potential to act on other objects were assigned to this meaning of force. These students answered that there is no force on stationary objects, and the force on moving objects disappears when the object stops moving. Ioannides and Vosniadou also included students who thought that force was only acquired by heavy moving objects, and claim that this response indicates that these students relate the Acquired Force to both the weight and the motion of the object. Additionally, Ioannides and Vosniadou included students who thought that unstable stones had more force because they could be set in motion more easily as well as those who treat both stable and unstable stones identically.

5. Acquired and Force of Push/Pull (synthetic): Students were assigned to the Acquired and Force of Push/Pull meaning if they gave answers meeting the criteria described above for the Acquired meaning of force, but also answered that there was force on an object acted on by an agent regardless of whether it moves.
6. Force of Push/Pull (core, synthetic): Only one student in Ioannides' and Vosniadou's study was assigned to the Push/Pull meaning of force. This student answered that a force was exerted only on objects being pushed by an agent, whether or not the object was moving.
7. Force of Gravity and Other (synthetic): None of the students in Ioannides' and Vosniadou's study could be assigned to the expected core gravitational force meaning described above; all of the students who mentioned gravity also mentioned other forces. However, Ioannides and Vosniadou claim that the answers given by students referring to gravity could be interpreted as composite force meanings consisting mainly of the Acquired Force meaning with the addition of gravity onto this core meaning.[18]

3.2. Results of I&V's study

The results of I&V's study are presented in Table 1. I&V do not explicitly state their criterion, but we believe they assigned students to a particular meaning only if all set codes were consistent with those projected by that meaning. In this regard, their coding had a degree of "softness" built in. For example, students were allowed to attribute no force in situations where the controlling parameter was small (e.g., if students have the Internal meaning and compare a large to a small object, they were allowed to say the small object had no force on it). Students with Internal/Affected by Movement and Internal/Acquired were allowed different interpretations as to whether instability (in contrast to stability) should imply the existence of a force, or the opposite. Students with the Gravity and Other meaning were allowed not to mention gravity in agentive pushing situations. Furthermore, I&V list five distinct submeanings of Gravity and Other, although the main submeaning contained about half of subjects who exhibited Gravity and Other answers. The degree of softness of coding Gravity and Other will need attention in our quasi-replication, as well.

Ioannides and Vosniadou's results are striking and apparently offer compelling support for their claims of a limited number of consistently applied meanings. By extension, these

Table 1
Summary of Ioannides and Vosniadou's data, frequencies of meaning of force as a function of grade

	Force meaning	Kindergarten	4th	6th	9th	Total
1	Internal	7	4			11
2	Internal/Affected by Movement	2	2			4
3	Internal/Acquired	4	10	9	1	24
4	Acquired		5	11	2	18
5	Acquired/Force of Push/Pull			5	10	15
6	Force of Push/Pull				1	1
7	Gravitational and Other		3	1	16	20
8	Mixed	2	6	4		12

results support their claims of the existence of a framework theory that guides and constrains children's understanding of the concept of force. Almost 90% of subjects "made use of a small number of relatively well-defined and internally consistent interpretations of force" (Ioannides & Vosniadou, 2002, 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 Ioannides and Vosniadou, Gravity and Other, contains components of those same four core meanings.

A second key result, consistent with I&V's theoretical framework, is that the students appeared to progress from the Internal meaning of force through a composite (but uninstructed) Internal/Acquired meaning to the Acquired meaning, evidence for what they claim is spontaneous conceptual change occurring in young children. Similarly, the ninth graders predominantly used the synthetic (affected by instruction) Acquired and Force of Push/Pull meanings, which Ioannides and Vosniadou claim is evidence of students attempting to synthesize instructed meanings of force onto the core (initial) Acquired meaning of force. According to Ioannides and Vosniadou, these composite meanings of force are "internally consistent," by which they appear to mean that different meanings are used in disjoint contexts: A student who appears to use the composite Internal/Acquired meaning of force consistently will answer questions about forces on stationary objects by referring to size and questions about moving objects by referring to the object's motion. Ioannides and Vosniadou cite their findings of older students tending to use composite meanings to support their claims that older students exhibit "increasing fragmentation" as they are exposed to instruction. This finding allows the Framework Theory hypothesis to account for data from older students that supports the Knowledge in Pieces hypothesis (e.g., diSessa, 1993, 1996). On the other hand, none of the ninth graders used "internally inconsistent" mixed models, a finding that appears to be somewhat at odds with I&V's increasing fragmentation hypothesis.

4. Empirical plan

4.1. Motivation and questions

On this background, we sketch our empirical plan. The first study (quasi-replication) was intended to be a near replication, and we expected to obtain very similar results to I&V. Differences we introduced were mainly for the purposes of (a) systematizing the design (always asking comparisons, rather than occasionally), (b) simplifying it slightly by removing the least informative questions and (c) by asking both about force and push/pull, providing the opportunity to glean a bit more information out of the study. We used nearly identical questions and the same kind of props—hand-drawn stick figures. In more detail, the main differences between our quasi-replication and I&V's study were:

- 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 eliminated contrasts of different height (e.g., to see how this dimension affected gravitational conceptualization), and eliminated questions about balloons (which disambiguate size and weight).

- We asked every question in a comparison set, rather than mixing comparisons with standalone force questions. Each of the 10 sets of questions that constituted our quasi-replication consisted of a pair of situations and a comparison. Every situation contained in the comparison sets was also in I&V's study. This design also has the advantage of more systematically collecting coarse quantitative information from students.
- Our questions were in a slightly different order, mainly as forced by asking comparisons always together with individual cases.
- Our interviews were conducted in English rather than Greek.
- We did not use the phrasing "is there a force/push–pull *exerted* on . . .," but, more simply, "is there a force/push–pull on . . ."
- We asked each subject alternately about force, or push/pull in each of the 10 question sets, and order was balanced across subjects.

Only the first and last of the above-mentioned differences should affect results, provided I&V's results are reasonably robust. The first should lower the sensitivity to fragmentation and thus should help I&V's case on this quasi-replication. The last item is theoretically motivated to provide some probe diversity that might increase apparent fragmentation, invoking p-prims somehow preferentially attached to different words. It might well result in less consistency in results. Appendix A lists the complete set of questions we asked our subjects in the quasi-replication. Appendix B gives the mapping between our questions and those of I&V.

The second, extension study was designed to show results that I&V cannot account for. Our view of their experiment—leaving aside questions about the meaning of coherence—is, to first approximation, that we can exploit their lack of specificity with respect to context. That is, we can show their consistent categorization of individuals breaks down on a wider range of problems. The contexts/questions in our extension study were also theoretically motivated to provide additional information on other aspects of the meaning of force, beyond the existential and coarse quantitative data I&V collected.

More specifically, in the extension study we sought to provide the basis for displaying the following patterns of data:

1. *Fragmenting categories of people*: Individuals classified as having a particular meaning (as specified by I&V) exhibit responses that are different from one another when asked a wider range of questions. This suggests that I&V's claim to have specified a (homogenous, among people in the category) way of thinking (a "meaning" for force) is an artifact of asking about too narrow a range of contexts.
2. *Fragmenting contexts*: Individuals are sensitive to attributes of context not described in I&V's categories, and they reason differently based on those attributes. The implication is that their accountability to contextuality is too weak to cover how individuals will reason outside the range of contexts involved in their experiment (which highlight only the attributes featured in I&V's meanings). Note that this pattern is independent of the first. Even if I&V categories of subjects behave homogenously, their specification of those categories may be inadequate to explain differences in reasoning across contexts.

3. *Unaccounted-for reasoning*: We intended to show that people have ways of reasoning about force and motion that are simply not prescribed by I&V's meanings. This is where our particular specification of aspects of the concept of force comes in. I&V's meanings do not prescribe how people should reason about several important aspects of force. Note that this pattern provides new opportunities for observations in the above two patterns. Consider pattern 2, for example. Not only will we see that subjects answer *the same questions that I&V asked* differently based on attributes not implicated in I&V's meanings, but subjects make distinctions among contexts beyond those named by I&V when answering *further questions* (implicating other aspects of specification), as well.

The extension consisted of 14 additional questions asked of each subject directly after the quasi-replication. We sketch those questions, their motivations and our expectations below. (Questions that are not analyzed here are presented in Appendix C, although we keep the numbering of the full set in order to facilitate cross-publication reference.) Each of these problem situations involved real props, rather than sketches. All the expectations in sets 11–19 and 23 are documented in diSessa (1993). Interviewing in these extension questions was a bit more open, starting with more neutral questions and moving toward the pointed ones we intended to code. We wanted to get more information than we could allow ourselves in the quasi-replication.

Set 13: Ball on string; Set 14: Ball in tube: This contrasting pair is motivated by the fact that we found in prior work that subjects seem to have a cluster of primitives that have to do with agentive, forceful interventions, and a different cluster involving constraint or geometric parallelism. Set 13 asked about a situation where a ball is spinning around in a circle on a string held by the opposite end (agentive, forceful). Set 14 asked about a ball running in a circle inside a tube (constraining, geometric). These situations are isomorphic from a Newtonian point of view. More relevant to the empirical issues here, they also do not differ in any of the attributes mentioned in I&V's meanings.[19] Thus, differential answers across these situations would show contextual fragmentation with respect to I&V's proposed meanings.

Set 15: Struck bell (clapper removed): There is no overt motion in a struck bell, although physicists know it is vibrating microscopically. However, from prior work we believed some subjects would see a kind of activity or agency in the bell, and describe it in terms of force, even though Acquired Force should not apply. In addition, we expected to see some different ontological characteristics of force showing up in this problem.

Sets 16–19: Leaning blocks: We asked about forces on two blocks leaning against one another: two big blocks leaning symmetrically; a big block leaning on a smaller one; a small block leaning on a big one; "fat" and "thin" blocks leaning symmetrically. We expected the "top" block in situations of asymmetry would be interpreted agentively, as applying a force, while the bottom would be interpreted only as supporting or resisting. None of I&V's attributes (with the possible exception of "instability," treated later) can distinguish "leaner" from "leanee," and thus we expected to document fragmentation of context not accounted for by I&V.

Sets 20–22: Pushing blocks: We asked subjects what would happen in case we pushed a block: in one direction; in opposite directions at the same time; in two orthogonal directions at the same time (pushing both equally, and also pushing in one direction "harder" than the other). These are the simplest cases of composing forces, hence should reveal a little about the compositional aspect of force.

Set 23: Yo-yo: We asked subjects to predict what would happen if the string on the yo-yo depicted in Fig. 6 (in Section 6) were gently pulled. As explained in diSessa (1993), the circular gestalt of this context should preferentially cue "motion to the left" (i.e., the tug spins the yo-yo, leading to leftward motion). This result would contradict the core inferential aspect displayed in Set 20, namely that force creates motion in the direction of its push or pull, and thus it would demonstrate contextual fragmentation with respect to causal inferential aspects of the concept. Again, I&V did not investigate causal/inferential aspects. Thus, the question reveals more about the meaning of force that individuals use, but it should also show fragmentation that I&V cannot explain by their meanings, in any case. (Once again, none of the attributes used in any of I&V's meanings—size, weight, motion, stability, etc.—can capture this consequential difference in contexts.)

4.2. Subjects and procedure

We administered the 24 question sets (quasi-replication plus extension) to 30 subjects across roughly the same age ranges as I&V. In addition, we added a debriefing for most subjects, which included general questions about whether they found the questions difficult or easy, and, in particular, whether they felt that forces are different from pushes or pulls (after noting to subjects that we had asked about both). We had nine preschool subjects (mean age: 5 years 1 month), nine elementary school subjects (mean age: 7 years 8 months), six middle school subjects (mean age: 12 years 6 months) and six high school subjects (mean age: 15 years 11 months.) 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. Interviews lasted between about 20 min to more than an hour. We split the interviews into two sessions with preschool subjects, in anticipation that they might find the full 24 question battery daunting.

4.3. Coding

We developed our own coding scheme for all 24 question sets that included three primary aspects: (1) whether 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; (3) judgment of comparative strength for each set that contained two focal objects (including all the quasi-replication questions). We did not include consideration of explanations, as I&V did. We found subjects' explanations diverse, often vague, and confusing. As a consequence, we wanted to avoid stretching our interpretive abilities, opening the way to bias or claims of bias. However, all the attributes of students' responses that we coded and used in model mapping were explicitly specified in I&V's codes. So every set of student responses that mapped to one of I&V's meanings should map onto the same meaning in our analysis. Said differently, the codes we used for model mapping used strictly an informational subset of what was used by I&V, and therefore, our coding was strictly "softer" and less sensitive to fragmentation, which should have biased our results in favor of I&V.[20]

In addition to coding aspects 1–3 (discussed above and used in the quasi-replication), for all 24 questions we coded (4) whether the focal object exerted a force on other objects; and, if so, (5) what the other objects were. These were intended for analysis beyond the quasi-replication. Other codes implicating direction (e.g., "in the direction of force") "diagonal between the direction of orthogonal applied forces" were added for sets 20–24. Set 24 had some unique codes for the nature of the force involved.

We developed elaborate coding sheets to make sure codes were applied in uniform terms. These included rules for coding in case subjects changed their minds (in general, we coded the final response), and rules for the use of implied attributions. (For example, if subjects explicitly told us that gravity applied in every case, we coded for gravity in every case, even if they did not explicitly mention it. If subjects had previously discussed a context, we allowed importing the coding of that situation, provided there was no evidence suggesting a contrary interpretation in the new case. In general, the latter was used only in adjacent sets, where a subject might well say—and many did—"I already told you that . . ..")

To test intercoder reliability, two coders independently coded a representative set of six (two each from elementary, middle and high school) subjects on the coding used in the quasi-replication study (the first 10 question sets). We had 99.4% intercoder agreement on codes relating to existence and comparison (158 out of 159). We had 98% agreement (265 out of 270) agreement on any codes that were used in our mapping to I&V models. We had 94% agreement on all codes (286 out of 293). 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.

As mentioned, 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 of a set. Therefore, 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 (e.g., size) was small]. See the list of "softnesses" in the first paragraph of our reporting of I&V's results. We 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. Appendix D shows representative mappings from codes to I&V meanings.

We report analysis of the quasi-replication and of the extension separately, in the next two sections.

5. Results of the quasi-replication

5.1. Main results

After coding subjects, we computed a *meaning deviation score*, the number of sets (out of 10) on which each subject's responses did not match the allowed responses for each of I&V's meanings (described in Section 3). Any mismatch in a set (e.g., a mismatched attribution of existence, or a non-allowed comparison) implied a "miss" for the set.[21] Then, we assigned each

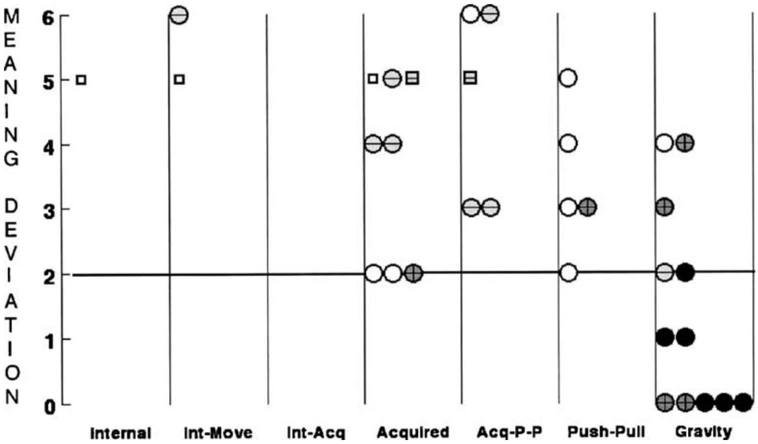


Fig. 3. Assignment of subjects to I&V meanings (horizontal), versus meaning deviation score (vertical, out of 10). Empty shape: preschool; light gray + horizontal bar: elementary; dark gray + cross: middle school; black: high school. Ninety percent of I&V's subjects were found along the bottom line, 0 meaning deviation score.

individual to the meaning on which they achieved lowest meaning deviation score. Fig. 3 shows the results. The vertical bands, left to right, show the force meanings in the order of development as projected by I&V. Each subject shows up as one or more shapes. Preschool subject show as empty shapes, elementary school subjects as light gray-filled shapes with horizontal lines, middle school subjects as dark gray-filled shapes with a cross, and high school student as black shapes. If a subject matched two or more meanings to the same degree of deviation, we split the corresponding marker into smaller shapes, and used squares rather than circles to emphasize multiple matches. At the level of 5 misses out of 10, one middle school subject equally matched Acquired and Acquired/Push–Pull, and one preschool subject matched three meanings.

Only 5 of the 30 subjects fully matched the specification of any meaning (compared to about 90% as found by I&V), and all of these were on the Gravity and 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 and Other meant essentially only that they mentioned gravity as a force in every case. We could have re-mapped to distinguish submeanings of Gravity, as specified by I&V, but later analysis will be as revealing of differences among the Gravity and Other group members and will simultaneously serve to aid our extended consideration of aspects of the concept of force beyond existential and coarse quantitative.

The dark line drawn at the level of meaning deviation score 2 represents a somewhat arbitrary, but also generous allowance for mistakes subjects might have made, a 20% error level. In pursuing the question of whether I&V's meanings really capture the reasoning of subjects, we will concentrate on subjects with two or fewer mismatches. Furthermore, we will also attend less to preschool subjects for several reasons. In general, we found coding responses of preschool subjects quite challenging, and we are less confident that results are meaningful. Indeed, several things suggested preschool subjects were playing an interesting game with us, rather than reporting on how they regularly thought about force and motion. One preschool

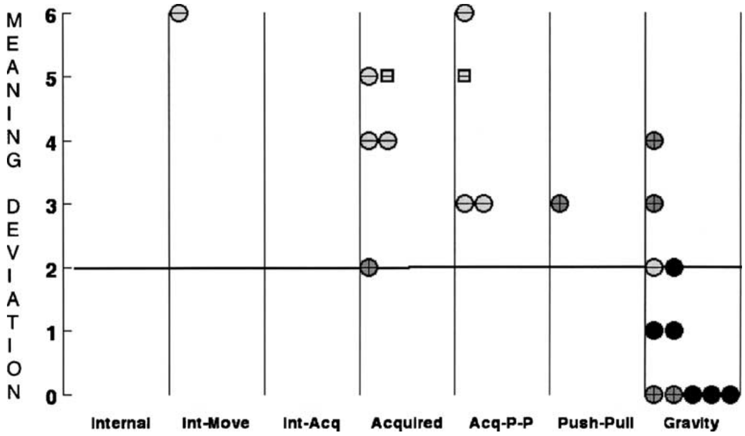


Fig. 4. Meanings assignments (horizontal) versus meaning deviation score (vertical) for elementary, middle, and high school subjects only.

subject, on being introduced to the task, immediately asked, “what do you mean by force?” although she was quite willing to answer our questions. Subjects often gave fanciful responses, such as “little lobsters” being responsible for forces on objects, and we had trouble keeping students attention during questioning, despite the fact that we split the questionnaire into two sessions for preschool subjects. To further emphasize the uncertainty of coding of preschool subjects, we note that they are spread widely across the models (unlike I&V’s results), while elementary, middle, and high school students show a discernable, if “smudged” developmental drift in the direction found by I&V. Finally, we note that the age of our preschool students averaged about 2 years less than the youngest group of I&V, suggesting less comparability, and perhaps pushing the interview into a less sensible age range.

Fig. 4 shows meaning mapping with the preschool students removed. Even with the generous allowance for mistakes, only 10 of 21 remaining subjects match any meaning, and 9 of these 10 match the vague category, Gravity and Other. For brevity of reference, we call the full set of subjects PEMH (preschool, elementary, middle school, high school), and the reduced set, without preschool subjects, EMH. Also for reference, the subjects matching the Gravity and Other meaning were E8, M3, M5, and H1–H6. (Subjects are denoted by a prefix letter, indicating their age group, followed by sequential numbering within each group.)

Only very crude statistics are needed to show that the difference between our results and I&V’s is significant. We modeled the data as binary variables corresponding to whether a subject matched any of I&V’s meanings within the allowed 20% error range, or not. Using a two-tailed *z*-test of the difference of two population proportions we found $p < 1 \times 10^{-7}$. The same data model and test applied to the data with preschool subjects removed results in a $p < 1 \times 10^{-5}$. A more refined data model would reduce these estimates.

Given the unexpected and dramatic difference in outcome of the quasi-replication compared to I&V, we undertook several analyses to see if we could determine the source of the difference. These analyses are discussed in the next three subsections.

5.2. Language: force versus push or 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 address this issue. Three separate questions ask about the forces or push/pulls in one situation (a large rock on the ground) and two in another situation (unstable rock on a hill). Because we alternated asking each subject about forces and push/pulls, we could compare responses. In the case of the rock on the ground, we asked about forces on one occasion, push/pull on 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 *identical* questions asked at different times in the interview. Whether these were mistakes or involved unarticulated contextual factors, we are in no position to say. Among the preschool subjects, four out of nine (44%) gave contrasting answers to force or push/pull versions, while three out of nine (33%) gave inconsistent answers to identical questions asked at different times.

The second question that was phrased in both force and push/pull forms, about an unstable rock on a hill, 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 nine (55%) preschool students switched answers.

Pooling these data suggests that only 10–15% of EMH subjects (skewed fairly strongly toward younger ages) gave different responses to the force and push/pull versions of a question, and this is in the same neighborhood as the number of differing responses that occurred either because of mistakes or unaccounted-for contextual factors (identical questions at different times). In addition we note that 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 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, allowing 20% error rate. 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.[22] Mapping subjects only on the basis of their answers to push/pull questions yielded similar results 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 dataset.

To summarize this line of investigation, although we lack the statistical power to chart 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 identical questions at different points in the interview.

5.3. The effect of asking comparisons

Besides eliminating questions about some dimensions—e.g., height and the distinction between geometric size and weight (balloon vs. stone)—the other main difference between our question set compared to I&V is that we asked more comparison questions. Could this be a source of difference? We repeated our analysis, eliminating the comparison questions that we asked and I&V did not ask. No substantial changes resulted. No subject changed best match meaning.[23] 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 meaning deviation score within the allowed 20% range, from meaning deviation score 2 to meaning deviation score 1.

5.4. Subject-to-subject correlation and "hidden meanings"

We performed one final analysis 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 apparent diversity of our data. For these purposes, we stripped the data of any subcodes, leaving only major coding of the existence of forces and their comparison. This clearly biases the data toward finding patterns as it eliminates the diversity created by subjects' takes on the source and nature of the force. Then, we established a "model" 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, not 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 minimum allowed meaning deviation score of 2. For calibration, assuming every subject matched one of seven models (following I&V's lead), the minimum number of matches would be 55.[24] Even if we allow 10% "mixed" meanings (again following I&V), which would not match any model, one would have to have *at least* 39 matches among the remaining subjects. Our data have a number of between-subject matches more consistent with from 15 to 20 models, as opposed to the seven meanings of I&V. Given that the number of subjects is 30, an estimate of 15–20 models (once again, suppressing all detail concerning source and nature of force) does not provide any support to the notion that there are hidden meanings that could substantially reduce the apparent diversity.

Our data are clearly not random. For example, all high school students thought there was a gravitational force on every object. To calibrate, a Monte Carlo simulation involving random choices of response yielded an expectation of about .02 matches per experiment (with 30 subjects). Yet, the data are clearly highly diverse, with no obvious explanation for the diversity.

There is no way definitively to rule out other meanings (including sensible ambiguities) that might make sense of these 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 dataset of the size they used. Unmotivated patterns of answers, however, make poor candidates for meanings, especially when one is almost guaranteed statistically to find matches at the level I&V found.

6. Results of the extension study

Our extension study was aimed at probing the coherence suggested by I&V by asking more questions, and asking about problems that we felt would violate the constraints of meanings proposed by I&V. In particular:

- We used prior research to suggest situations in which we expected subjects to give divergent answers, compared to those suggested by I&V's meanings. In particular, we introduced attributes not attended to in I&V's analysis nor salient in their set of questions, expecting to show their relevance to force and motion issues.
- We used real-world props instead of sketches, presuming more attributes relevant to forces would arise.
- The interviews were more open-ended since we did not consider it important to constrain ourselves narrowly to I&V's methodology.
- We asked questions motivated by getting some better data on specification, including at least a little data on all the aspects we listed: existential, coarse-quantitative, ontological, compositional, and causal aspects. Because of the still narrow range of questions and limited discussion with subjects, our expectations continue to be limited in terms of how much specificity could result. Instead, we intended only to show that much more specification than that introduced by I&V is appropriate.

We organize our exposition around the listed aspects of specification, discussed in Section 2. Given the results of the quasi-replication, there appears to be no point in strictly adhering to model mapping (even if we could project how I&V's meanings should constrain answers to these new questions). Instead, we use our coding, described above, to identify patterns in the data that I&V cannot account for.

6.1. Existential and coarse quantitative aspects

6.1.1. Sling versus circular tube

Expectations. This pair of questions asked about forces on a ball constrained by a string to move in a circular path as opposed to a ball constrained by a tube. There is no distinction between these situations on the basis of the attributes involved in I&V's meanings, e.g., size, motion, stability, etc. As noted, we expected subjects to be less prone to assert the existence of a force in the tube situation. Again, we concentrate on EMH subjects, and, particularly, on those in the Gravity and Other category. [Recall that Gravity and Other is the only meaning that had a substantial number of EMH subjects in it (9). The only other meaning to meet the "meaning deviation score ≤ 2" criterion had only one subject.] In this section, the fact of fewer data points than subjects implicates subjects who were not asked the question.

Table 2
67% of subjects said forces were different in isomorphic tube and string situations

	E	M	H	G	All
Equal	1	1	2	3	4
Bigger (string)	0	2	3	3	5
True (string)/false	1	0	0	0	1
Different (incomparable)	4	1	1	2	6
False/false	1	1	0	0	2
Collapsed					
Same	2	2	2	3	6
Different	5	3	4	5	12
Total	7	5	6	8	18

Results. Consistent with our expectations, only 1/3 of responding subjects indicated that the forces would be the same in these two situations. In addition, also consistent with our expectations, when a difference was specified, the string was implicated in a stronger force (or, in one case, the string situation involved a force on the ball, and the tube situation did not). Those classified in the quasi-replication as Gravity and Other were spread out nearly equally across three possibilities. Three subjects said the forces would be equal; three said the force on the ball in the string case would be bigger than in the tube case; and two said merely that the forces would be "different," declining to assign a relative magnitude. The latter suggests that subjects might believe forces may be ontologically different in such a way that comparison of magnitude is impossible. To bolster this interpretation, we note that no subjects explicitly said that one force would have to be greater than the other, but that they simply did not know which was which.

Table 2 shows the complete breakout of data for this question by age level, by Gravity and Other grouping (denoted G), and collapsed into summary categories. This analysis provides clear evidence of contextual fragmentation in that subjects are responding to attributes not included in I&V's analysis. Further, it shows significant fragmentation of subject categories in that responses of even the G subjects were spread across several different categories, fairly equally spread across three different attribution patterns. On the basis of these questions, there is little reason to group the single largest I&V meaning group, Gravity and Other, together.

6.1.2. Blocks

Expectations. The blocks set of questions (16–19) asked about various setups of two blocks leaning against one another. For consistency across nearly identical situations, all questions were phrased in terms of forces. Our conjecture was that the leaning block would be construed as agentive, lending to a trend in the data identifying the leaning block as the one applying force, or a greater force if both were seen as applying a force. We call this the "canonical pattern." For present purposes, we analyze only the contrasting pair of a big block leaning on a small block, and a small block leaning on a big one. The canonical pattern should have a force exerted only by the leaning block, or a greater force by that block compared to the force exerted by the support. By contrast, if size were the controlling parameter, there should be no change in attribution of force exerted by the big block on the small one compared to the force of the small exerted on the big one, and we describe this as the "no change" (in attribution based on leaning status) outcome. For reference we note that the canonical pattern is incorrect. The forces are identical as they are "action and reaction" forces.

I&V meanings do not distinguish between "leaner" and "support" roles. There is no motion in these setups, and the use of differing sizes should sort out the relevance of that variable. One might potentially argue that stability could be an issue, that big-leaning-on-little is less stable than little-leaning-on-big, or that the leaning block is, by its nature, less stable. However, among our focal subjects, the Gravity and Other group, only one of five submeanings listed by I&V for Gravity and Other [Gravity, Push/Pull, and Acquired (b)] construed stability as relevant to amount of force, and this was a relatively unpopular construal. Looking directly at our data, of the nine Gravity and Other subjects, all but one (89%) identified forces in stable and unstable situations (question Set 2) as identical. The only Gravity and Other subject to change attribution across stable and unstable situations is an outlier in our data. He was the lone elementary school student who matched the Gravity and Other criteria.

In net, given our own data and those of I&V, and given the set of meanings identified by I&V, we should not expect any effect of leaning status for the Gravity subjects; force attributions should, at best, depend on size.

Results. Seven of eight (87.5%) Gravity and Other subjects changed attribution of force based on leaning status. The sole exception was a high school student whose responses are consistent with the principle that the force of the small block on the big block was always smaller than the reverse. The canonical pattern is quite prominent, with 50% of the Gravity and Other subjects responding consistently with it. This is in line with typical misconception studies that show somewhere around 50% of students have the most prominent misconception among the set of wrong answers (e.g., McCloskey, 1983). However, the "reverse canonical pattern" (leaner applying less force) appeared in two cases (25%). If we uncollapse the smaller/no force distinction, the results are strikingly fragmented. Among the eight Gravity and Other subjects, there are seven distinct patterns of attribution.

In the larger EMH dataset, 13 out of 19 responding subjects (68.5%) changed attribution based on leaning status. Seven out of 19 (37%) showed the canonical pattern, and 2 out of 19 (10.5%) showed reverse canonical. Within the larger EMH dataset, I&V's potential rebuttal—that instability rather than leaning/agency might account for switches—is more viable; some of these subjects did, indeed, respond to the stability/instability distinction. Table 3 shows the complete breakout of data by age level, by Gravity and Other grouping (denoted G), and collapsed into summary categories. Table 4 shows the complete coding for the eight responding Gravity and Other subjects.

Once again, we have validated a predicted contextual boundary and new relevant attribute (leaning construed as agentive). Furthermore, the one I&V meaning that showed up in our data, Gravity and Other, appears substantially fragmented; subjects who were classified as having that meaning did not think alike about this situation.

Table 3
87.5% of Gravity and Other subjects specified a change in force attribution based on leaning status; 68.5% of all EMH subjects specified a change

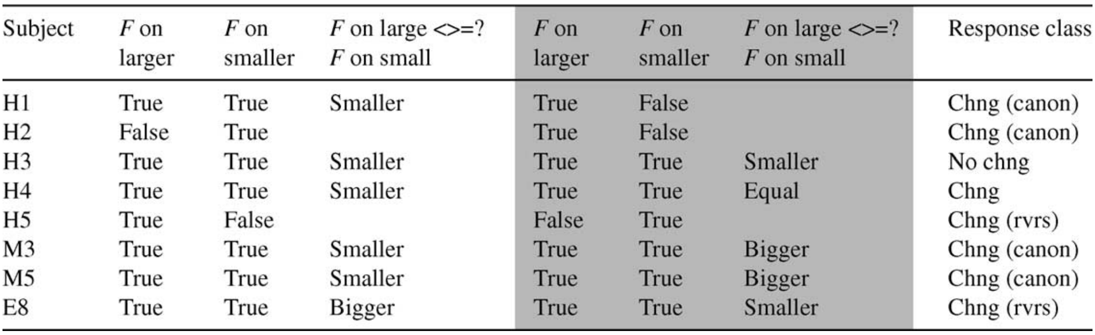
	E	M	H	G	All
Canonical pattern	1	4	2	4	7
Reverse canonical	1	0	1	2	2
Collapsed					
Change (leaning related)	5	4	4	7	13
No change (size related)	3	2	1	1	6
Total	8	6	5	8	19

6.2. Ontological aspect

6.2.1. Expectations

The ontological aspect of a concept is determined by its spatial–temporal properties, and by its relations to objects. The methods used for this study are not well-adapted to identifying ontologies. In particular, categorical questions (“Is there a force? Is this force greater than that one?”) do not provide much information about the nature of forces. Nonetheless, we did ask one question directed at eliciting some ontological information, with expectations of exposing ontological properties at variance with any of I&V’s meanings. We asked whether there was a force in the situation of a struck bell. The protocol made clear to subjects that we meant to attend to the situation after the bell was struck (we were not interested in the strike of the bell). Our expectation was that the sound of the bell would cue an agentive attribution, and thus subjects would implicate a force in situation where none of the characteristics that implicate forces involved in I&V’s meanings are apt. A force stemming from vibration or from sound from the bell cannot be the direct result of size or weight (e.g., it dies away with time); there is no overt motion; stability is not an issue; there is no persistent agent

Table 4
Response patterns of Gravity and Other subjects to leaning blocks questions



Subject	*F* on larger	*F* on smaller	*F* on large <>=? *F* on small	*F* on larger	*F* on smaller	*F* on large <>=? *F* on small	Response class
H1	True	True	Smaller	True	False		Chng (canon)
H2	False	True		True	False		Chng (canon)
H3	True	True	Smaller	True	True	Smaller	No chng
H4	True	True	Smaller	True	True	Equal	Chng
H5	True	False		False	True		Chng (rvrs)
M3	True	True	Smaller	True	True	Bigger	Chng (canon)
M5	True	True	Smaller	True	True	Bigger	Chng (canon)
E8	True	True	Bigger	True	True	Smaller	Chng (rvrs)

Notes. White background denotes larger block leaning on smaller. Gray denotes smaller leaning on larger. Response class is marked “Chng” if attribution is changed based on leaning status. Response is “canon” (canonical) if it matches the predicted trend, “rvrs” (reverse) if it is opposite the predicted trend, and “No chng” if there was no change.

pushing or pulling; and, it turned out, no subjects mentioned any relation between vibration and gravity.

6.2.2. Results

About half of the subjects did, indeed, identify a force in/on/around the bell, somehow associated with vibration (primarily) or the sound produced (mentioned usually secondarily, in conjunction with vibration) H: 4/6 (67%); M: 3/6 (50%); E: 4/8 (50%); P: 0/7 (0%). The force died away with the sound of the bell. Of those subjects identifying a force, most identified “vibration” (and used that word) as the force involved.

We believe that a “force of vibration” necessarily has a different ontology, in the sense we use the term here, than any of the forces mentioned by I&V. Consider, in sequence, the ontological (or near ontological) natures of each of I&V’s core meanings:

- Internal Force: Internal Force is inherent in an object by virtue of its weight or size. But neither weight nor size is implicated in vibration, a point that is further underscored by the fact that the vibration force autonomously dies away with time.
- Acquired Force: Acquired Force is inherent in a moving object by virtue of its gross (rectilinear or circular) motion. Vibration is not a gross (center of mass) motion. Furthermore, several subjects did not mention and did not seem to know that vibration involves any sort of object motion at all. Finally, one interpretation of Acquired Force is that it is a characteristic of one object (e.g., a tossing hand) that is imparted to the object. Yet, the hammer that imparts vibration is not vibrating.
- Push/Pull Force: Push/Pull Force is inherent in a persistent relation between an object and the agent that pushes or pulls it. No between-object relation of any sort is implied by vibration.
- Gravity: Gravity is probably a particular persistent relation between an object and the earth. At least, it is by Newtonian standards. Vibration implicates no such relation, and, whatever gravity’s ontological characteristics, no subject implicated a gravitational source or connection to vibration.

More simply, we observed that most subjects who saw a vibration force also believed that the force moves around, on its own, independent of objects. The modal spatial description of this motion was “outward in all directions” from the bell. Some subjects had the force also circulating within the bell. A force that moves on its own, independent of object motion, is ontologically distinct from any of the meanings discussed by I&V: it is neither inherent in the object (Internal, Acquired), nor relational (Push/Pull, Gravity).

Samples from interviews:

H1: **Circulation of vibration force**.

H1: The sound obviously comes from the shape of the thing, and the vibration force goes throughout it.

H2: **Vibration force dies away**.

I: After the hit, is there a force on the bell?

H2: I think there’s a little bit, that gradually goes away.

M6: Vibration force, but no indications the force moves around.

I: Would you say there are any forces here?

M6: Yeah.

I: What would that be?

M6: I'd say it's vibrating, which is causing it to ring.

I: So that's a force?

M6: Yeah.

M5: Emanating, circulating force. Rejected relational and inherent ontologies.

I: Is there any force.

M5: Yes. The sound vibrations that are coming off of the bell.

I: Would you say that's a force on the bell?

M5: No, it's more of a force radiating from the bell. But it does bounce inside, like it bounces in and around.

M2: Emanation of vibration force. No indication that vibration, itself, is a motion.

I: After I hit it, would you say there's a force.

M2: The vibration.

I: Is that a force?

M2: Yeah.

[later]

I: Is that a force on the bell?

M2: Yes, it's vibrating out. [gestures with one had moving away from bell]

[later]

I: While it's ringing, is there any movement there?

[clarification omitted]

M2: The vibrations are moving around causing you to hear it.

I: The vibration's moving around. Do you know how they move around?

M2: They're moving out.

[A few moments later, she clarifies "out" with a gesture—outward from the bell.]

I: Is there anything moving?

M2: Yeah, the sound waves coming off the bell are moving.

E1: Vibration force in and emanating from the bell.

I: Would you say there are any forces, now?

E1: M-hm. *[Yes]*

I: What forces are those?

E1: Vibration from the bell that goes out into the air.

[later]

I: Is anything moving, after the hit?

E1: Yes, when you hit it, it vibrates.

I: And is that a force?

E1: Umm, yes.

I: That vibration is a force?

E1: M-hm. *[Yes]*

Overall, we identify a distinct ontology in our subjects' responses. A (vibration) force that moves around according to its own principles—independent of objects or objects' motions—is neither relational nor inherent in an object. Four of the six citations, above (H1, M5, M2, E1) are entirely consistent with this ontology. Vosniadou explicitly excludes the ontology as relevant to I&V's meanings when she says, "All of these meanings of force were constrained by the underlying presupposition that force is a property (inherent or acquired) of physical objects" (Vosniadou, 2002, pp. 74–75). We also saw this ontology in responses to questions about multiple pushes on a block. In the case of two oppositely directed pushes simultaneously applied, several subjects explained that the forces moved through the block toward the center and cancelled at that point. This is a strikingly sophisticated ontology—a moving, conserved vector quantity, just as we described the expert ontology in Section 2 (see also diSessa, 1980).[25]

Finally, we note both fragmentation of I&V's "meaning groups" and a core contextuality. Fragmentation appeared in that the attribution of a force evenly split the Gravity and Other subjects: 5 saw a force; 4 did not. Contextuality is evident in that the vibration force (or, in general, autonomously moving force) appeared in only very particular contexts—prominently only in this bell question, and, to a much more limited extent, in multiple force questions (directly below).

6.3. Compositional aspect

6.3.1. Expectations

With respect to composition, we mentioned that experts see forces combining in one general and very well-defined way: vector addition. In prior studies we observed subjects seeing forces combine in a variety of ways. One force acts on another, reducing or modifying the effect of the other. Forces might dominate, or overcome other forces, and so on. In these interviews, also, some subjects showed interference and influence of one force on another, as opposed to pure composition (on questions other than the simple block composition, below), but we do not systematically report these here.

To look at compositional effects, we asked subjects about cases of a single or multiple simultaneous forces on an object in a simple and familiar situation: a person pushing on a block. Fig. 5 shows schematically the various configurations. We expected to see a general development from less sophisticated strategies of composition, such as "overcoming" (stronger "gets its way"), or "alternating effects" toward more sophisticated ones, which approximate vector addition.

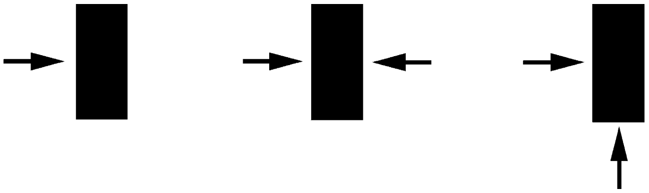


Fig. 5. Pushes are applied to a block. In the case of opposite forces, we asked about pushes of equal magnitudes. In the case of perpendicular forces, we asked about equal pushes, and with about half the subjects also about unequal pushes.

6.3.2. Results

All EMH subjects answered that a single force resulted in motion in the direction of push. All high school subjects answered in the canonical way: A single force results in motion in the direction of the force; equal opposing forces cancel; orthogonal forces result in diagonal motion. We did not ask high school subjects about unequal orthogonal forces.

Progressing toward younger subjects, there was a general trend toward more non-canonical answers overall, and younger students gave more non-canonical answers to simpler questions (although, not even the preschool subjects gave a non-canonical answer to the single force question). Of the middle school subjects, four of six subjects gave canonical answers to all questions, excluding for the moment the unequal orthogonal forces question. One of those four also gave a canonical response to the unequal orthogonal forces (weighted diagonal motion), one said the block would spin under the influence of unequal forces, and two were not asked about the unequal case. The two middle school subjects who gave non-canonical responses to equal orthogonal forces (alternate, cancel) also gave non-canonical answers to unequal orthogonal forces (bigger wins).

More non-canonical answers were provided by elementary students, and they answered simpler questions in non-canonical ways. Three subjects (out of seven respondents) gave the canonical answer on one force, opposing forces, and equal orthogonal forces. A greater proportion of elementary subjects gave categorical answers ("bigger wins," "smaller wins") to the unequal orthogonal forces question compared to middle school subjects, although the numbers are small and certainly not statistically significant. One elementary school subject gave "cancels," and one subject gave "smaller wins" response to unequal orthogonal forces, which we consider particularly primitive. Marginally more elementary school subjects gave "cancels" or "alternates directions" answers to equal orthogonal forces (3/7) compared to middle school students (2/6). While no middle school subjects gave non-canonical answers to equal orthogonal forces, three of seven elementary school students did. Non-canonical answers to this question included "back and forth," "wiggle," and "don't know." Table 5 gives the full results for elementary and middle school subjects. Recall, all high school subjects answered canonically for all questions asked, but none were asked about unequal orthogonal forces.

This analysis did not split the high school subjects, who all answered canonically. Of the other three Gravity and Other subjects, only one answered all questions asked canonically. On the other hand, we see a fairly complex pattern of development, about which I&V's meanings are silent. One of the more prominent non-canonical responses is "bigger wins." An apparently

Table 5
Responses to forces on a block

Subject	Single force	Opposite equal	Orthogonal equal	Orthogonal unequal
M1	Parallel to force	Cancel	Diagonal	Spin
M2	Parallel to force	Cancel	Diagonal	*Weighted diag*
M3	Parallel to force	Cancel	*Cancel*	*Bigger wins*
M4	Parallel to force	Cancel	*Alternate*	*Bigger wins*
M5	Parallel to force	Cancel	Diagonal	
M6	Parallel to force	Cancel	Diagonal	
E1	Parallel to force	Cancel	Diagonal	
E2	Parallel to force	*Back and forth*	Diagonal	*Weighted diag*
E3	Parallel to force	*Don't know*	*Alternate*	*Bigger wins*
E5	Parallel to force	*Wiggle*	*Cancel*	*Smaller wins*
E6	Parallel to force	Cancel	*Cancel*	*Bigger wins*
E8	Parallel to force	Cancel	Diagonal	*Bigger wins*
E9	Parallel to force	Cancel	Diagonal	*Cancel*

Notes. Non-canonical answers are italicized. Blank cells represent questions not asked. All high school subjects (not shown) gave canonical answers to all they were asked.

even more primitive class of responses might be glossed as “the object cannot decide” (or, possibly, “interference” is seen): the object either stays still (cancel) or alternates between directions. Preschool responses included the category “one way or the other.” A few non-straight-line responses were recorded, including “wiggle,” “spin,” and “object will break” (one preschool subject).

6.4. Causal aspect

6.4.1. Expectations

The final analysis we present concerns causal inference. What follows from the existence of a force? The Newtonian response is that acceleration is determined by force (and mass). A typical novice response is that speed, not acceleration, is determined by force, which is a well-documented misconception (e.g., Viennot, 1979). In our extension study, we aimed to display a simple, previously documented fact (diSessa, 1993). Students are sensitive to the context of application of a force, and distinguish their expectation accordingly. In particular, forces in circularly oriented situations cause spinning, not rectilinear motion. Fig. 6 shows the situation we used. We used a physical prop, not a drawing.

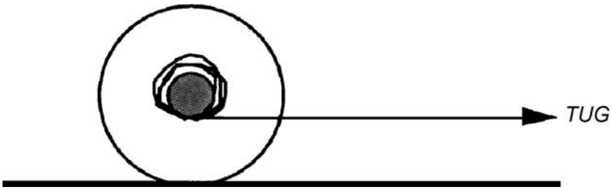


Fig. 6. The string on a yo-yo is gently pulled toward the right.

Table 6
Responses to the yo-yo problem

	Back and forth	Aligned	Opposite
High school	0	0	100% (6/6)
Middle school	0	17% (1/6)	83% (5/6)
Elementary school	12.5% (1/8)	50% (4/8)	37.5% (3/8)
Preschool	0	67.5% (5/8)	37.5% (3/8)

What direction will the yo-yo move, compared to the tug?

6.4.2. Results

As documented above, all PEMH subjects gave canonical answers to a single force on a block: The block's motion is aligned with the force on it. With respect to the circularly oriented yo-yo problem, development across different ages appeared to progress from equally-likely selection of motion aligned or opposite the force—or possibly even favoring "aligned"—to uniform expectation that the yo-yo will spin in the "obvious" way, and thus roll *opposite* to the force. Table 6 shows the detail.

Strikingly, this represents a reverse development in the sense that the consensus answer of high school students is incorrect. The yo-yo actually moves in the direction of the force. Thus, a U-shaped development is implied here; younger students are more correct than older ones, and, presumably at some point in instruction students will return to favoring the correct answer. Our interpretation of this U-shaped development is that, as people get older, they have more experience with spinning things and become more conceptually competent with their behavior. They gradually come to understand that, in such contexts, forces spin, they do not (primarily) move. This is a contextual refinement that is important to carry into physics instruction. Circularly oriented situations are different from rectilinear ones. However, even the existence of an appropriate contextual divide does not entail correct use of the relevant inferences.[26]

This example makes a simple point. People can and do have contextual boundaries on causal inferences with respect to force and motion. Of course, no claim can be made that circular versus linear is the only divide. In fact, Knowledge in Pieces suggests that that is unlikely. With respect to I&V's analysis, they take on no accountability for the specification of any inferential aspects of the meaning of force, so their data and theory are silent on this aspect. Even if their meanings survived contextual scrutiny, it seems extremely unlikely that the here-documented reverse developmental phase in an inferential aspect of force could be brought under the explanatory umbrella of their apparently monotonic developmental sequence of meanings.

7. Discussion

7.1. Quasi-replication

The principal results of our quasi-replication, shown in Figs. 3 and 4, are strongly at variance with I&V's results. Even allowing a generous error allowance of 20% (two of our 10 questions), very few subjects matched any of I&V's meanings. The statistics on this are clear. There is virtually no chance ($p < 10^{-5}$) that I&V's data and ours could have arisen from a population having the same proportion of classifiable subjects. Furthermore, the only remotely viable meaning, Gravity and Other (which embraced 9 of 10 EMH subjects who met the 2 deviations out of 10 criterion), seems vague and too forgiving of differences in conceptualization to be a viable "meaning." Our extension study consistently showed non-trivial differences in the way Gravity and Other subjects thought about force and motion.

Further analysis of our data suggested that obvious candidates for an underlying cause of difference do not pan out. Essentially nothing changed if we removed the comparison questions I&V did not ask. With regard to the question of how much effect introducing linguistic variety in terms of alternating "forces" with "pushes or pulls" had, our internal analysis suggests that it made little difference. Few subjects changed when the same question was asked in two different ways (10–15%), and about the same number of subjects changed in the case an identically phrased question was asked at two different times in our analysis. Furthermore, classifying subjects only on the basis of questions phrased in terms of force produced negligible change in classification (e.g., three more EMH subjects moved into the 20% deviation range, while two moved out of it). Finally, we found comparatively weak alignment among subjects, independent of I&V meanings, not enough to suggest (much less determine) "hidden meanings" that might reduce apparent diversity.

While these results are strongly in favor of "fragmentation," they surprised us. In reaction, we double-checked our mappings, our codings, 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 possibilities worth considering:

- *Interviewing technique*: I&V expected students' answers would provide simple, direct confirmation of a few meanings. We, obviously, had no such expectation. 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.
- *Coding*: We noted earlier that I&V's coding is in some small degree holistic. That is, they do not independently code different items in a question set (e.g., forces on two objects plus comparison), but assign codes that entail a pattern of answers. While holism is minimal, "best match" coding on I&V's part might amount to forcing unobserved coherence. I&V coded explanations in addition to the features we coded (existence, relative size, and source). However, this should have inevitably resulted in our results being *less* fragmented than theirs.

  The next two possible explanations involve differences in subject populations.
- *Instructional differences*: Few U.S. students receive any systematic instruction in force and motion before high school, and none of our subjects 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 these subjects were. We note also that

many more U.S. subjects invoked gravity compared to I&V subjects, which is opposite of the effect more instruction is likely to have.

- *Language*: Perhaps the most intriguing idea emerges from the differences between Greek and English. In particular, the Greek word for force, *dynamis*, connotes strength and power, in addition to being the technical term for physicists' force. "Strength" is more like what young I&V subjects attributed to objects (Internal Force). We are in no position, yet, to evaluate this hypothesis. If it were true, however, it would suggest a striking Whorf–Sapirian effect—the natural language one speaks deeply affects conceptualization.

The language issue is generally, as we noted, thorny. 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 we and I&V have entirely escaped the problem of language *as opposed to* conceptualization, it would be quite surprising.

7.2. Extension study

Our extension study used a simple strategy, "ask more questions," in order to show: (1) fragmentation of categories of people (e.g., Gravity and Other adherents), (2) fragmentation of contexts (systematic differences in response based on features not mentioned in I&V meanings), and (3) inferences and reasoning that are not predictable on the basis of I&V's meanings (extending to ontological, compositional and causal aspects of force). As we pointed out, while our listing of aspects of specification involved in item 3 is new, results in most of these categories constitute the basis of many misconceptions that are documented in the literature, and are claimed to be blocks to instruction. While I&V could in principle claim that these aspects are irrelevant to an apt characterization of the meanings they find—that is, their meanings simply do not have these aspects—such a claim would be problematic for the use of such meaning to understand conceptual change leading to the instructed meaning of force. Manifestly, the instructed meaning *does* have these aspects, and furthermore, we see, in virtue of our experiment, that people of various ages have them as well, even if they are not part of an integrated "concept of force." Conceptual change theories must eventually encompass how students get from their uninstructed ways of thinking about force and motion to the normative concepts. In this light, the best we could say about I&V's meanings (even if they proved cogent) is that they are substantially incomplete in mapping the full naïve and novice knowledge state concerning force and motion. This observation is elaborated in the last section, "Qualifications and Final Words."

In the following, we review our results by problem context, indexing them by the failure patterns (1–3) described above.

Ball in tube versus ball on string: These situations are identical with respect to all of the attributes mentioned in I&V's study. Nonetheless, 70% of EMH subjects said that forces were different in these situations (pattern 2). Furthermore, the Gravity and Other subjects spread their answers relatively equally across three different responses, lending no support to the contention that their thinking about force is homogeneous (pattern 1).

Leaning blocks: Seven of eight Gravity and Other subjects (87.5%) changed attribution based on leaning status, and nearly 70% of all EMH subjects did so (pattern 2). Uncollapsing "small" versus "no force" categories, seven of eight Gravity and Other subjects exhibited distinct patterns of attribution (pattern 1). On several bases, leaning status cannot be absorbed by the influence of the "stability" attribute discussed by I&V.

Sound in a struck bell: About half of subjects implicated a "force of vibration" that had ontological properties unlike those in any of I&V's meanings (pattern 3). In particular, the force flowed in a manner (outward from the bell) that was independent of object motion; it tracked neither the back and forth of vibration nor the center of mass motion.

Simultaneous forces on a block: Our cross-sectional study showed a complex development from non-canonical answers such as "bigger wins," "alternates," "will break, wiggle," or even "equal orthogonal forces cancel" before high school subjects settle in on a qualitative version of vector addition. These responses are not specified by I&V's meanings (pattern 3), and it is difficult to see how they could be at all constrained by those meanings. Thus, we argue that composition represents an at least tentatively independent dimension of the maturing concept of force.

Yo-yo: The yo-yo problem shows definitively that subjects distinguish between "circular" and "rectilinear" situations in projecting the causal effects of a force (patterns 2 and 3). Circular situations yield a primacy of "spinning" over "moving" as the causal result of a force for older subjects. Once again, there is no obvious relationship between the developmental trajectory shown in these data (including a U shape pattern, more wrong answers in older students) and I&V's meanings (pattern 3). Like the simultaneous forces question, this one did not show pattern 1.

8. Summary and conclusion

8.1. Review: theory

The framing theoretical issue in this paper is how much and what kind of accountability for details in conceptual change must conceptual change researchers take on. In this frame, we have argued, in particular, that two foci for accountability exist, and are highly consequential for data collection and theoretical analysis. First, *contextuality* concerns how people respond to different situations, different ways of asking questions, and so on. It involves what attributes of contexts are responded to, and in what way. Statically, we argue that researchers of conceptual change need to specify something about the ranges of application of elements (concepts, theories, etc.), overlap, and more. Dynamically, we want to know if subjects have multiple ways of conceptualizing a situation, what ways are initially given and, probably more importantly, which (if any!) are preferred after extended consideration. We need also to attend to perspective (as we called it), including modality of presentation, modality of response, focus of question (explanation, prediction, etc.), stability in response to counter-suggestion, and so on.

The second focus for accountability is *specification*: What does one need to say to have adequately specified the nature of a concept, meaning, or theory? We offered one particular delineation of specification, which happens to be particularly apt for the concept of force, but which also applies to any physical quantity (and therefore to many of the concepts attended to in conceptual change research, like force, velocity, acceleration, heat, and temperature).

A specification of a quantity involves multiple aspects: (1) existential—under what conditions is the construct seen to exist in a situation?; (2) coarse quantitative—what absolute or comparative specification of size can be given among various instances of the construct?; (3) ontological—how do we describe "the nature" of the construct, in particular, its defining spatial–temporal properties (directionality, conservation, and locational properties) and its family of relations to objects?; (4) compositional—how do multiple instances combine, if co-present in a situation (interference, influence, or pure composition)?; (5) causal—what implications (e.g., among entities of other ontologies) can be inferred from the presence of an instance of the construct? We noted that, with respect to the concept "force," each of these aspects is easy to specify for the instructed concept—hence plausibly constitute relatively independent goals for instruction—and that the body of literature on conceptual change of force documents naïve-novice/expert differences in all these aspects, with the exception of (previously uninvestigated) composition. The literature also implicates these differences as consequential for instruction.

The particular context in which we argue for the importance of accountability for both contextuality and specification is the long-standing debate about whether naïve concepts (theories, etc.) are fragmented, or coherent. On the background of contextuality and specification, we argued that fragmentation and coherence are probably not well-defined in themselves, and in any case they require much more complex description than simply "fragmented" or "coherent." Some of the issues we raised concerned: the difficulty in determining range of applicability, without (at least somewhat arbitrarily) taking on a preferred conceptual "frame of reference"; the importance and plausibility of nuanced issues such as the "logic" of contextual boundaries (e.g., whether "principled" and internally treated by the theory, or "unprincipled" and dictated substantially by complexity of the domain and developmental history); overlaps in context, and instability of boundaries. We also raised the issue of relational structure—how do the different elements and aspects of a concept or theory relate to one another?—in addition to contextuality and specification. While relational structure is highly relevant to settling the debate about fragmentation versus coherence, its treatment here is minimal, owing to the limited impact of the sort of data collected for this work on relational questions.

In order to avoid some of the problems of incommensurability among different paradigms of conceptual change research, we aimed to do empirical work using the same conceptual focus, the same age ranges, the same empirical and analytical methods, and even using many of the same questions as prior work. In particular, we capitalized on recent work by Ioannides and Vosniadou (2002). I&V's work inherited the theoretical framework long advocated and substantially empirically supported by Vosniadou (e.g., Vosniadou & Brewer, 1992, 1994). In brief, Vosniadou claims that specific theories, models or meanings used by subjects are strongly constrained by framework theories, which are entrenched and difficult to change. The constraints that are posed by framework theories on specific theories (models, meanings) are such that only a few possibilities exist, typically about a half-dozen. I&V's recent empirical work seems strongly to support claims of the Framework Theory model. Almost 90% of subjects were classifiable as systematically adhering to one of seven "relatively narrow, but coherent" meanings of force.

Theoretically speaking, I&V do not take on many of the accountabilities that we argue are important. This is not a debate among different ways of being accountable to specification or contextuality. We have intended to remain relatively open about how one treats these, while still being committed to treating them (and adding what we consider relatively uncontroversial specificity so as to achieve empirical tractability). I&V give no account of their principles of specification. In terms of our principles of specification, I&V's data deal mainly with existential aspects of force. They deal with coarse quantitative aspects to a limited degree, and not at all with ontological, compositional, or causal aspects. In terms of contextuality, I&V's treatment is minimal. They provide no specification of the intended contextual range of the meanings they present (although, we feel the implication is clear that they are staking a claim to a relatively full specification of the meaning of force across a sensibly broad range of contexts). The logic of their arguments strongly suggests that other contexts should not evoke a large number of other meanings and responses. Otherwise, how could these data be put forward, as they are, as bearing significantly on the fragmented versus coherent debate? I&V do say these meanings are "narrow but coherent." But "coherent" is not specifically defined, and neither detail on what is meant by "narrow" nor any operationalization of the term is given.[27]

I&V use our own Knowledge in Pieces perspective (diSessa, 1988, 1993) to represent the "fragmented" side of the "fragmented versus coherent" debate. Not surprisingly, Knowledge in Pieces takes on more accountability with respect to contextuality than does the Framework Theory model. In particular, we argue that hundreds or thousands of intuitive elements are involved in the naïve take on the conceptual field of Newtonian mechanics. However, at least some elements of wide scope exist (e.g., extending well-beyond the contextual range of Newtonian mechanics, and being frequently evoked). Furthermore, because of developmental complexities (e.g., the developmental forces that shape contextuality of different elements are different), the contextual boundaries of elements should be expected to be unprincipled (by physics standards), which leads to overlapping contexts, dynamical instability, and other relevant phenomena. Compact specification of the content of "intuitive physics" (say, a compact natural language exposition) is impossible owing to these contextual complextities, to the number of elements involved, and to the problematic relation of p-prims to natural language. These theoretical commitments make element-by-element specification of contextuality difficult, although system-wide issues such as conflict, many elements, dynamic instability, and so on, are much easier to document. By the same token, a global specification of the concept of force (which we use as a reference to interrogate the completeness of I&V's theory and data) is much more difficult and less useful concerning an element-by-element (p-prim by p-prim) specification of intuitive conceptualization.

8.2. Review: empirical results

The empirical work developed here has two parts. First, we attempted a quasi-replication of I&V's work, which included minor modifications. The second empirical strategy, involved in our extension study, was built on prior work in order to: (1) show a diversity in ways of thinking among even those who came closest to matching I&V's categories, (2) demonstrate that more features of the world were involved in judgments about force and motion than contained in I&V's seven-meaning analysis, and that differences in those attributes draw out different patterns of reasoning in subjects, and (3) to expose aspects of the concept of force to which I&V did not attend. Item (3) begs the question of how I&V's meanings could be extended to cover the aspects of subjects' reasoning exposed in these new questions (e.g., what are subjects' takes on composition; how does a "flowing quantity" ontology relate to Acquired, Gravity, or other meanings?).

Our quasi-replication failed in the sense that our results are statistically exceedingly far from I&V's results. Rather than 90% of subjects classifiable as answering in accordance with one of I&V's meanings, less than 50% of our elementary, middle and high school subjects were classifiable, even allowing an error rate of 2 questions out of 10. Furthermore, 90% of those who met the 2 out of 10 mismatches criterion belonged to the Gravity and Other meaning category, which is vague (and which the subsequent extension study revealed to be a diverse category, conceptually speaking). None of the differences we introduced between our quasi-replication and I&V's seemed capable of accounting for the differences in the data.

Of course, from the Knowledge in Pieces perspective, finding data that suggest I&V's experiment will not robustly replicate, and that the systematic difficulty is inherent diversity in subjects' responses, is congenial. Nonetheless, the stark contrast is unsettling, and we plan follow-up work investigating methodological issues.

Even disregarding the divergent results from our quasi-replication, our extension study provides, we believe, a very strong suggestion that additional accountability in the way of specification and contextuality is needed if adherents of coherent, non-fragmented views of conceptual change are to make their case. Essentially all of our expectations were confirmed. (See question-by-question listing in "Motivation and Questions," in Section 4, and subsequent reporting of results.) The results undermined I&V's meanings as being a full specification of a concept of force, and they highlighted lack of accountability for contextuality by showing diversity in the data across contextual boundaries that were unspecified by I&V, and across subjects.

8.3. Qualifications and final words

While the theoretical analysis and subsequent empirical work here should give pause to coherence or Theory Theory advocates, we caution against drawing too extreme implications from this work. First, we do not believe conceptual change is a homogeneous phenomenon. In particular, for example, "models of the earth's shape" (viz. Vosniadou's earlier work), we feel, is dramatically less connected to the kind of rich phenomenology of everyday experiences of force and motion that lie beneath intuitive conceptions in mechanics. The Knowledge in Pieces perspective was developed specifically to deal with experientially rich domains, such as mechanics. Application to other domains is at least somewhat speculative, and possibly even doubtful, without that inherent richness. Thus, results outside mechanics favoring the coherence view, including earlier work by Vosniadou, are not immediately threatened, although the quality of their "validation" is undermined by the arguments we gave for increased accountabilities.

It is also easy to caricature the meaning and entailments of Knowledge in Pieces as prescribing knowledge that is completely "inconsistent, unstable over time and problem context, and infinitely malleable" (Samarapungavan & Wiers, 1997, p. 147), or a "fragmented, inconsistent jumble" (p. 179). Individual p-prims constitute cognitive regularities that are repeatable and stable. While we warned against uncontrolled perspective in question asking or problem posing, we believe p-prims are stable enough to be relevant both to clinical interviews, to actions of subjects in physical situations (unmediated by language), and to classroom instruction (e.g., diSessa & Minstrell, 1998). Furthermore, the resulting knowledge subsystem is not at all without character, has experimentally calibrated systematicities (diSessa, 1993), and, far from being infinitely malleable, results, for example, in learners' finding instructed physics often to be incomprehensible, or worse: comprehensible but seeming to be irretrievably false. The problem is not that there is pure incoherence in naïve thought, but that Theory Theories seem grandly to overestimate coherence and simplicity, running roughshod over contextual boundaries, and expecting that a few sentences can characterize a rich, complexly adapted knowledge system. If our claims are valid, implications of Theory Theory views for instruction will be limited to rough heuristics for dealing with the fact of naïve preconceptions, rather than any detailed account of particular conceptual difficulties.

We do not believe I&V's meanings are without merit or sense. The Acquired meaning in the guise of impetus-like responses to trajectory problems is well-established. The idea that an object (viewed as a generalized "agent") shows weakness if it is moved is consistent with some data in Piaget's work (Piaget & Garcia, 1974), and with work on the core conceptualizations underlying language (Talmy, 1988). That big things are strong ("forceful") and the propensity to cause damage is associated with strength ("forcefulness") is entirely consistent with p-prim-like attributions children might make. The problem comes when these ideas are taken to be well-developed (e.g., relationally strongly coherent), few in number, broad in scope (comparable to Newtonian mechanics), and compactly specifiable with respect to their meaning and contextuality. Hence, we would be strongly tempted to reanalyze responses I&V obtained in terms of p-prims, confluences of p-prims (e.g., the "impetus theory," see diSessa, 1993, 1996) or in terms of other structural elements of intuitive knowledge (diSessa, 1996).

An alternate interpretation of the realities of I&V's meanings might be that they are not at all at the p-prim level (as above), but represent broad, general frameworks that are coherent, but simply do not specify the "details" represented by our extension study. "Narrow," in this interpretation, might not have anything to do with contextuality at all, but, instead, it might describe specification. Some aspects of students' ideas are specified by the general framework, but others are not. So, the data obtained here might attend to details, but ignore the core background: vague, but influential framework theories.

This interpretation, however, is weak. What is fundamental and what is "details" requires operationalization and demonstration, not gestalt judgment. If I&V give preference to their meanings, they then imply that other aspects of the concept of force that they do not specify (e.g., causal, ontological, and compositional) are less important or unimportant. As we pointed out, conceptualization associated with these other aspects has been implicated in the literature as important barriers to conceptual change. Furthermore, earlier in the paper we identified the relevant expert aspects of conceptualization, and contrasted them with what we found in subjects. So it is perfectly clear that significant change in them must take place. At a minimum, I&V simply have not made the case that their meanings (dealing only with existential and coarse quantitative aspects) are more important than other aspects of meaning.[28]

More generally, the crucial "importance" is prospective, what role do the specified conceptualizations play in coming to understand Newtonian physics? In this regard, "details" outside I&V's "narrow but coherent" conceptualizations might turn out to be centrally involved in conceptual change. Brown and Clement's results (1989), cited earlier, provide a nice model. A minor aspect of naïve conceptualization (appreciation of compressed springs' pushing back) might seed a key Newtonian conceptualization simply by instruction's widening the range and importance of a "little" naïve idea. diSessa (1980) observes that the flow ontology, which was not noted by I&V and probably is, indeed, a minor part of naïve conceptualization, might well-seed a core and important Newtonian conceptualization. These are not small matters, in our view. If a substantial proportion of the resources available for conceptual change do not come from the few "coherent meanings" that are charted by researchers (even assuming that such meanings constitute a central part of naïve conceptualizations), but instead come from what might be interpreted as "noise" or "minor" aspects of naïve conceptualization, then, ironically, researchers will have charted more what is *not* used in conceptual change, rather than what *is* used. Where *do* the resources for *new* conceptualization come from?

These issues of importance and completeness do not even arise if I&V's meanings do not survive empirical tests of cogency and reasonable contextual range. The present work implicates a large number of contextual attributes that may change students' conceptualization. Even the quasi-replication, which did not involve new attributes, did not confirm I&V's results. So the very integrity of I&V's meanings remains unsubstantiated, before we get to the question of relative importance compared to other aspects of conceptualization.

Given all these considerations, we much prefer the interpretation of I&V's meanings as being p-prim-like (contextually limited), or p-prim related regularities for which I&V take insufficient responsibility for establishing contextuality, and which ignore a wide range of competing conceptualizations that establish how subjects reason about other aspects of force, and in other circumstances.

The theoretical framework developed here highlights how much is still not known about the intricacies of conceptual change. We do not put forward our discussions of contextuality, specification and relational structure as definitively specified and decisive. We present no general view on specification, beyond what seems sensible and necessary at this time in the context of quantity concepts, like force. Our comments on contextuality were limited to example issues, and even those were treated in a rough-and-ready way. Relational structure was named, but scarcely opened at all to possibilities and results. Nonetheless, the empirical work here emphasizes that unless the field comes to grips with increased accountabilities along these lines, excellent but accidentally narrow work that is insufficiently specified and insufficiently defended may proliferate. Following up these three foci—contextuality, specification, and relational structure—constitutes a substantial, long-term program of work.

Notes

1. There are other reasons, which will be introduced later in the discussion.
2. Classroom debate is taken to be profitable in both these views (e.g., diSessa & Minstrell, 1998). However, with the "fragmented" view, it is implausible that students could "decide to reject theirs and believe the correct theory," as makes sense within a view of conceptual change that implicates coherent, systematic ideas (theories) both pre- and postinstruction.

3. "Limited in number" means both (1) a limited total number across subjects, and (2) a limited number within subjects. Within subjects, Vosniadou claims subjects typically have exactly one model or theory, which is part of her argument for coherence.
4. Vosniadou does not explicitly address the issue of compact characterization, but we feel that, in view of the way she describes naïve ideas, she plainly adheres to the view that such characterizations of students' conceptions exist.
5. A clean example is given in diSessa (2002). A subject at first sees "re-equilibration" as governing a situation, but then loses track of that way of viewing the situation. See also the description of J's alternation of a one-force and a two-force model of a toss in diSessa, Elby, and Hammer (2002).
6. Brown and Clement show that students understand that objects supporting others must push up in some (springy, agentive) circumstances, but they do not naturally extend this to the case of "hard" supporting objects. With a relatively small amount of instruction, however, students learn to see springiness in "hard" situations, which makes the normative idea of pushing up compelling in the case of hard support, as well.
7. We are suppressing another real and complicating issue. A subject might well ask, "for what purpose?" if we ask whether two situations are the same. Purpose sensibly indexes same or different—it is an aspect of what we called *perspective*—hence it influences the "counting" and "size" of contexts.
8. In diSessa (1993) relational structure is called systematicity. See that publication for a more extended treatment.
9. Coherence, as defined technically by Thagard (2000), is a good example of a specification of relational structure. Thagard specifically mentions the distinctions that we used here to illustrate different relational structures. In his scheme, "deductive coherence" is not the same thing as "analogical coherence."
10. "Incoherent" in this sense is not at all necessarily a demeaning characterization. Instead, it may mark the best-possible response to essentially rich and complex domains—where integration (e.g., by a common scientific theory) is the ultimate and rare human accomplishment.
11. The embedding of concepts in larger structures, such as theories, is important to many theorists of conceptual change, such as Carey (1985) and Medin (1989).
12. For example, Newton's third law—that the existence of a force of A on B implies the existence of an equivalent but opposite force of B on A—might be argued to be true by the very nature of the physical meaning of a force (hence, classified as ontological). This position is argued as tenable in diSessa (1980). Newton's third law might alternatively be considered causal [novices, at any rate, interpret the existence of one force (action) to cause the existence of the other (reaction)], coarse quantitative (it establishes a reference value for certain pairs of forces), or perhaps it should provoke the creation of a new inferential aspect, *intrinsic inferential*, which consists of cross quantity (but not cross ontology) inferences relating forces or aspects of them.
13. Thagard's (2002) model of explanatory coherence might fail to apply to this case precisely here. Explanatory coherence applies to *propositional* systems, and not obviously to "sensory-motor" ones.

14. See the development section of diSessa (1993) for some partial results.
15. Ioannides and Vosniadou actually focus on “meanings of force,” but they believe these meanings are very strongly determined by the specific theories in which “force” is embedded.
16. The nature of the inconsistency escapes us. The gist of I&V’s argument is that internal forces resist motion, but acquired force happens only after some force creates motion. “If we think of an object that has been set in motion by an agent as having an acquired force, such an object cannot be thought of as having an internal force also, because if it did, the agent should not have been able to move it. . .” (p. 58). In other words, how could something resist motion, but be entailed by motion at the same time? But, the scientific concept of inertia acts both to “resist” initiation of motion and it also perpetuates motion that already exists, so there seems to be no strict conflict here. More generally, surely we must allow that there could be different kinds of forces that act a little differently. diSessa et al. (2002) document multiple meanings of force in one individual. If there is force to I&V’s claimed inconsistency, we believe it entails hidden assumptions about contextuality (or the nature of concepts), for example, that an entity cannot resist motion in one context and constitute or enhance it in another.
17. While moveable objects’ having less force sounds anomalous to adult ears, the children’s implicit rationale, according to I&V, is that motion implicates insufficient force to resist another influence or tendency.
18. We do not believe it is sensible to *assume* that gravity is a distinct meaning of force. From a scientific point of view, gravity has all the properties of any other force. One should at least allow this possibility and subdivide the issue into the nature of force, and *then* whether gravity is one.
19. The balls we used in the different situations were different. However, no subject commented on the difference in size and/or weight, and no subject explained any difference of force on this basis.
20. Anderson et al.’s results (1992), that explanations are more consistent than predictions, may seem relevant here. We do not believe it is. First, “explanations” here really refers to justifications, not to the kind of “explanations of what is happening” that Anderson elicited. Secondly, there is no prediction condition at all in our or I&V’s study. In fact, the perspective of our questions was not different from I&V’s; we asked for and received justifications, just as I&V did. It is only that we coded just the answers, not the justifications. I&V’s codes strictly entailed what we coded, but they entailed more as well (a justification), which we did not code.
21. In principle one could count the percentage of all codes that matched a particular model, rather than what we did, the slightly more aggressive counting by set, and counting a mismatch in a set if any code did not match. However, our accounting more closely matches I&V’s procedure since they integrated each set into a single code for each subject. In addition, accounting details are essentially irrelevant since I&V report a very high threshold: About 90% of subjects had *zero* mismatches. Zero I&V mismatches should translate strictly into a zero meaning deviation score.

22. 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.
23. Again, we ignored changes in the number of tied meanings.
24. To make calculations easy, we made the false assumption that if two subjects are in the same meaning category, then they match each other within the specified tolerance. This assumption is correct if one uses meaning deviation score 1 to classify subjects, then looks for individuals matching each other within 2. The minimum number of individual matches occurs when the subjects are spread most evenly across the meaning categories.
25. However, the contextuality of this ontology is far from an expert's. It appears limited to a few situations. Further, the compositional principle (cancellation of oppositely directed elements) is only a fragment of the expert's equivalent, vector addition.
26. The relevant physics distinction is not exactly that, in circular situations, things spin and in rectangular situations, things move. Instead, (still put fairly crudely) it is that in circular situations, analysis on the basis of spinning (torque) may be easier than analyses using linear analysis (force).
27. It appears to us that judgments of coherence (relational structure) of meanings is made implicitly by I&V. Model mapping, in our view, operationalizes consistency rather than relational coherence, but it operationalizes it over a contextual range that is only implicitly defined, by the set of questions asked and their perspective.
28. Note that, if we take the delineation of specification given in this paper seriously, and if we ignore difficulties in validating even the contextual range purported by I&V (difficulties seen in our quasi-replication, which did not introduce new contexts), then our work would roughly calibrate the "narrow" that I&V use to describe their meanings. Crudely speaking, "narrow" would describe about one fifth (existential) of the full range of aspects that must be consolidated in the development of an adequate Newtonian concept of force.

Acknowledgments

We are indebted to Stella Vosniadou and Christos Ioannides for putting their theories on the line with well-documented empirical data. We also gratefully acknowledge access to prepublication versions of their article and conversations that helped clarify what they did and how we should pursue our work. Preschool subjects were arranged with the help of the Berkeley Institute for Human Development. Marie-Eve Thomas (Hoover Elementary School) and Suzy Loper (Longfellow Middle School) also helped arrange subjects. A number of people contributed to formulating the empirical plan, including David Kaufman and Joe Wagner. Jim Greeno and two thoughtful, anonymous reviewers provided helpful feedback on an earlier version of the paper. Others who provided helpful feedback on drafts include Flavio Azevedo, Barbara White, Orit Parnafes, Nathaniel Brown, Michael Ranney, Sherry Seethaler, and Melinda diSessa. Funds from the UC Berkeley Chancellor's Professorship, provided to Andrea diSessa, supported students during this work.

Appendix A. Question sets 1–10 (quasi-replication)

See Table A1 .

Table A1
The 10 questions in our quasi-replication of I&V

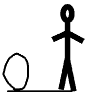
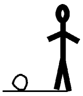
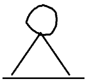
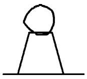
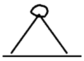
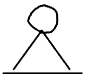
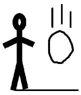
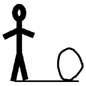
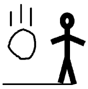
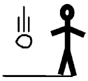
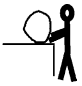
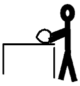
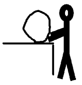
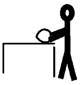
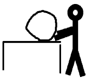
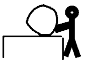
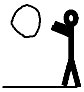
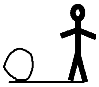
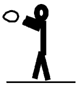
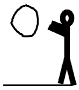
Set	Drawing A	Question A	Drawing B	Question B	Comparison question
1		“This stone is standing on the ground. Is there a force on this stone? Why?”		“This stone is standing on the ground. Is there a force on this stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
2		“This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? Why?”		“This stone is standing on a hill. It is stable. That means it won’t easily fall down. Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
3		“This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? Why?”		“This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
4		“This stone is falling. “Is there a force on the stone? Why?”		“This stone is standing on the ground. Is there a force on this stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
5		“This stone is falling. Is there a force on the stone? Why?”		“This stone is falling. “Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
6		“This man is trying to move this stone. Is there a force on the stone? Why?”		“This man is trying to move this stone. Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
7		“This man is trying to move this stone and it won’t move. Is there a force on the stone? Why?”		“This man is trying to move this stone and it won’t move. Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”
8		“This man is trying to move this stone and it won’t move. Is there a force on the stone? Why?”		“This child is trying to move this stone and it won’t move. Is there a force on the stone? Why?”	“Is the force on this stone the same or different than the force on this stone? Why”

Table A1 (*Continued*)

9		"This man has thrown this stone. Is there a force on the stone? Why?"		"This stone is standing on the ground. Is there a force on this stone? Why?"	"Is the force on this stone the same or different than the force on this stone? Why"
10		"This man has thrown this stone. Is there a force on the stone? Why?"		"This man has thrown this stone. Is there a force on the stone? Why?"	"Is the force on this stone the same or different than the force on this stone? Why"

Appendix B. Mappings of I&V's questions to ours

See Table B2.

Table B2
Mappings between I&V's questions and ours

I&V question number	Focus	Our question number
1	Large stone	1a
2	Small stone	1b
	(Compare above)	1c
3	*Large balloon*	
4	*Small balloon*	
5	Man pushing large stone	6a
6	Man pushing small stone	6b
	(Compare above)	6c
7	Man pushing large balloon	
8	*Man pushing small balloon*	
9	Man pushing large versus small stone, no movement	7a–c
10	Man versus child pushing large stone, no movement	8a–c
11	Unstable large stone on hill	3a
12	Unstable small stone on hill	3b
	(Compare above)	3c
13	Unstable large balloon on hill	
14	*Unstable small balloon on hill*	
15	*Unstable stone on hill versus stone on ground*	
16	*Unstable stone on hill versus unstable stone on lower hill*	
17	Unstable versus stable stones on hills	2a–c
18	Falling large stone	5a
19	Falling small stone	5b
	(Compare above)	5c
20	Falling large balloon	
21	*Falling small balloon*	
22	Falling stone versus stone on ground	4a–c
23	Thrown large stone	10a

Table B2 (*Continued*)

I&V question number	Focus	Our question number
24	Thrown small stone	10b
	(Compare above)	10c
25	*Thrown large balloon*	
26	*Thrown small balloon*	
27	Thrown stone versus stone on ground	9a–c

Notes. Our questions are divided into 10 sets that concern (a) the existence of a force on one object, (b) the existence of a force on another, and (c) the comparison of these forces, if they exist. Questions added are shown in bold face. Questions eliminated are shown in italics.

Appendix C. Unanalyzed questions in the extension study

Set 11: Balance scale: We asked about the "return to equilibrium" motion of a Lego-constructed balance scale. We expected to cue a forceless "return to equilibrium" p-prim, and thus get answers that deny Acquired Force even by subjects who showed Acquired Force meaning on I&V's questions.

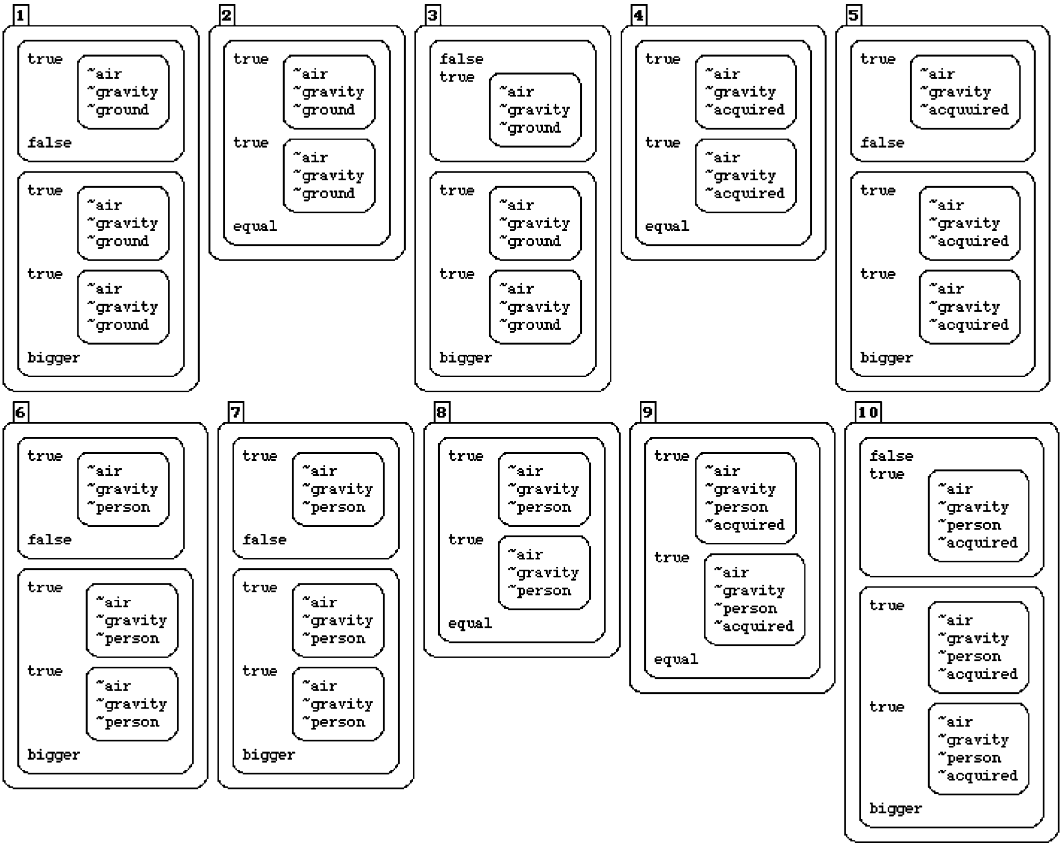


Fig. D3.1. Mapping for Internal meaning.

Set 12: Drinking straw: We asked subjects about the motion of liquid sucked up a drinking straw. Our prior work suggests the existence of an independent "sucking" primitive that is likely to behave differently than Acquired or Push/Pull meanings.

Set 24: Tug of war: We asked subjects what happened if the rope in a tug of war broke, and whether a force accounted for what did happen. We expected this to reveal further ontological and inferential aspects of force. In particular, we expected to see a "recoil" force that was generated in opposition to and out of release of pre-existing tension.

Appendix D. Sample mappings

How to read these mappings: Each question set is in a numbered box. Each subbox of a numbered box, stacked vertically, represents an acceptable alternative in terms of responses. A true or false indicates there is or is not a force (push/pull) on the relevant focal object (two focal objects per question set). In the case a force is specified on each focal object, a comparison code (bigger, equal, smaller) is given. Boxes that appear after a code are subcodes. Within subcodes, a ∼ indicates the following subcode is not allowable. For example, in case of the

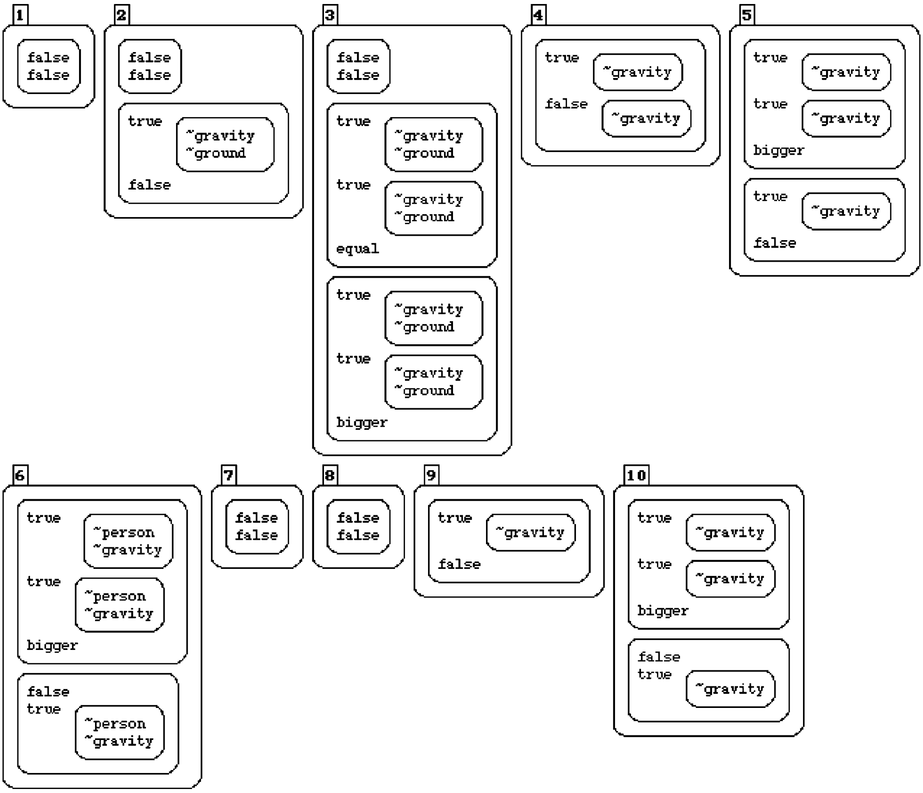


Fig. D3.2. Mapping for Acquired meaning.

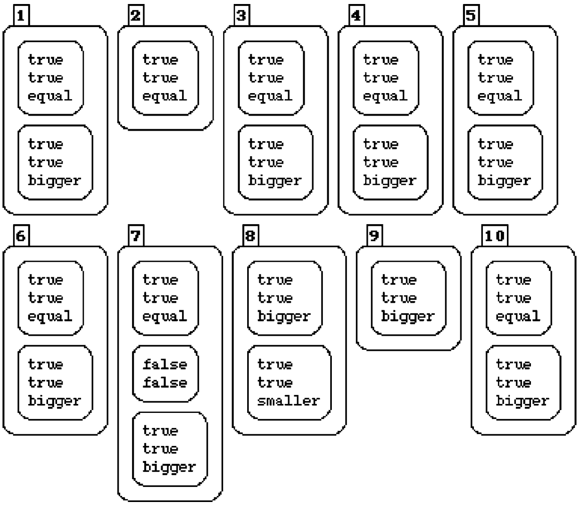


Fig. D3.3. Mapping for Gravity and Other meaning.

Internal meaning, the force on an object cannot involve gravity or any external objects, such as air or the ground. When motion is involved, the subject must not implicate that motion as the cause or evidence of force (∼acquired).

The full set of mappings is available from the authors (see Figs. D3.1–D3.3).

References

Anderson, T., Tolmie, A., Howe, C., Mayes, T., & Mackenzie, M. (1992). Mental models of motion. In Y. Rogers, A. Rutherford, & P. A. Bibby (Eds.), *Models in the mind: Theory, perspective, and application* (pp. 57–71). New York: Academic Press.

Brown, D., & Clement, J. (1989). Overcoming misconceptions via analogical reasoning. *Instructional Science*, *18*, 237–261.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (1999). Sources of conceptual change. In E. K. Scholnick, K. Nelson, & P. Miller (Eds.), *Conceptual development: Piaget's legacy* (pp. 293–326). Mahwah, NJ: Lawrence Erlbaum Associates.

Carey, S., & Smith, C. (1995). On understanding the nature of scientific knowledge. In D. N. Perkins (Ed.), *Software goes to school: Teaching for understanding with new technologies* (pp. 39–55). New York: Oxford University Press.

Chi, M. T. H. (1992). Conceptual change in and across ontological categories: Examples from learning and discovery in science. In R. Giere (Ed.), *Cognitive models of science* (pp. 129–160). Minneapolis, MN: University of Minnesota Press.

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

Clark, D. B. (2000). *Scaffolding knowledge integration through curricular depth*. Unpublished doctoral dissertation. Berkeley, CA: Graduate School of Education, University of California.

Clark, D. B. (2003). Analyzing student knowledge integration: Theories or pieces? In *Proceedings of the National Association of Research in Science Teaching Conference*. Philadelphia, PA.

Confrey, J. (1990). A review of student conceptions in mathematics, science, and programming. In C. Cazden (Ed.), *Review of research in education* (Vol. 16, pp. 3–56). Washington, DC: American Educational Research Association.

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

diSessa, A. A. (1980). Momentum flow as an alternative perspective in elementary mechanics. *American Journal of Physics*, *48*, 365–369.

diSessa, A. A. (1988). Knowledge in pieces. In G. Foreman & P. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Mahwah, NJ: Lawrence Erlbaum Associates.

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

diSessa, A. A. (1996). What do "just plain folk" know about physics? In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development: New models of learning, teaching, and schooling* (pp. 709–730). Oxford, UK: Blackwell Publishers, Ltd.

diSessa, A. A. (2000). Does the mind know the difference between the physical and social worlds? In L. Nucci, G. Saxe, & E. Turiel (Eds.), *Culture, development and knowledge* (pp. 141–166). Mahwah, NJ: Lawrence Erlbaum Associates.

diSessa, A. A. (2002). Why "conceptual ecology" is a good idea. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 29–60). Dortrecht: Kluwer.

diSessa, A. A., Elby, A., & Hammer, D. (2002). J's epistemological stance and strategies. In G. Sinatra & P. Pintrich (Eds.), *Intentional conceptual change* (pp. 238–290). Mahwah, NJ: Lawrence Erlbaum Associates.

diSessa, A. A., & Minstrell, J. (1998). Cultivating conceptual change with benchmark lessons. In J. G. Greeno & S. V. Goldman (Eds.), *Thinking practices in mathematics and science learning* (pp. 155–187). Mahwah, NJ: Lawrence Erlbaum Associates.

diSessa, A. A., & Sherin, B. L. (1998). What changes in conceptual change? *International Journal of Science Education*, *20*(10), 1155–1191.

Driver, R. (1989). Students' conceptions and the learning of science. *International Journal of Science Education*, *11*, 481–490.

Forbus, K. (1985). Qualitative process theory. *Artificial Intelligence, 24*, 85–169.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York, NY: Cambridge University Press.

Hunt, E., & Minstrell, J. (1994). A cognitive approach to the teaching of physics. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practice* (pp. 51–74). Cambrigde, MA: MIT Press.

Ioannides, C., & Vosniadou, C. (2002). The changing meanings of force. *Cognitive Science Quarterly*, *2*, 5–61.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge, UK: Cambridge University Press.

Kaiser, M. K., Proffitt, D. R., Wheelan, S. M., & Hecht, H. (1992). Influence of animation on dynamical judgments. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 669–689.

Keil, F. (1989). *Concepts, kinds, and cognitive development*. Cambidge, MA, USA: MIT Press.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist*, *44*(12), 1469–1481.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Mahwah, NJ: Lawrence Erlbaum Associates.

Minstrell, J. (1982). Explaining the "at rest" condition of an object. *Physics Teacher*, *20*(1), 10–14.

Minstrell, J., & Stimpson, V. (1996). A classroom environment for learning: Guiding students' reconstruction of understanding and reasoning. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 175–202). Mahwah, NJ: Lawrence Erlbaum Associates.

O'Malley, C., & Draper, S. (1992). Representation and interaction. In Y. Rogers, A. Rutherford, & P. A. Bibby (Eds.), *Models in the mind: Theory, perspective, and application* (pp. 73–91). New York: Academic Press.

Piaget, J., & Garcia, R. (1974). *Understanding causality*. New York: Norton.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, *66*(2), 211–227.

Ranney, M. (1987). *Changing naïve conceptions of motion*. Doctoral dissertation. Pittsburgh, PA: University of Pittsburgh, Learning Research and Development Center.

Ranney, M. (1988, November). *Contradictions and reorganizations among naïve conceptions of ballistics*. Paper presented at the annual meeting of the Psychonomic Society, Chicago.

Ranney, M. (1994). Relative consistency and subjects' "theories" in domains such as naïve physics: Common research difficulties illustrated by Cooke and Breedin. *Memory and Cognition*, *22*(4), 494–502.

Ranney, M. (1996). Individual-centered vs. model-centered approaches to consistency: A dimension for considering human rationality. *VIVEK: A Quarterly in Artificial Intelligence*, *9*(2), 35–43.

Samarapungavan, A., & Wiers, R. (1997). Children's thoughts on the origin of species: A study of explanatory coherence. *Cognitive Science*, *21*(2), 147–177.

Talmy, L. (1988). Force dynamics in language. *Cognitive Science*, *12*, 49–100.

Thagard, P. (1992). *Conceptual revolutions*. Princeton, NJ: Princeton University Press.

Thagard, P. (2000). *Coherence in thought and action*. Cambridge, MA: MIT Press.

Viennot, L. (1979). *La raisonnement spontané en dynamique elementaire* [Spontaneous reasoning in elementary dynamics]. Paris: Hermann.

Vosniadou, S. (2002). On the nature of naïve physics. In M. Limón & L. Mason (Eds.), *Reconsidering conceptual change: Issues in theory and practice* (pp. 61–76). Dordrecht: Kluwer Academic Publishers.

Vosniadou, S., & Brewer, W. F. (1992). Mental models of the earth: A study of conceptual change in childhood. *Cognitive Psychology*, *24*(4), 535–585.

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

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, *43*, 337–375.

Wiser, M. (1987). Novice and historical thermal theories. In S. Strauss (Ed.), *Ontogeny, phylogeny, and the history of science*. Norwood, NJ: Ablex.

Wiser, M. (1995). Use of history of science to understand and remedy students' misconceptions about heat and temperature. In D. Perkins, J. Schwartz, M. West, & M. Wiske (Eds.), *Software goes to school* (pp. 23–38). New York: Oxford University Press.

Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 267–297). Mahwah, NJ: Lawrence Erlbaum Associates.