

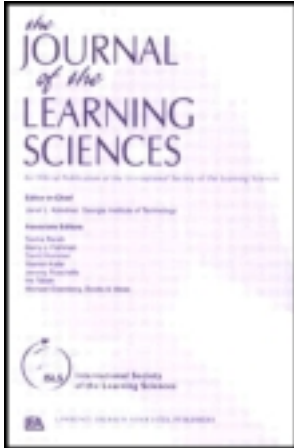
This article was downloaded by: [VUL Vanderbilt University]

On: 20 October 2011, At: 08:50

Publisher: Routledge

Informa Ltd Registered in England and Wales Registered Number: 1072954

Registered office: Mortimer House, 37-41 Mortimer Street, London W1T 3JH, UK



Journal of the Learning Sciences

Publication details, including instructions for authors and subscription information:

<http://www.tandfonline.com/loi/hlns20>

Comparison of Students' Knowledge Structure Coherence and Understanding of Force in the Philippines, Turkey, China, Mexico, and the United States

Douglas B. Clark^a, Cynthia M. D'Angelo^b & Sharon P. Schleigh^c

^a Department of Teaching and Learning, Vanderbilt University

^b Wisconsin Center for Education Research, University of Wisconsin

^c Department of Mathematics, Science & Instructional Technology Education, East Carolina University

Available online: 20 Apr 2011

To cite this article: Douglas B. Clark, Cynthia M. D'Angelo & Sharon P. Schleigh (2011): Comparison of Students' Knowledge Structure Coherence and Understanding of Force in the Philippines, Turkey, China, Mexico, and the United States, *Journal of the Learning Sciences*, 20:2, 207-261

To link to this article: <http://dx.doi.org/10.1080/10508406.2010.508028>

PLEASE SCROLL DOWN FOR ARTICLE

Full terms and conditions of use: <http://www.tandfonline.com/page/terms-and-conditions>

This article may be used for research, teaching, and private study purposes. Any substantial or systematic reproduction, redistribution, reselling, loan, sub-licensing, systematic supply, or distribution in any form to anyone is expressly forbidden.

The publisher does not give any warranty express or implied or make any representation that the contents will be complete or accurate or up to date. The accuracy of any instructions, formulae, and drug doses should be independently verified with primary sources. The publisher shall not be liable for any loss, actions, claims, proceedings, demand, or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with or arising out of the use of this material.

Comparison of Students' Knowledge Structure Coherence and Understanding of Force in the Philippines, Turkey, China, Mexico, and the United States

Douglas B. Clark

*Department of Teaching and Learning
Vanderbilt University*

Cynthia M. D'Angelo

*Wisconsin Center for Education Research
University of Wisconsin*

Sharon P. Schleigh

*Department of Mathematics, Science & Instructional Technology Education
East Carolina University*

This study investigates the ongoing debate in the conceptual change literature between unitary and elemental perspectives on students' knowledge structure coherence. More specifically, the current study explores two potential explanations for the conflicting results reported by Ioannides and Vosniadou (2002) and diSessa, Gillespie, and Esterly (2004) in terms of differences in coding schemes and differences in student populations. The current study addresses these questions by applying the coding schemes from both studies to interviews with 201 students drawn from the United States, the Philippines, Turkey, China, and Mexico. The analyses focus first on the coding schemes, suggesting that differences in coding schemes seem unlikely to account for the differences in the original studies. The analyses then focus on potential differences between student populations, suggesting that some differences exist in terms of consistency and meanings that might result from language, culture, or educational systems, but that these differences are too small to

account for the radical differences in the findings of the original studies. Two additional explanations are then proposed and explored involving the instruments and the epistemological stances invoked for the students. Overall, the results align more closely with the findings of diSessa, Gillespie, and Esterly (2004). [Supplemental materials are available for this article. Go to the publisher's online edition of *Journal of the Learning Sciences* for the following free supplement: Coding Schemes and Rules.]

Is a student's knowledge most accurately characterized as a coherent unified scheme of theory-like character (e.g., Carey, 1999, 2000; Gopnik & Schulz, 2004; Gopnik & Wellman, 1994; Ioannides & Vosniadou, 2002; McCloskey, 1983a, 1983b; Vosniadou, 2002; Vosniadou & Ioannides, 1998; Wellman & Gelman, 1992; Wiser & Carey, 1983)? Or is a student's knowledge more accurately characterized as an ecology of quasi-independent elements (e.g., Clark, 2000, 2006; diSessa, 1983, 1988, 1993; diSessa, Gillespie, & Esterly, 2004; diSessa & Sherin, 1998; Dufresne, Mestre, Thaden-Koch, Gerace, & Leonard, 2005; Hammer, Elby, Scherr, & Redish, 2005; Harrison, Grayson, & Treagust, 1999; Hunt & Minstrell, 1994; Linn, 2006; Linn, Eylon, & Davis, 2004; Linn & Hsi, 2000; Minstrell, 1982, 1989; Minstrell & Kraus, 2005; Özdemir & Clark, 2009; Parnafes, 2007; Thaden-Koch, Dufresne, & Mestre, 2006; Wagner, 2006)?

The preceding statements are simplifications of actual theoretical perspectives that are considerably more nuanced as a result of substantial research and ongoing debate among their respective proponents. Proponents of "theory-like" or "unitary" perspectives, for example, do not argue that students' knowledge is theory-like to the degree that scientists' knowledge is theory-like (e.g., including meta-conceptual awareness or availability to hypothesis testing). These proponents do argue, however, for an overarching hierarchical conceptual structure with theory-like properties that constrains a student's interpretation of subordinate models and ideas. Similarly, the "elemental" or "manifold" perspectives should not be incorrectly caricatured as the random interaction of independent elements. Rather, elements interact with one another in an emergent manner such that the combinatorial complexity of the system constrains students' interpretations of phenomenon.

The researchers in each camp also vary in terms of other important issues (e.g., conceptual grain size, ages of students, methods, and scientific content areas). Comparing findings among researchers in this debate has been difficult because of these differences in research methodologies and contexts. Recently, however, two groups of researchers have begun to address some of these issues around the concept of "force" in science. Ioannides and Vosniadou (2002; hereafter, I&V) published a study in *Cognitive Science Quarterly* about Greek students' understanding of force, suggesting that the students in their study demonstrated coherent understandings in terms of the consistent answers they expressed across multiple contexts. diSessa et al. (2004; hereafter, DG&E) published the results from their U.S. quasi-replication in *Cognitive Science*. Their findings suggested

that students' explanations lacked ontological coherence and varied significantly across contexts.

The current study applied the coding schemes from I&V (2002) and DG&E (2004) to 201 student interviews from the Philippines, Turkey, China, Mexico, and the United States (approximately 40 students in each country) in order to investigate students' understandings of force and the knowledge structure coherence of those understandings. Through this analysis, the current study contributes to the resolution of the controversy regarding the structure and coherence of students' science knowledge by clarifying the role of methodological approaches and student population differences in the findings of researchers on opposing sides of the controversy.

SCOPE AND IMPORTANCE OF THE DEBATE

This is a fundamental debate in the conceptual change literature. diSessa (2006) detailed fully the proponents and perspectives involved in this debate in the *Cambridge Handbook of the Learning Sciences* (Sawyer, 2006) in terms of the debate's roots in Piaget (Gruber & Voneche, 1977), Kuhn (1970), and Toulmin (1972) and the evolution of the debate through the early misconceptions research (e.g., McCloskey, 1983a, 1983b; Wisner & Carey, 1983) and conceptual change research (e.g., Posner, Strike, Hewson, & Gertzog, 1982) to the present.

The debate about knowledge structure coherence involves important theoretical and practical implications. Much research focuses on the conceptual processes through which students revise and build on their existing knowledge. Deeper understanding of the nature and structure of students' knowledge would contribute substantially to these efforts. Essentially, in order to understand the processes through which concepts change, it is important to understand the nature and structure of what is changing.

Similarly, understanding the nature and structure of students' knowledge would facilitate the design of curricula to better support students' learning processes as they build upon this knowledge. In addition, understanding how students from Mexico and other countries think about science topics like force and motion in comparison to U.S. English-monolingual students (who are more frequently studied) will contribute to the development of curricula that better support the underserved diverse student populations in classrooms around the world.

THEORETICAL COMMITMENTS AND SIGNIFICANT FINDINGS FROM I&V (2002)

Ioannides and Vosniadou's (2002; Vosniadou, 2002; Vosniadou & Ioannides, 1998) theoretical perspectives share several core theoretical commitments with

the other unitary perspectives described previously (e.g., Carey, 1999, 2000; Chi, 2005; Gopnik & Schulz, 2004; Gopnick & Wellman, 1994; Keil, 1994; McCloskey, 1983a, 1983b; Wellman & Gelman, 1992; Wiser & Carey, 1983).

I&V hypothesize that students' ontological and epistemological presuppositions and observations are organized into "framework theories." I&V define *framework theories* as causal explanatory frameworks for organizing physical phenomena that constrain the process of knowledge acquisition in ways analogous to the way in which paradigms have been thought to constrain the development of scientific theories. I&V do not claim that framework theories have the same status as scientific theories in terms of conscious awareness or availability to hypothesis testing, but they do claim that framework theories are coherent and consistently applied. Framework theories give rise to "specific theories." These specific theories consist of interrelated propositions or beliefs describing the properties or behaviors of physical objects that are generated through observation or other information provided by the culture.

Framework and specific theories provide the basis for generating situation-specific representations of mental models for problem solving. Even when constructed on the spot, the mental models are assumed to contain relatively consistent features because they are constrained by underlying framework and specific theories. I&V acknowledge that students may create "synthetic mental models" by fusing an existing mental model with new information to create a new interim model. I&V argue, however, that students will primarily apply the same mental model across contexts. Conceptual change should therefore generally involve a clear developmental progression from mental model to mental model.

I&V tested this perspective in Greece in terms of how students from four age groups conceptualized force. I&V showed students standardized sets of questions involving pictures with simple stick models and asked them about the forces on the objects in the pictures (see Figure 1). I&V also asked comparison questions to further explore students' interpretations of force (e.g., "big stone vs. small stone" or "big stone falling vs. big stone standing on the ground"). I&V proposed that consistency in individual students' explanations across contexts would constitute evidence for their theoretical perspective.



FIGURE 1 Sample question from Ioannides and Vosniadou's (2002) questionnaire: (a) standing big stone and (b) falling small stone.

I&V found that 88.6% of their participants' responses fell into one of seven internally consistent "meanings" of force. This included 86.7% of pre-kindergarten (pre-K) students, 80.0% of elementary school students, 86.7% of middle school students, and 100% of high school students. The seven meanings include (a) internal force, (b) internal force affected by movement, (c) internal and acquired force, (d) acquired force, (e) acquired force and force of push-pull, (f) force of push-pull, and (g) gravity and other forces. Students exhibiting an internal force meaning, for example, would explain that force is something innate to an object and/or is related to an object's size or weight. These students would make predictions and explanations consistent with this internal meaning across contexts. Full descriptions of each force meaning are provided in "I&V's Coding Scheme" in the Methods section. I&V interpreted the high levels of consistency observed in their Greek students' explanations as evidence of students' coherent knowledge structures.

SIGNIFICANT FINDINGS FROM DG&E (2004)

DG&E proposed that students maintain a more elemental knowledge structure (diSessa, 1983, 1988, 1993, 1996; diSessa et al., 2004) that shares a number of core theoretical commitments with other elemental perspectives (e.g., Clark, 2000, 2006; Dufresne et al., 2005; Hammer et al., 2005; Harrison et al., 1999; Hunt & Minstrell, 1994; Linn, 2006; Linn et al., 2004; Linn & Hsi, 2000; Minstrell, 1982, 1989; Minstrell & Kraus, 2005; Özdemir & Clark, 2009; Parnafes, 2007; Thaden-Koch et al., 2006; Wagner, 2006). These perspectives hypothesize that students' conceptual ecologies include a wide range of elements such as subconceptual *p*-prims,¹ beliefs, facts, facets,² and mental models, among others. These elements are cued by context and interact with one another in a network of positive and negative connections. These core mechanisms and interactions result in the potential for conflicts between ideas, sensitivity to contexts, differential weighting of ideas, and the systematicities created by the interaction of prominent elements.

Systematicities and local coherences in students' explanations arise when (a) contexts or questions cue the same subsets of elements, resulting in parallel interpretations by a student, particularly when the contexts are similar enough (e.g., Clark, 2006; diSessa, 1993); and/or (b) the students view their goals in an epistemological manner encouraging the pursuit of explanatory coherence (e.g., Ranney

¹*P*-prims (or phenomenological primitives) are unarticulated explanatory primitives in a student's conceptual ecology that provide the basis for many of the student's explanations about science phenomena (e.g., diSessa, 1993; diSessa & Sherin, 1998).

²*Facets* are independent explanatory facts or "rules of thumb" that students use to understand and explain situations and phenomena (e.g., Hunt & Minstrell, 1994; Minstrell & Kraus, 2005).

& Schank, 1998; Rosenberg, Hammer, & Phelan, 2006; Thagard, 1989, 2007; Thagard & Verbeurgt, 1998). This point is important because elemental perspectives are often caricatured as involving random interactions with no consistency. In fact, cuing the same sets of elements should result in consistent interpretation and explanation by the student, although these consistencies should typically involve less broad scopes than predicted by unitary perspectives. The question therefore becomes one of scope in terms of these causal systematicities. Learning occurs through a process of reorganization as students hopefully develop a more parsimonious and coherent understanding of normative theory-like character over time.

DG&E conducted a quasi-replication of I&V's (2002) study with U.S. students using a condensed version of I&V's study. DG&E found that their U.S. students' explanations of force did not demonstrate the same consistency as reported by I&V for the Greek students. More specifically, DG&E found that only 16.6% of their 30 students were fully consistent for one of the seven force meanings. DG&E then broadened their criterion for consistency with a 20% error allowance (which allowed a student to be categorized as consistent for a meaning if the student was coded for that meaning on at least 8 of the 10 question sets rather than 10 out of 10). When this softer criterion was used, 13 of the 30 students (43.3%) could be counted as consistent, but 9 of these 13 were consistent for the gravity and other meaning, which DG&E considered to be an ambiguous category. DG&E interpreted these results as evidence for an elemental knowledge structure in which students maintain ecologies of contextually cued knowledge pieces.

PURPOSE OF THE CURRENT STUDY

Subsequent discussions between Vosniadou and diSessa (e.g., Wagner, 2005) regarding the contradictory findings of the original studies focused on (a) cultural, semantic, or other differences between participant populations; and (b) differences in coding methods. In the Greek language, for example, the word for *force*, *dynamis*, also means "strength" or "power" in everyday speech. This might have contributed to higher levels of coherence in Greek students' explanations across interview contexts. Although semantic and cultural differences have been shown to impact students' thinking about specific science concepts (Aikenhead & Jegede, 1999; Costa, 1995; George, 1999; Inagaki & Hatano, 2002; Lubben, Netshisaulu, & Campbell, 1999), other differences in the national educational systems might also contribute to differences in outcomes. The explanation regarding methodological differences also warrants careful attention. Clearly, even slight differences in analytic methods can profoundly impact interpretations (Burkhardt & Schoenfeld, 2003; Nisbett & Ross, 1980; Stigler, Gallimore, & Hiebert, 2000; van de Vijver & Leung, 1997).

The current study examines these two explanations regarding the contradictory findings of I&V's and DG&E's studies. The first explanation explores the possibility that I&V's and DG&E's coding schemes code student responses differently and thus result in different findings. The second explanation focuses on potential differences between sample populations in terms of schooling, culture, and/or language across five countries. In answering these questions, the current study focuses on identifying the presence or absence of differences in knowledge structure coherence resulting from either differences in student populations or differences in coding methods. The current study thus provides a foundation for future studies to investigate the nature and underlying sources of the specific differences identified, contributes to the debate over knowledge structure coherence, and clarifies the conflicting findings between I&V's study and DG&E's quasi-replication.

METHODS

This study replicates DG&E's (2004) and I&V's (2002) work by applying the methodologies from those studies across interviews with students of the same four age groups from five countries (the United States, Turkey, the Philippines, China, and Mexico). The study builds upon the work and methods outlined in Özdemiş and Clark (2009).

Instrument

Students were asked the same 10 replication question sets that DG&E (2004) condensed from I&V's (2002) questions. Each question set includes two drawings comparing various combinations of sizes and positions of stones and people to explore the contexts in which the participants would assign forces and how they would describe those forces. Figures 2 and 3 present the 10 question sets.

DG&E reorganized I&V's questions such that each set consisted of three questions: two simple questions and one comparison question. The simple questions in each set asked about the existence and nature of forces in each picture (e.g., "Is there a force on this stone? Why?"). Comparison questions asked students to compare the forces in the two pictures (e.g., "Is the force on this stone in the first picture the same or different than the force on this stone in the second picture? Why?"). Comparison questions were asked when students indicated the existence of a force on both stones. The comparison question, when applicable, provided more information related to the student's understanding of force in terms of strength and contextual-related differences. Although DG&E omitted some questions, and rearranged the remaining questions into the 10 comparison



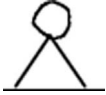
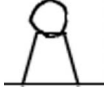
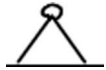





Question Set	Drawing A	Question A	Drawing B	Question B	Comparison Question
1		“This stone is standing on the ground. Is there a force on this stone? Why?”		“This stone is standing on the ground. Is there a force on this stone? Why?”	“Is the force on this stone the same or different than the force on 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?”

FIGURE 2 Question sets 1 through 5 from diSessa, Gillespie, and Esterly's (2004) study.

question sets, DG&E retained the same contexts and representations for the questions. DG&E did change the syntax in terms of alternating between using the term *force* and the phrase *push or pull* in their study but found that it did not result in differences and suggested that only the term *force* be used in subsequent studies. In the current study, we used only the term *force* except in rare occurrences for pre-K students who did not have any familiarity with the term *force* (almost exclusively in the Philippines). These exceptions are detailed later in terms of the specifics of each language in the Results and Discussion section.











<i>Question Set</i>	<i>Drawing A</i>	<i>Question A</i>	<i>Drawing B</i>	<i>Question B</i>	<i>Comparison Question</i>
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?"
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?"

FIGURE 3 Question sets 6 through 10 from diSessa, Gillespie, and Esterly's (2004) study.

Participants and Procedures

For the current study, 201 students from five countries were interviewed (37 students from the United States, 46 students from Turkey, 39 students from Mexico, 40 students from mainland China, and 39 students from the Philippines). As with I&V's and DG&E's studies, the current study involved students of four different age groups, with approximately 10 pre-K, 10 elementary school, 10 middle school, and 10 high school students per country. The mean student ages for these groups were 5, 10, 13, and 16 years, respectively. Students in each country were selected as being socioeconomically representative of middle-class students as defined for their country. No more than three students were selected from the

same school at any grade level. The Turkish cohort involved a new set of students rather than the original Turkish cohort from our group's first study (Özdemir & Clark, 2009) in order to standardize methods across the five countries. All students were interviewed individually for roughly 20 to 25 min. Students were asked all of the questions in one session. All interviews were videotaped. Interviews were translated into English by the interviewers prior to coding.

Selection of Countries and Interviewers

Countries for the study were chosen based on a combination of theoretical and practical considerations. The study was funded through a National Academy of Education/Spencer Foundation postdoctoral fellowship. To maximize the grant budget, we focused on countries of potential interest in which we had contacts with experience conducting science education research who were willing to conduct and translate interviews for reasonable stipends or exchanges of work. Four of the five interviewers were currently working on their doctoral degrees, and the fifth was an assistant professor of science education. In addition to science education research experience, interviewers were required to be natives of the focal countries so that the nuances of the language and culture would be well understood by the interviewers. Both of these criteria were considered essential in terms of the quality and fidelity of the interviews and the interviewers' ability to accurately translate and transcribe the interviews into English. Within these practical parameters, China and Turkey were chosen because of the significant differences in language and culture in comparison to the United States. The word for *force* in Turkey, furthermore, shares many colloquial similarities to the term for *force* in Greek. Although Spanish is not as distinct from English as is Chinese or Turkish, Mexico was chosen because of the important pragmatic value of understanding how students from Mexico, who make up a large proportion of immigrant students in U.S. classrooms, might think similarly or differently from U.S. English-monolingual students about the science concepts at the heart of this study. Furthermore, the word for *force* in Spanish also shares many colloquial similarities to the term for *force* in Greek. The Philippines was chosen as an interesting comparison point to the United States because instruction in the Philippines is largely conducted in English, but significant cultural differences between the Philippines and the United States exist. Greece would have been an ideal choice, given that I&V conducted their study in Greece, but budgetary constraints and events ultimately prevented data collection in Greece for this study.

Description of I&V's and DG&E's Schemes and Data Analysis

Students' responses were examined across question sets to check whether each student consistently applied the same meaning of force across the 10 question

sets. To address the possibility that the differences in the findings between I&V (2002) and DG&E (2004) resulted from differences in their coding schemes, the current study separately applied both coding schemes to each student's responses. We used largely the same approach applied in our group's initial study in Turkey (Özdemir & Clark, 2009) with a few refinements as described in the following sections.


I&V's coding scheme. I&V's coding scheme first coded students at the "question set level" and then at the "overall level." At the question set level, students' responses to each question set were scored as a group based on a scoring key containing a set of response categories for each set of questions. We transferred I&V's question set level coding rubrics to the revised organization of question sets used in DG&E's quasi-replication. This involved applying the same question set level scoring categories used by I&V in their rubrics. The question set level scoring categories for Question Set 1 are presented in Table 1. The Appendix presents the coded and annotated transcript of the interview with one student as an example of the application of I&V's as well as DG&E's schemes.

After scoring all of the questions at the question set level for the student's specific responses, we used the overall level rubric to assign the student's responses to potential matches from the seven force meaning categories (e.g., internal, push-pull, gravity and other). The overall level rubrics therefore involved a second scoring key outlining the pattern of expected responses for each force meaning. Again, we applied I&V's rubric categories and criteria to the question sets DG&E used in their quasi-replication. Table 2 provides an example of overall level scoring for Question Set 1.

The criteria used by I&V for assigning students to each of the meanings of force are as follows:

1. *Internal force.* Students were assigned to this meaning of force if they gave answers indicating that there is a force on or in all objects or only on big/heavy objects because they have weight or are big/heavy. Students do not refer to gravity, the object's motion, or another agent.
2. *Internal force affected by movement.* Students were assigned to this meaning of force if they gave answers indicating that force is due to the size/weight of an object but also that moving objects and objects that are likely to fall have less internal force than stationary objects.
3. *Internal and acquired force.* Students were assigned to this meaning of force if they indicated that there is a force on or in stationary objects due to size/weight and that these objects acquire an additional force when they are set in motion. I&V included students in this meaning who were ambivalent about unstable objects and who interpreted unstable objects as either lacking internal force or likely to acquire additional force.

TABLE 1
 Rubric for Assigning Categories of Responses Based on Ioannides and Vosniadou's (2002)
 Coding Scheme for Question Set 1

Set 1 	Big vs. Small Stones Standing on the Ground <ul style="list-style-type: none"> • This stone (big) is standing on the ground. Is there a force on this stone? Why? • This stone (small) 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?
<i>Response Category</i>	<i>Explanation</i>
A. Force only on the big stone	Because the big stone is big and/or heavy and/or you cannot move it. No force on the small stone because it is small and/or light and/or you can move it easily.
B. Force on both stones but greater force on the big stone	Because both stones are heavy or they have weight but the first stone is bigger and/or heavier and/or you cannot move it.
C. Force of gravity on both stones	Same force on both stones. It is the force of gravity, the Earth's attraction.
D. Alternative interpretation of the force of gravity ^a	Greater force of gravity/Earth's attraction on the big stone because it is heavier and its weight.
E. No force on any stones	Because they are not moving.
F. Force on the small stone, no force on big stone	Because the big stone is heavy and/or no one can move it easily. Because small stone is light and/or you can move it easily.
G. No force on any stones because no one pushes them	Because no one pushes them.
H. Force from the air on both stones	It is the force from the air above the stones. Same force on both stones because both stones are standing on the ground.

^a“Alternative” in this case is meant to distinguish Category D from Category C (“force of gravity on both stones [same]”). It is not meant to imply that if a student thinks there is more force on the larger stone that this is an alternative conception (i.e., a nonnormative or naïve conception) of gravity.

4. *Acquired force.* Students who indicated that force is a property of objects that explains motion and potentially acts on other objects were assigned to this meaning of force. These students answered that there is no force on stationary objects and that the force on a moving object disappears when the object stops moving. I&V also included students who thought that force is acquired only by heavy, moving objects and claimed that this response indicates that these students relate the acquired force to both the weight and the motion of the object. In addition, I&V included students in this meaning who thought that unstable stones have more force because they

TABLE 2
 Rubric for Assigning Force Meanings at the Overall Level Based on Ioannides and Vosniadou's (2002) Coding
 Scheme for Question Set 1

<i>Question Set</i>	<i>Internal</i>	<i>Internal/Move</i>	<i>Internal/Acquired</i>	<i>Acquired</i>	<i>Acquired/Push-Pull</i>	<i>Push-Pull</i>	<i>Gravity and Other</i>
Set 1: Big vs. small stones standing on the ground	A, B: Force only or greater on the big stone because bigger and/or heavier and/or you cannot move it	A, B: Force only or greater on the big stone because bigger and/or heavier and/or you cannot move it	A, B: Force only or greater on the big stone because bigger and/or heavier and/or you cannot move it	E: No force on any stones because they are not moving G: No force on any stones because no one pushes them G: No force on any stones because no one pushes them H: Force from the air	E: No force on any stones because they are not moving F: Force only on the small stone	G: No force on any stones because no one pushes them D: Greater force of gravity on the big stone	C: Force of gravity on both stones

Note. A–G correspond to the response categories in Table 1.

can be set in motion more easily as well as those who thought that all stones (stable and unstable) can be set in motion easily.

5. *Acquired force and force of push–pull.* Students were assigned to this meaning if they gave answers meeting the criteria described previously for the acquired meaning of force but also answered that there is a force on an object when it is acted on by an agent regardless of whether or not it moves.
6. *Force of push–pull.* Students were assigned to this meaning if they indicated that a force is exerted only on objects being pushed by an agent whether or not the object is moving.
7. *Gravity and other forces.* Students were assigned to this meaning if they mentioned gravity or gravity and other forces. According to I&V's coding, students could be considered consistent with this meaning for Question Sets 7 and 8 even if they did not mention the word *gravity* in these sets.

DG&E's coding scheme. DG&E were concerned that they could not reliably apply I&V's scheme to their interviews and thus adapted I&V's coding scheme. DG&E attempted to design their coding scheme to be more liberal in coding students as consistent for a meaning. DG&E's scheme is more "coarse quantitative" than I&V's in the sense that students' explanations were not coded for each question and integrated into an overall code. DG&E instead developed a "model mapping" technique that included all of I&V's meanings and specific codes. More specifically, DG&E compared students' responses to expected patterns for I&V's meanings at the coarse quantitative level by comparing combinations of the existence, absence, and relative sizes of forces on each object with potential exemptions based on the inclusion of specific sources of force expressed by the students. DG&E's scheme is exemplified for Question Set 1 in Table 3 (see also the Appendix).

We used DG&E's exact set of replication questions and could therefore use the exact coding scheme from DG&E. One point of clarification, however, involves the gravity and other category. DG&E expressed specific concerns about the lack of specificity of the category in their study. In DG&E's study, if a student mentioned gravity as a force on the object, the student was automatically precluded from being coded into any other meaning. Also in DG&E's study, a student could be coded as compatible with the gravity and other meaning based solely on the existence of forces on both stones without specifically mentioning gravity or attraction from the earth. Often students in our interviews outlined multiple independent force explanations within a question set (as was also seen in DG&E's study and in our initial study that precisely applied DG&E's scheme). We therefore modified our interpretation in the current study to code each independent force separately if there were multiple forces. Other separate meanings within a question set that were not specifically connected to gravity in the student's

TABLE 3
 Rubric for Assigning Force Meanings from diSessa, Gillespie, and Esterly (2004) for Question Set 1

<i>Question Set</i>	<i>Internal</i>	<i>Internal/Move</i>	<i>Internal/Acquired</i>	<i>Acquired</i>	<i>Acquired/Push-Pull</i>	<i>Gravity and Other</i>
Set 1: Big vs. small stones standing on the ground	Force only on the big stone, but not due to air, gravity, or ground	Force only on the big stone, but not due to air, gravity, or ground	Force only on the big stone, but not due to air, gravity, or ground	No force on any stone	No force on any stone	Equal force on both stones
	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	No force on any stone	No force on any stone	Force on both stones but greater force on the big stone
	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	Force on both stones but greater force on the big stone, but not due to air, gravity, or ground	No force on any stone	No force on any stone	Force on both stones but greater force on the big stone

explanations were counted for appropriate meanings. In addition, we did not assign students into the gravity category for DG&E's scheme unless the students explicitly referred to gravity or an attractive force between the earth and the object. Although our application of DG&E's coding scheme was generally symmetrical with its application in DG&E's original study, we did therefore adjust DG&E's coding scheme slightly to increase the specificity of the gravity and other category in alignment with the concerns expressed by DG&E and our initial study (Özdemir & Clark, 2009).

Online Repository of Materials and Invitation to Access the De-Identified Transcripts

We include I&V's and DG&E's rubrics for coding Question Set 1 in Tables 1 through 3, but space considerations preclude presenting the full coding rubrics for I&V and DG&E here. We do, however, make these available as supplemental materials in the publisher's online edition of *Journal of the Learning Sciences*. In addition to including the full set of rubrics for both schemes, we also include a document that outlines the rules we used to resolve specific coding challenges.

Our human subjects protocol does not allow us to post the entire corpus of transcribed interviews on the *Journal of the Learning Sciences* server, but we invite researchers interested in analyzing the full corpus of de-identified transcripts to contact us by e-mail. Our human subjects protocol stipulates that we can collaborate with researchers at other institutions to compare analyses of the data. We would welcome this opportunity and would work with other researchers to get them approved to work with the de-identified transcripts in order to advance discussion and consensus about the structure of students' understandings in science.

Coding of Individual Students in the Current Study

Two different coders coded each interview individually. The coding consisted of marking the data cells for each question for each possible force meaning using the methods developed for each coding scheme. This corresponded to a total number of 140 cells for each student (10 question sets multiplied by 7 possible force meanings per question set multiplied by 2 coding schemes). Any differences were discussed and resolved for each question set. After coding each interview, we tabulated across the question sets to determine how many times each student matched each force meaning category according to each scheme.

Interrater Reliability in the Current Study

The overall interrater reliability between the two coders before discussion was 93.1%, calculated by the percentage of matched cells in the coding schemes. Note

that this is a measure of reliability between the coders and not a comparison of agreement between coding schemes. We can also break down this number by country, by age group, and by coding scheme. By country, there were minor differences, with all countries having an average interrater reliability between 90.7% (the United States) and 95.7% (the Philippines). Among the age groups there was a similar spread, between 90.3% (elementary) and 95.0% (pre-K). There was a slightly higher interrater reliability using DG&E's scheme as opposed to I&V's scheme (94.3% vs. 92.1%, respectively).

Criterion Levels for Determining Consistency With and Without the 20% Error Allowance

Each student's consistency across the 10 question sets was checked for all seven force meanings. Additional possible consistent meanings were explored when they arose (e.g., force from being alive). Students were first checked for the meanings that they applied consistently across all 10 question sets. If a student matched for a force meaning across all 10 question sets, the student was coded as consistent for that force meaning. DG&E also used a looser criterion to code a student as consistent with an error allowance if the student used the same meaning on at least 8 of the 10 sets. In other words, a student could be classified as "consistent with allowance" if the student matched for at least 8 out of 10 question sets.

RESULTS AND DISCUSSION

We first explore the two most likely explanations for the radical differences between I&V's and DG&E's findings: (a) differences in coding schemes and (b) differences between student populations. We then propose and discuss possible limitations to this study and the original studies. Finally, we propose and explore two additional potential explanations related to the other two fundamental components of I&V's and DG&E's studies: (a) differences in the interview instruments and (b) differences in the impact of the interviewers themselves.

Explanation 1: Differences in Findings Resulting From Differences in Coding Schemes?

The first and simplest possible explanation for the differences in the findings of I&V's and DG&E's original studies would involve differences between their coding schemes. This was suggested by Vosniadou in a symposium at the annual meeting of the American Educational Research Association in 2005 (Wagner, 2005). DG&E were concerned about their ability to reliably apply I&V's coding scheme. They thus adapted I&V's coding scheme as discussed in the Methods

section. Although DG&E adapted the coding scheme in a manner that they felt would code for consistency more liberally than I&V's, it is possible that it did not.

To investigate this possibility, we compared the coding agreement between I&V's and DG&E's coding schemes at the levels of (a) cell-by-cell codes, (b) best-match meanings for individuals, (c) coding of individuals as either consistent or not consistent, and (d) overall percentages of individuals coded as consistent. The cell-by-cell codes represent the most atomized level of the coding. Best-match meanings aggregate the cell-by-cell codes. Coding individual students as consistent or not consistent aggregates the individual students one step further. Finally, comparing overall percentages of students coded as consistent by each scheme aggregates the overall impact of the two schemes at the highest level. Analyzing agreement between the coding schemes at these four levels of granularity clarifies the degree to which differences in the coding schemes might have accounted for the differences in findings of I&V's and DG&E's original studies from the most atomized level of the coding process to the most composite.

Agreement at the level of cell-by-cell coding. One approach for comparing agreement between the two coding schemes involved a "cell-by-cell" comparison of the coding charts for every student for every question for every possible force meaning (10 questions per student multiplied by 7 potential meanings per question for each coding scheme). Using this method with the students in our current study, we found that the two coding schemes agreed on 85.3% of the cells overall. The two schemes agreed for individual cells 84.5% of the time for U.S. students, 84.9% of the time for Turkish students, 85.6% of the time for Mexican students, 82.7% of the time for Chinese students, and 86.4% of the time for Filipino students. Overall, the age groups had similar levels of cell-by-cell agreement, ranging from 87.9% (pre-K students) to 82.1% (middle school students).

Agreement at the level of best-match meanings. Each student's "best-match meaning" is the force meaning matching the largest number of question sets for that student. In terms of best-match meaning agreement between I&V's and DG&E's schemes, if the best-match meaning assigned by DG&E's scheme for a student was the same as the best-match meaning assigned by I&V's scheme, then agreement was considered positive for that student. If two meanings tied for a scheme, then either could be used to determine agreement under this method. According to this criterion, the two coding schemes agreed on the best-match meanings for 89.1% of the students overall in the current study. The two schemes agreed in their characterization of 81.1% of the U.S. students, 91.3% of the Turkish students, 92.3% of the Mexican students, 92.5% of the Chinese students, and 87.2% of the Filipino students. Overall, the pre-K and high school students demonstrated the highest percentages of best-match agreement (approximately

93% each) and the elementary and middle school students demonstrated the lowest (approximately 85% each).

Agreement at the level of coding individual students as consistent or not consistent. Another approach for analyzing agreement involved comparing the categorical code for consistency assigned by each coding scheme to each student. The two coding schemes agreed for 86.6% of the students for the “fully consistent” criterion and 73.1% of the time for the “consistent with allowance” criterion. There was variation across countries and ages for the rate of agreement, but no statistically significant or obvious patterns manifested themselves.

Overall percentages of students coded as consistent by each scheme. Finally, we compared the overall percentages of students coded as fully consistent or consistent with allowance. In terms of the overall number of students categorized as fully consistent, the differences between the two coding schemes were less than 1% and not statistically significant. In terms of the consistent with allowance criterion, roughly 5% more students were coded as consistent overall with I&V's scheme than with DG&E's, but this difference was also not statistically significant. We present and discuss these percentages in greater detail in our examination of potential differences between student populations later on.

Sources of disagreement between the coding schemes. Overall, the two coding schemes demonstrate high levels of overall agreement. The sources of difference are worth considering, however, in terms of possible future work. The two coding schemes code student responses using somewhat different approaches. These differences manifest themselves in at least three ways. First, DG&E's scheme uses exemptions such as “force only on the big stone, but not due to air, gravity, or ground” to assign responses to force meanings. This can result in the exclusion of certain meanings, particularly if the student mentions gravity or is unsure about the forces in the question set. Second, the response categories and coding options are generally written in anticipation of a response focusing on a single force rather than multiple sources of force. This can lead to discrepancies between the two coding schemes, especially in combination with DG&E's system of exemptions. Third, the questions implicitly assume that force has an amount (and thus can be compared by amount). Coding students who do not think about force in this manner can therefore pose challenges. This difficulty manifests itself more frequently with DG&E's scheme because of the way the anticipated exemptions are worded in DG&E's scheme. These three sources of differences between the coding schemes account for most of the coding disagreements between the schemes. Although the two coding schemes code students very similarly overall, future work in this area would benefit from investigating these differences in greater detail.

Implications: Coding agreement between I&V's and DG&E's coding schemes. The two schemes code students (a) very similarly at the overall level and (b) fairly similarly at finer levels of granularity. These findings mirror the findings of our group's initial study in Turkey (Özdemir & Clark, 2009). We therefore conclude that differences between the coding schemes by themselves seem unlikely to account for the extreme differences in findings of the original studies.

Explanation 2: Differences in Findings Resulting From Differences Between Countries?

A second likely explanation for the differences between I&V's and DG&E's findings involves differences between the students in the two countries. Vosniadou and diSessa have discussed this possibility, particularly in terms of differences in languages or meanings of the word for *force* in each language (e.g., Wagner, 2005). In Greek, the word for *force*, *δύναμη* (*dynamis*), is commonly used colloquially even by young children and means "force," "might," "potency," "power," "strength," "vigor," and "virtue." When cross-referenced, some of these words have multiple entries in the Greek dictionary (in other words, *power* has more than one associated Greek word, each with a slightly different meaning), although *strength* does not. The word *force* in English, however, is not used colloquially as frequently by young children and colloquially has meanings related to police forces, armed forces, and "making someone do something," as discussed in more detail later on.

These differences should have less impact on the high school students in each country, who likely have been taught about the meaning of the term *force* in science classes, but these differences could potentially result in large differences for young children in terms of their understanding of the term, their interpretations of forces in the interview questions, and the consistency of their answers in their interviews. This perspective is supported by DG&E's discussion of how difficult it was for them to make sense of the whimsical and random nature of some of the pre-K students' answers (e.g., the "little lobsters"; see diSessa et al., 2004, p. 870). The current study explored the possible implications of differences between student groups by including students from five countries (i.e., the United States, Mexico, Turkey, China, and the Philippines) with a variety of colloquial meanings³ associated with the term for *force*.

³The descriptions of the colloquial meanings of the terms for *force* in each language were written by the interviewers involved in this study. As described in "Selection of Countries and Interviewers" in the Methods section, the interviewers were all native speakers of their focal languages. They conducted the interviews as well as the translations. All of the interviewers had completed some or all of their doctoral training in science education at the time of the interviews. Their descriptions were cross-checked by one or more additional native speakers of each language for verification.

United States. The word *force* can have many meanings in English. It can be used in its normative sense in a science context, but everyday language assigns it many other uses as well. It can be used in the sense of someone forcing someone to do something, a force of nature, a police force, a forced entry, or even “May the Force be with you.”

Turkey. The Turkish word for *force*, *kuvvet*, implies “power,” “strength,” “constrain,” and “firm” in addition to its normative meaning in physics. The colloquial meanings for *force* in Greece and Turkey therefore exhibit many parallels and similarities. The analysis of these interviews confirms that almost all students held nonnormative ideas and beliefs about force that were mostly related to other meanings and daily uses of the term *kuvvet*.

Mexico. The Spanish word *fuerza* is used for *force*. It has very similar meanings to the English word *force*, including “strength,” “influence,” “power,” to compel someone to do something, or a police force. The word for *strength* is also *fuerza*, comparable to Greek and Turkish.

China. Participants in China were interviewed using Mandarin. In Mandarin, the single character 力 (*li*) corresponds to the meaning of *force* in physics. In everyday life, this single character is seldom used. It is often combined with another character to further define the meaning. For instance, 权力 means “power;” and 力量 or 力气 mean “strength.” Interviews with middle school and high school students were conducted using the word *force* because they have learned physics in schools. For younger students who were not familiar with force, the interviewer had to use 用力推 (meaning “using force to pull”) or 用力拉 (meaning “using force to push”). Interviewees at this level tended to explain that there was a force pushing or pulling the object only when a person was actively pushing or pulling it. They typically explained that there was no force because the person did not touch the stone.

The Philippines. Although the majority of the interviews from the Philippines were conducted in English, most of the interviews with pre-K students needed to be conducted in Tagalog, the common language spoken in the Philippines. Some of the students interviewed were not familiar with the word *force* or its translation in Tagalog, which is *puwersa*. In these interviews the interviewer had to use *humihila* (“pulling”) and *tumutulak* (“pushing”) instead of using the word *force* or *puwersa*. According to the interviewer, all of the interviewees in this age group tended to say that something was pushing or pulling on the object in the situations in which they actually saw a person doing the pushing or pulling (similar to the young Chinese students). According to the interviewer, when asked whether they thought something was pushing or pulling on the stone

on the ground, the students would typically say that nothing was, because the person was not touching the stone. In other words, for these students, pushing or pulling came about when someone was in direct contact with the stones. This was not necessarily the case, however, for the students in the Philippines in the higher grade levels, who interviewed in English. To be clear, all of the students of all ages in all other countries were interviewed in their native language.

Difference Between Countries: Synthesis. One might expect to see different patterns of meanings expressed in the interviews as a result of the varying colloquial meanings in each language. One might also expect to see higher levels of consistency for young students in countries in which the terms for *force* include common colloquial meanings for young students than in countries where the term for *force* does not have such common or consistent colloquial meanings. Students in these latter countries might be less consistent in their explanations because they would not have a clear sense of the term for *force*. These differences in levels of consistency would be expected to diminish for older students, however, who presumably would become more familiar with the meaning of the term for *force* in the context of science.

Other cultural or educational differences among countries beyond language could also result in differences in consistency. The current study focuses on establishing the degree of differences observed among countries in the context of the current debate. If the current study were to document substantial differences, future work could then explore the nature of these differences in greater detail.

Best-Match Meanings Across Ages and Countries

Tables 4 through 8 combine all students from each country into a grid that is similar in format to that used by DG&E to display their coding results. The columns represent the seven force meanings. Each student's best-match meaning is the force meaning matching the largest number of question sets for that student. The rows show how many question sets a student matched for a best-match meaning (between 2 and 10 question sets). Whereas DG&E needed only one table because they focused on one country with one coding scheme, we include separate tables for each country and each coding scheme. In our tables, the letters represent individual students of each grade. Lowercase letters represent a student who had two or more meanings that "tied" for best match for that scheme. The first row in Table 4, for example, contains two lowercase *ks* because a U.S. pre-K student's best match was a tie between internal/movement and internal/acquired (both of which matched for two question sets).

These tables are useful because they help visually identify overall trends by country and coding scheme. Looking at the patterns in the grids, one can see many resemblances between coding schemes for each country, which lends support to the idea from the previous section that I&V's and DG&E's schemes

TABLE 4
Compiled U.S. Best-Match Meanings for Each Student

<i>Number of Question Sets Matched</i>	<i>Internal</i>	<i>Internal/Move</i>	<i>Internal/Acquired</i>	<i>Acquired</i>	<i>Acquired/Push-Pull</i>	<i>Push-Pull</i>	<i>Gravity and Other</i>
Ioannides and Vosniadou's (2002) coding scheme							
2		k	k				
3					e	e	
4		k m	k m	m			
5	m				K e m m	K e	e m m
6		K E		E	E E		
7		K E	E M		H	K	H
8				M	E M		M M
9			M				M H H
10					m M		E m H H H
							H
diSessa, Gillespie, and Esterly's (2004) coding scheme							
2				K			
3					K		
4							
5		K k e			K E e M	k e	
6				E m m	m m		M
7	K	K E	M m	h	E E M M m h	K	M H
8				e	e M		H
9				E			E H H
10							M M H H H

Note. Letters represent students, columns represent best-match meanings, and rows represent the number of question sets the student matched for that meaning. Lowercase letters represent a student who had two or more meanings that "tied" for best match for that scheme. Boldface indicates students who matched a meaning for eight or more question sets, thus qualifying as consistent or consistent with allowance for that meaning. K = pre-K; E = elementary school; M = middle school; H = high school.

code students similarly. However, there are also differences in force meanings and consistency levels among countries. The following sections investigate these potential differences among countries in greater depth.

How Consistent Are Students in Their Meanings?

We first examine levels of consistency in terms of the fully consistent and consistent with allowance criteria for (a) overall levels of consistency, (b) consistency by age groups, and (c) consistency by country. We then examine consistency using the best-match scores underlying the fully consistent and consistent with allowance codes.

Overall levels of consistency. At the most basic level, we see that the overall percentage of consistent students in the current study was about 12% in terms

TABLE 5
Compiled Turkey Best-Match Meanings for Each Student

<i>Number of Question Sets Matched</i>	<i>Internal</i>	<i>Internal/ Move</i>	<i>Internal/ Acquired</i>	<i>Acquired</i>	<i>Acquired/ Push–Pull</i>	<i>Push– Pull</i>	<i>Gravity and Other</i>
Ioannides and Vosniadou's (2002) coding scheme							
2							
3							
4			K		K		
5		h	h	H	K	E	h
6		K		m	m m h	m m h	
7	k E H		K k E	E	E M	K	H
8			E	M H	H H H H H H		M M
9		H		E E	E H H		
10					M H H H		H H H H H
diSessa, Gillespie, and Esterly's (2004) coding scheme							
2							
3		k	k				
4							
5		h	h	H	h	E	
6				K	K M H H H	M	
7	k k h	h	k k e e h	E M M H	K e e H h	K	H h
8	E H h				M H H h		H H
9				E E	E H H		H H
10				E H	H		M M H

Note. Letters represent students, columns represent best-match meanings, and rows represent the number of question sets the student matched for that meaning. Lowercase letters represent a student who had two or more meanings that “tied” for best match for that scheme. Boldface indicates students who matched a meaning for eight or more question sets, thus qualifying as consistent or consistent with allowance for that meaning. K = pre-K; E = elementary school; M = middle school; H = high school.

of the fully consistent criterion and approximately 53% in terms of the consistent with allowance criterion (see Figure 4). As discussed in terms of coding schemes differences, the differences between the two coding schemes for the fully consistent criterion were less than 1% and were not statistically significant. Students were coded 5.4% more frequently for the consistent with allowance criterion using I&V's scheme than DG&E's, but this difference was also not statistically significant.

At this point, it should be noted that the overall frequencies of fully consistent and consistent with allowance students in the current study for both coding schemes are more similar to the frequencies reported by DG&E than those reported by I&V (see Figure 4). Even the percentage of students who are consistent with allowance in the current study does not match I&V's reported percentage of fully consistent students.

TABLE 6
Compiled Mexico Best-Match Meanings for Each Student

Number of Question Sets Matched	Internal	Internal/ Move	Internal/ Acquired	Acquired	Acquired/ Push-Pull	Push- Pull	Gravity and Other
Ioannides and Vosniadou's (2002) coding scheme							
2							
3				e		e	
4				k		K k	
5		m M	m				E
6		E	M		m		m
7	k		k E m	K E E M	M H h H	E	M h
8	K E			M	M M M H H		E E M H
9	K K						H
10	K K						H
diSessa, Gillespie, and Esterly's (2004) coding scheme							
2							
3		m	m				
4					H		
5	e	e	e m	k e e m m	e e m m	k	
				m			
6					E	E	E
7	K k		k	K E m	m H	E	H
8					H	K	E M M H H
9	K K E			M	M m H		
10	K K			M	M H		

Note. Letters represent students, columns represent best-match meanings, and rows represent the number of question sets the student matched for that meaning. Lowercase letters represent a student who had two or more meanings that "tied" for best match for that scheme. Boldface indicates students who matched a meaning for eight or more question sets, thus qualifying as consistent or consistent with allowance for that meaning. K = pre-K; E = elementary school; M = middle school; H = high school.

Consistency by age group. When comparing frequencies of consistent students across countries, we see that the overall patterns were the same according to each coding scheme (see Figures 5 and 6). Students in older age groups were more frequently consistent. These differences among age groups were statistically significant for the DG&E coding scheme for the consistent with allowance criterion, $\chi^2(3, N = 201) = 18.38, p < .01$; and for the fully consistent criterion, $\chi^2(3, N = 201) = 11.60, p < .01$. These differences were also statistically significant for I&V's scheme for the consistent with allowance criterion, $\chi^2(3, N = 201) = 24.67, p < .01$; and for the fully consistent criterion, $\chi^2(3, N = 201) = 17.85, p < .01$.

Consistency by country. There appeared to be differences among countries in terms of consistent students (see Figures 7 and 8). These differences were

TABLE 7
Compiled China Best-Match Meanings for Each Student

<i>Number of Question Sets Matched</i>	<i>Internal</i>	<i>Internal/ Move</i>	<i>Internal/ Acquired</i>	<i>Acquired</i>	<i>Acquired/ Push–Pull</i>	<i>Push– Pull</i>	<i>Gravity and Other</i>
Ioannides and Vosniadou's (2002) coding scheme							
2							
3							
4						E	
5		k	k				
6		k E	k k	K	k	K	E M
7	K m				K e	e	M m H H
8		m	m		E	K e	M M M M M m H H H H H H
9		K H			E E e	K E	M H
10					K		
diSessa, Gillespie, and Esterly's (2004) coding scheme							
2							
3						E	
4							
5	e	e	e	k e	k e	k e	
6	h	K	K	K			E h
7			h		E E h	K	h
8	K			E	K E m	K E E	M m H H H H
9	M m	K			K M M m m m m		M m m m m H H
10	H				E		H

Note. Letters represent students, columns represent best-match meanings, and rows represent the number of question sets the student matched for that meaning. Lowercase letters represent a student who had two or more meanings that "tied" for best match for that scheme. Boldface indicates students who matched a meaning for eight or more question sets, thus qualifying as consistent or consistent with allowance for that meaning. K = pre-K; E = elementary school; M = middle school; H = high school.

statistically significant for the DG&E scheme for the consistent with allowance criterion, $\chi^2(4, N = 201) = 10.07, p = .04$, but not for the fully consistent criterion. This suggests the possibility of some differences between student populations in terms of levels of consistency, potentially resulting from educational systems, languages, or cultures among countries.

Main and Simple Main Effects Analysis for Age and Country

Thus far, the analyses have focused on the fully consistent and consistent with allowance criteria used by the original studies. The categorical nature of the

TABLE 8
Compiled Philippines Best-Match Meanings for Each Student

<i>Number of Question Sets Matched</i>	<i>Internal/ Internal</i>	<i>Internal/ Move</i>	<i>Internal/ Acquired</i>	<i>Acquired/ Acquired</i>	<i>Acquired/ Push-Pull</i>	<i>Push-Pull</i>	<i>Gravity and Other</i>
Ioannides and Vosniadou's (2002) coding scheme							
2							
3				k	k E		
4							
5				K			E
6			m	k	k m		
7		M	H	K E E		K	M
8			E	K E E E	M M M M M		
9					H		
10			K M	K E	K H	E	M H H H
						K	H H H
diSessa, Gillespie, and Esterly's (2004) coding scheme							
2							
3							
4				K			
5				K			
6				k k E e m	k k e m	k	E M
7		M		e	E E e H	K E	H
8			M H	K K E m	M M m H		
9			K	E	E M M M H		
10					H	K	H H H H

Note. Letters represent students, columns represent best-match meanings, and rows represent the number of question sets the student matched for that meaning. Lowercase letters represent a student who had two or more meanings that "tied" for best match for that scheme. Boldface indicates students who matched a meaning for eight or more question sets, thus qualifying as consistent or consistent with allowance for that meaning. K = pre-K; E = elementary school; M = middle school; H = high school.

fully consistent and consistent with allowance criteria, however, limits analysis to nonparametric tools. Focusing instead on the best-match scores that underlie the fully consistent and consistent with allowance designations allows a broader range of analytic tools for examining the data in their finer grained semi-continuous format. Best-match score is the number of question sets that a student matched for his or her best-match meaning. If a student was 100% consistent for a force meaning across all question sets, that student matched for the same force meaning on all 10 question sets, and that student's best-match score would therefore be 10 out of 10.

We performed a 4 × 5 analysis of variance to look at the effects of age and country on best-match score. For the data using I&V's scheme, there was a significant main effect for age, $F(3, 181) = 12.34, p < .01$; but not for country, $F(4, 181) = 1.75, p = .14$. For DG&E's scheme, there was a significant interaction

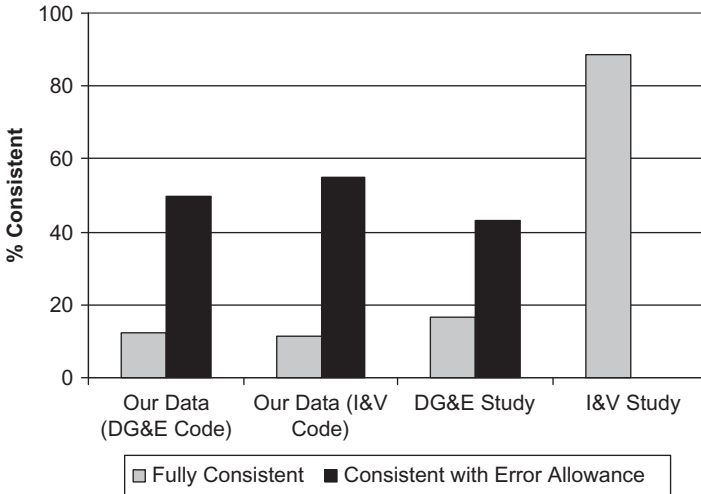


FIGURE 4 Overall levels of consistency for the current study using Ioannides and Vosniadou's (I&V) and diSessa, Gillespie, and Esterly's (DG&E) schemes compared to the levels of consistency reported in I&V's (2002) and DG&E's (2004) original studies.

effect between age and country, $F(12, 181) = 2.43, p < .01$. Further analysis was required to make sense of the best-match data.

Looking more closely at the results using I&V's coding scheme, we see that although there was a significant main effect for age, not all of the pairwise comparisons were significant. The significant pairwise comparisons were pre-K with middle school, pre-K with high school, elementary with high school, and middle school with high school. The high school scores were the highest ($M = 8.43$), and the pre-K scores were the lowest ($M = 6.73$). Although there was not a significant main effect for country, there was a significant pairwise difference between the United States ($M = 7.08$) and the Philippines ($M = 7.85$).

For the results using DG&E's coding scheme we needed to look at the simple main effects (because of the Age \times Country interaction). Looking at age within country, we see that there were significant differences within the United States, $F(3, 181) = 7.35, p < .01$; China, $F(3, 181) = 3.90, p = .01$; and the Philippines, $F(3, 181) = 3.36, p = .02$. As can be seen in Figure 9, the United States had the most differences between age groups, most of which were significant. For the country within age simple main effect, only within pre-K was there a significant effect, $F(4, 181) = 4.17, p < .01$. This tells us that the majority of the differences in consistency among countries occurred at the pre-K level, ranging from 5.13 (the United States) to 8.00 (Mexico). Although there were definite pairwise differences among many countries and age groups, overall the largest

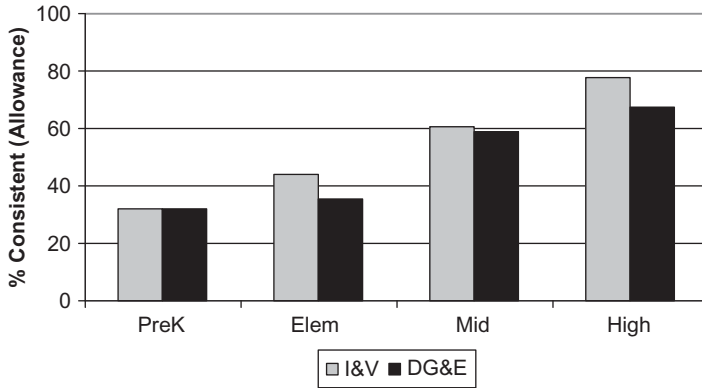


FIGURE 5 Percentage of consistent with allowance students in the current study by age group. PreK = pre-kindergarten; Elem = elementary school; Mid = middle school; High = high school; I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004).

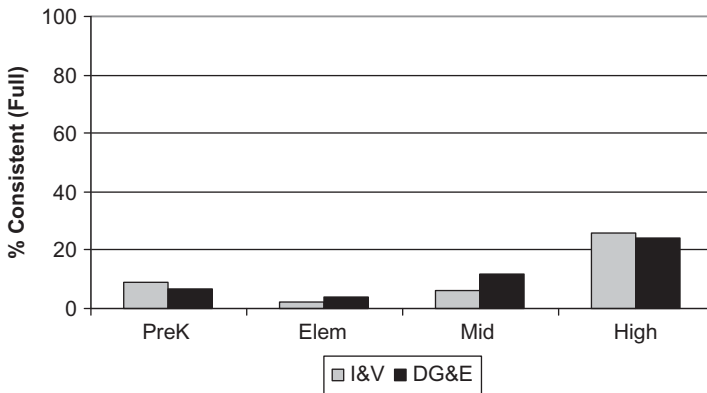


FIGURE 6 Percentage of fully consistent students in the current study by age group. PreK = pre-kindergarten; Elem = elementary school; Mid = middle school; High = high school; I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004).

differences across countries involved the youngest students. This aligns with the idea that differences due to language might manifest themselves most strongly for younger students and diminish for older students. Although these differences are certainly of interest for future studies, the scale of these differences in consistency is not substantial enough to account for the radical differences between I&V's and DG&E's findings.

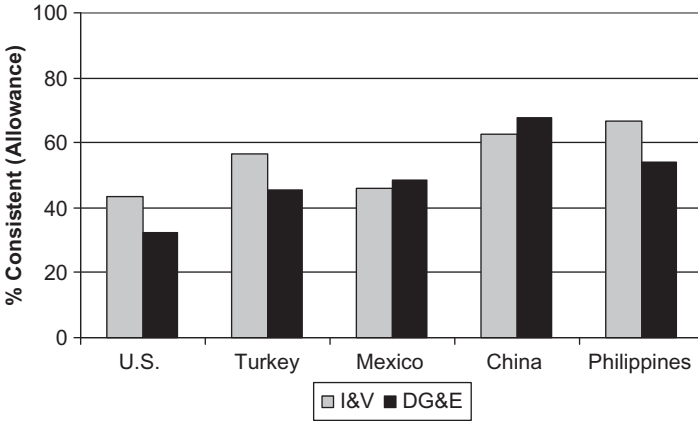


FIGURE 7 Percentage of consistent with allowance students in the current study by country. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004).

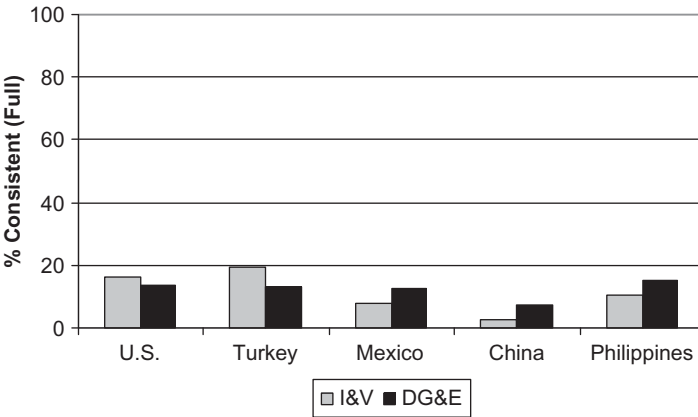


FIGURE 8 Percentage of fully consistent students in the current study by country. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004).

Possible Differences in Progressions of Meanings by Country

Although the debate over differences between I&V's and DG&E's findings focuses on differences in consistency, it is also worth exploring possible differences in progressions of meanings. Overall, the results of the current study indicate the same general age progression of meanings across countries as reported by I&V and DG&E. The pre-K students were spread across all of the

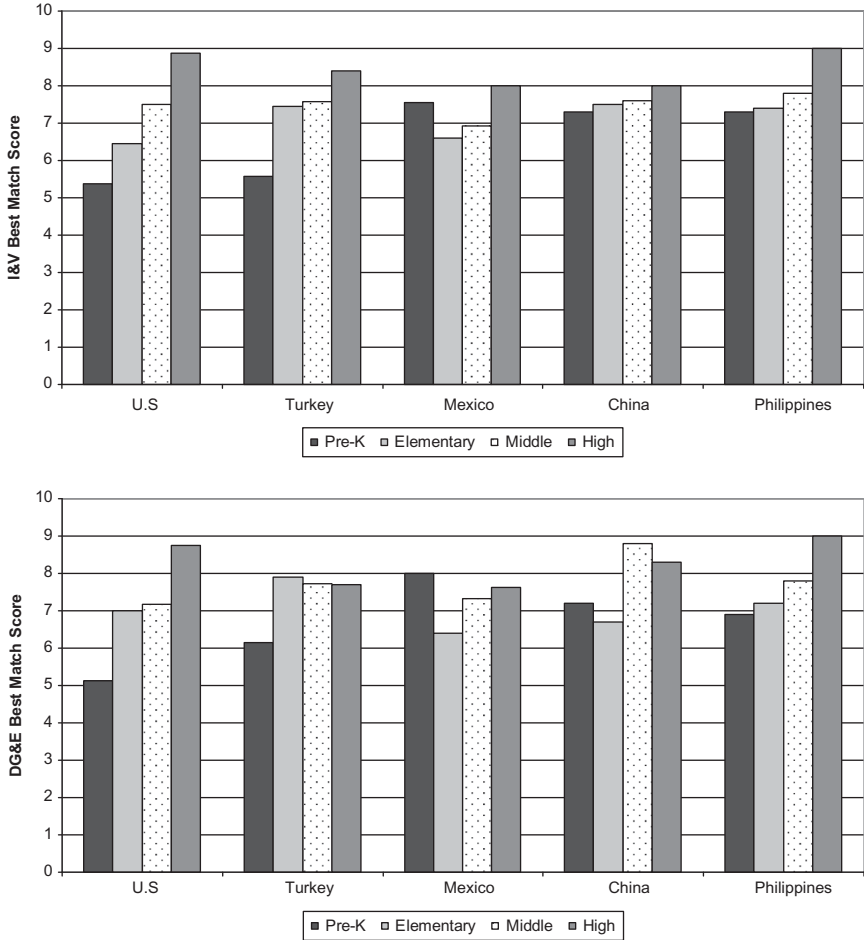


FIGURE 9 Best-match scores for each coding scheme by country and age group. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); PreK = pre-kindergarten; Elementary = elementary school; Middle = middle school; High = high school.

best-match meaning categories except gravity and other. The elementary students primarily expressed acquired-related meanings, with a substantial scattering of other best-match meanings. The middle school students primarily expressed acquired/push-pull and gravity and other meanings, with few middle school students expressing other best-match meanings. The high school students primarily expressed gravity and other or acquired/push-pull meanings, with minimal expression of other best-match meanings.

Comparing the best-match meanings expressed by students across countries suggests significant similarities across countries, but there are some differences. Students in Turkey and Mexico had a similar distribution of meanings, with many students in acquired-related meanings (although they differed in the internal meaning). Students in China and the United States exhibited gravity and other as their best-match meanings more frequently than students in the other countries and also had a high number of students in the acquired/push-pull meaning. The push-pull force meaning was the best-match meaning for about the same number of students in each country (about 10%). The acquired/push-pull force meaning had the largest percentage of students in each country (around 30%), except for China, where gravity and other was the best-match meaning for the largest percentage of students. The acquired and acquired/push-pull meanings were especially pronounced for the Philippines. The patterns of meanings described here based on I&V's scheme are very similar for DG&E's scheme and we therefore do not present them for DG&E's scheme in the interest of space. Overall, the noted differences among countries manifest themselves across age groups, which suggests differences that might include language but that likely also include other variables, such as educational systems, at least for the older students.

Summary: Can Differences Across Countries Account for the Significant Differences in the Findings of DG&E and I&V?

I&V found that 88.6% of the 105 students in their original study were fully consistent for a single meaning across all question sets. DG&E found in their original study that only 16.6% of their 30 students were fully consistent and only 43.3% could be counted as consistent with allowance (with most of these falling in the gravity and other meaning). Overall, the results of the current study align much more closely with DG&E's findings than I&V's findings, but we do see some differences between the United States and the other countries. For example, although China and the Philippines somewhat mirror the United States in terms of increasing consistency from the youngest students to the oldest students, Turkey and Mexico show relatively constant levels of consistency across age groups. When we shift the analysis to focus on the finer grained best-match scores underlying the consistency criteria, we see significant interactions of age and country. These differences focus on the youngest students and could be explained by differences in language. In Spanish, as in Greek, the word for *force* is exactly the same as the word for *strength*, and the word for *force* is colloquially familiar to young children. Similarly, in Turkish, the word for *force* is colloquially familiar to young children. This familiarity could support the increased consistency for the youngest students in these two countries and in Greece. That said, the levels of consistency for students of all ages in all countries align much more closely with DG&E's findings than with I&V's.

In summary, (a) there are some differences noted among countries that might result from language, culture, or educational systems; (b) the overall results from the five countries in the current study suggest that differences among countries do not seem likely to explain the substantial differences in findings between I&V's and DG&E's studies; and (c) the results of the current study for student consistency seem much more closely aligned with the overall findings of DG&E's study than I&V's study.

LIMITATIONS OF THE CURRENT STUDY AND THE ORIGINAL STUDIES

Before we explore two other potential explanations for the radical differences in the findings of DG&E and I&V, five issues require further discussion. These are (a) the absence of Greek students from the current study, (b) the implementation of the coding schemes in the current study, (c) the framing of the questions developed by I&V and adopted by DG&E and the current study, (d) the numbers of students and likely variation within a country, and (e) problematic issues with the gravity and other coding category.

Absence of Greece

Including Greece in the current study would have been desirable, but this study was conducted as part of a National Academy of Education/Spencer postdoctoral fellowship and did not include a substantial personnel budget. The current study thus focused on partnerships and relationships that we already had in place, which precluded including Greece as a site. We argue, however, that these five countries represent a large range of cultures, educational systems, languages, and colloquial meanings of the terms for *force*. The terms for *force* in Mexico and Turkey share many colloquial similarities to the term for *force* in Greece. Furthermore, although some differences in terms of consistency and meanings are suggested across the countries, these differences are of such small magnitude in comparison to the substantial differences in findings of DG&E and I&V that it seems unlikely that potential differences between Greece and the United States could account for the substantial differences. The students in Greece would need to be radically, qualitatively, and anomalously different from the students in the five countries of the current study to account for the substantial differences in findings. Therefore, although it does remain a possibility that the students in Greece are indeed radically different from the students in the five countries of the current study, the results of the current study suggest that such differences seem unlikely as an explanation.

Implementation of Coding Schemes

The current study adopted and applied the two coding schemes from DG&E's and I&V's original studies. We outlined our interpretations of the schemes in the Methods section and detailed them more fully in Özdemir and Clark (2009). Some interpretation in the implementation of the schemes was unavoidable. We have tried to make clear the ways in which we think our implementations may have diverged from those of I&V and DG&E. In summary, (a) we used the replication questions that DG&E condensed from I&V's original questions; (b) we applied I&V's coding scheme to DG&E's replication questions by copying the categories, definitions, and codes from I&V's coding schemes for the analogous question sets in their original study into the reorganized question sets; and (c) we directly applied DG&E's coding scheme as implemented by DG&E with the exception of the modifications to the gravity and other category coding described earlier. Although we therefore acknowledge that our application of the two schemes was not in perfect alignment with the original studies, we do claim that the application in our current study is a reasonable representation of the coding schemes from the original studies.

Framing of the Questions

Following the quasi-replication portion of their study, DG&E added additional types of questions for the extension portion of their study to move beyond requiring only existential and coarse quantitative descriptions of the forces involved. These extension question sets included specification of ontological, compositional, and causal aspects of force. DG&E's goals involved outlining "a plausible set of requirements for specifying important aspects of the content of a concept that is a physical quantity, such as force" (diSessa et al., 2004, p. 854) that extend beyond existential and coarse quantitative aspects. We fully agree with this need for increased specification. Beyond these aspects of increased specification of the nature of students' thinking about forces, however, another problematic issue with the current and the original studies involves the way in which the question sets are framed and phrased.

By asking a student to explain the forces involved in the question contexts, the framing and phrasing of the questions focuses not only on the student's thoughts about the underlying physical mechanisms but also on the student's understanding of the word for *force*. The current study adopted the questions and framings developed in the original studies to allow for comparison and to contribute to the resolution of the ongoing debate, but future work should focus on framing questions in a way that does not rely on students' specific definitions for technical terms. A better approach, for example, would focus on asking students to predict and explain what will happen next in an everyday context. For example,

asking “Where will the stone go after it is thrown? Why will it go there? What determines how far or how fast it will travel along the way?” would allow the interviewer to investigate how students think about the mechanisms driving the event without introducing terminology or depending on students’ definitions or understandings of that terminology. The current format of questions is effective, particularly for the older students, but this alternative approach seems to offer additional affordances. We have completed the initial phases of development for a set of thermodynamics questions based on this approach.

Numbers of Students and Variation Within a Country

Individually interviewing each student requires substantial resources for data collection. Modes of data collection can therefore limit the number of students involved in conceptual change studies. I&V interviewed 105 students. DG&E interviewed 30 students. The current study interviewed 201 students. Larger numbers of students from wider populations of schools and geographic areas would strengthen claims and generalizability. We attempted to increase generalizability in our data collection procedures for the current study by selecting no more than three students in a grade level from any single school, but we would like to draw from a broader population of students in future work.

We will begin to address this question of generalizability by investigating generalizability and variability within Turkey by comparing the Turkish cohort from Özdemir and Clark (2009) with the Turkish cohort from the current study. By comparing variation within the same country, we will begin to account for the possible role of language, educational system, and culture in observed patterns of consistency and meanings across age groups and countries in the current study (Clark, Menekse, Özdemir, D’Angelo, & Schleigh, 2010). We are also exploring the potential of using a written instrument to collect similar data to facilitate data collection across larger groups of students (Schleigh & Clark, 2010).

Problematic Issues With the Gravity and Other Meaning

Many students, especially the middle school and high school students, matched for the gravity and other meaning. As discussed by DG&E, the gravity and other category is a hybrid category that can include many other ideas in addition to gravity. I&V had originally anticipated a strict gravity meaning, but no students in their cohort consistently assigned only gravitational forces in the question sets. I&V therefore modified the original strict gravity category to be applied if a student expressed force ideas about gravity as well as any other force meanings. This became the hybrid gravity and other meaning.

Students were often coded for this force meaning even if they did not express a solid conceptual understanding about the nature of gravity. In addition, the “and

other” component of the gravity and other force meaning leaves a broad range of possibilities lumped within a coding category that is supposed to represent a coherent meaning. We attempted to increase the specificity of the coding for the gravity and other category in this study as described earlier. Further work and specification of gravity and other into subcategories representing specific meanings would strengthen the research community’s ability to make progress in disentangling students’ thinking given the prevalence of this code across the current and original studies.

OTHER POSSIBLE EXPLANATIONS, IMPLICATIONS, AND CLOSING THOUGHTS

The results presented in this article suggest that differences in coding methods and student populations seem unlikely to account for the radical differences in findings of DG&E and I&V. What else could explain these radical differences? Taken together, the (a) coding methods, (b) student populations, (c) interview instruments, and (d) interviewers would seem to represent the four fundamental components of these studies. After ruling out coding methods and the student populations, we propose that the interview instruments and the interviewers become prominent candidates because few other sources of potential difference remain. Our discussion of these two latter possibilities builds on theoretical projection and inference, rather than data, but lays the groundwork for future research.

Interview Instrument Differences

Vosniadou suggested that differences in the interview instruments might explain the differences between their study and DG&E’s (e.g., Wagner, 2005). Differences in the interview instruments could potentially explain the differences in findings between I&V and DG&E as well as the similarity of the current study’s findings with DG&E’s results because (a) DG&E condensed and reorganized the interview question instrument developed by I&V and (b) the current study adopted the interview set condensed by DG&E. In reorganizing I&V’s interview questions, DG&E removed some questions and reorganized the remaining questions into a standardized question set structure involving what they referred to as two “simple questions” and one “comparison” question. It is certainly possible that this reorganization affected how students thought about the questions.

If these differences in the interview instrument could explain the differences in findings, however, what would be the implications for unitary and elemental theories of students’ knowledge structures? On the surface, it would seem that current unitary theories would need to be much more complex because the content of the

interview instruments remains very similar if somewhat abbreviated and reorganized. It is not clear how removing a subset of the questions and reorganizing the remaining questions while retaining the content and representations of the remaining questions nearly verbatim could result in such dramatic changes in levels of consistency from the perspective of a unitary theory. Students should retain and apply their same framework theories because the differences in context are very superficial.

Alternatively, however, current elemental theories can begin to explain the differences in I&V's and DG&E's findings and the similarity of DG&E's and the current study's findings in terms of the changes to the interview instrument. One possible explanation is that I&V's original full set of questions shifted contexts gradually enough to continue to cue the same set of core ideas (e.g., Clark, 2006; diSessa, 1993) and to support the students in maintaining explanatory coherence (Ranney & Schank, 1998; Thagard, 1989, 2007; Thagard & Verbeugt, 1998) across the questions in a manner similar to Clement's (1993, 1998, 2008) bridging analogies. Essentially, each subsequent question in I&V's original instrument could have been similar enough to the preceding questions that students viewed them as similar, which supported the cuing of the same patterns of ideas consistently across questions. In DG&E's reorganized and abridged instrument, however, the gulfs between questions may have become greater, making the question sets appear more distinct, thus removing the "bridging analogy" quality and flow between questions, thus reducing students' perceived requirements for explanatory coherence. Thus, reorganizing the questions could have allowed different sets of cuing priorities to arise and become salient for students in answering each question set. In summary, if the issue involves differences in the interview instruments, we might explain the differences in local ontological coherence between the studies as resulting from differences in the cues inherent across the structure of the interview instrument.

Interviewer Differences

Another potential explanation focuses on the interviewers themselves. Interviewers (along with teachers, peers, parents, and bosses) can unknowingly and inadvertently frame expectations for the goals and purposes of an interaction. It is possible that the interviewers in the current study and in DG&E's study communicated a different set of expectations than did the interviewers in I&V's study, which resulted in students in I&V's study approaching the interviews differently.

Essentially, how interviewers inadvertently framed the interview task through nonverbal cues and verbal responses could have changed how students thought and engaged in the interview process, resulting in radically different outcomes for the interviews. As with differences in the interview instrument, this possible

explanation is potentially more difficult to explain with current unitary theories than with current elemental theories because of the sensitive contextuality implied by the explanation.

In particular, this potential explanation aligns closely with Hammer and Elby's (2002, 2003) and Rosenberg et al.'s (2006) account of elemental epistemological resources. According to their epistemological resources account, people maintain elemental or multiple epistemological components and beliefs that are in some ways analogous to the role of p-prims in ontological understanding (e.g., diSessa, 1993). These resources are cued by the nature of a task, by members of a group, or by authority figures associated with a task. The combination of cued epistemological resources organizes how a person will interpret the nature and goals of a task and how the person thinks about that task. Combinations of epistemological resources can be self-reinforcing, allowing people to maintain local epistemological coherence in how they approach a task (Rosenberg et al., 2006). Thus, the interviewers in I&V may have triggered a different set of expectations and epistemological resources than the interviewers in the other two studies, such that ontological coherence was pursued to a higher degree, resulting in greater consistency across questions in I&V's study.

The current study involved one interviewer in each of the five countries, as described in the Methods section. Two of these five interviewers (those in the United States and Turkey) were closely involved in our work and research group, whereas the interviewers in China, Mexico, and the Philippines were not. This inevitably resulted in some differences in how the interviews were conducted in each country, although as discussed in the section about differences in student populations, these differences were relatively small overall. It is possible, however, that the interviewers in I&V's study, who were closely involved theoretically in the debate from the unitary perspective, inadvertently communicated their expectations of consistency and coherence among the questions to the students through their verbal and nonverbal interactions. This would explain why students were less consistent in DG&E and subsequently in the current study. Essentially, if the issue involves interviewer differences, we might explain the differences in local ontological coherence between the studies as arising from differences in the epistemological resources cued by the interviewers in those studies.

Closing Thoughts and Implications for the Debate

In terms of the ongoing debate over knowledge structure coherence, the results from the five countries in the current study are in much closer alignment with the relatively low levels of consistency observed in DG&E's study regardless of the coding scheme used or student nationality. Although the levels of consistency seen by I&V may not be common, the data do evidence important systematicities in students' thinking.

This study therefore suggests that coding scheme differences and student differences among countries do not appear to be likely candidates for explaining the disparities between I&V's findings and the findings of DG&E and the current study. We propose that, if not coding scheme differences or student differences, the two most likely candidates would involve differences in the interview instruments or the epistemological stances invoked for the participants by the interviewers. These two proposed explanations require further testing and exploration, but, as outlined earlier, unitary theories would require substantial revision to account for the observed combination of local systematicities and fragmentation, whereas current elemental theories could begin to account for the differences in terms of either explanation. Essentially, as elaborated in our discussion of these two proposed explanations, current elemental perspectives account readily for fragmentation as well as local ontological and epistemological coherence in terms of element cuing (Clark, 2006; diSessa, 1993; Hammer et al., 2005), explanatory coherence models (e.g., Ranney & Schank, 1998; Thagard, 1989, 2007; Thagard & Verbeurgt, 1998), epistemological coherence models (Hammer & Elby, 2002, 2003; Rosenberg et al., 2006), and coordination class research (e.g., diSessa & Sherin, 1998; diSessa & Wagner, 2005; Dufresne et al., 2005; Parnafes, 2007; Thaden-Koch et al., 2006; Wagner, 2006). Elemental perspectives thus seem to provide initial steps toward an explanation.

Regardless of theoretical perspective, these results suggest that researchers from both camps now need to focus on adjusting or developing theoretical models to more fully account for the nature of these local systematicities and fragmentations. The global consistencies reported in I&V's study for their Greek students seem rare at the very least. It is now time for theories to account for local coherence and fragmentation in terms of (a) the application of more stringent models and criteria for coherence beyond mere consistency; (b) the structural configurations and relationships of local coherences; (c) the scope and nature of the domains across which students display them; and (d) the impact of educational, cultural, and linguistic variables on them.

Beyond the theoretical implications for the ongoing debate over students' knowledge structure coherence, the results of this study and related studies can help shape educational policy and curriculum design. The major current theoretical perspectives on knowledge structure coherence differ fundamentally in terms of advocating top-down versus bottom-up instructional approaches for scaffolding conceptual change. For example, should curricula focus on helping students revise their existing ideas and connections, or should curricula focus on instilling new perspectives incommensurate with students' existing interpretations? The results of the current study emphasize the importance of working with students to refine their existing ideas and the connections they draw between those ideas as they develop increasingly integrated, parsimonious, and coherent accounts for core science concepts. Similarly, the current study suggests that "conflict" approaches

that focus primarily on “disproving” a student’s misconception and replacing it with a more desirable candidate seem unlikely to succeed. Essentially, curricula should focus on supporting students as they reorganize and reprioritize ideas and the connections they make between ideas rather than on replacing core framework theories.

Finally, ongoing research about differences in how young students from Mexico and other countries may think about force in comparison to U.S. English-monolingual students (who are more frequently studied) can provide insights into developing curricula to better support the diverse underserved student populations in classrooms around the world. Although the current study was not designed to identify the specific sources of differences among students in different countries, it suggests that differences among countries seem to be greatest for younger students. Excellent work has been done about the nature of young students’ alternative conceptions across countries (e.g., Inagaki & Hatano, 2002). The current study provides a foundation for future exploration of the nature of differences in coherence across countries. Ultimately, understanding more about the structure of students’ knowledge will facilitate research and curriculum development to support students as they restructure and build upon that knowledge.

ACKNOWLEDGEMENT

This work was funded by a National Academy of Education/Spencer Postdoctoral Fellowship awarded to Douglas Clark.

REFERENCES

- Aikenhead, G., & Jegede, O. (1999). Cross-cultural science education: A cognitive explanation of a cultural phenomenon. *Journal of Research in Science Teaching*, *36*, 269–287.
- Burkhardt, H., & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher*, *32*(9), 3–14.
- 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: Erlbaum.
- Carey, S. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, *21*, 13–19.
- Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. *Journal of the Learning Sciences*, *14*, 161–199.
- Clark, D. B. (2000). *Scaffolding knowledge integration through curricular depth*. Unpublished doctoral dissertation, University of California at Berkeley.
- Clark, D. B. (2006). Longitudinal conceptual change in students’ understanding of thermal equilibrium: An examination of the process of conceptual restructuring. *Cognition and Instruction*, *24*, 467–563.
- Clark, D. B., Menekse, M., Özdemir, G., D’Angelo, C. M., & Schleigh, S. P. (2010). *Comparison of knowledge structure coherence and force meanings across sites in Turkey*. Manuscript in preparation.

- Clement, J. (1993). Using analogies to deal with students' preconceptions in physics. *Journal of Research in Science Teaching*, 30, 1241–1257.
- Clement, J. (1998). Expert novice strategies and instruction using analogies. *International Journal of Science Education*, 20, 1271–1286.
- Clement, J. (2008). *Creative model construction in scientists and students: The role of imagery, analogy, and mental simulation*. Dordrecht, The Netherlands: Springer.
- Costa, V. (1995). When science is “another world”: Relationships between worlds of family, friends, school, and, science. *Science Education*, 79, 313–333.
- diSessa, A. A. (1983). Phenomenology and the evolution of intuition. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 15–33). Hillsdale, NJ: Erlbaum.
- diSessa, A. A. (1988). Knowledge in pieces. In G. Forman & P. B. Pufall (Eds.), *Constructivism in the computer age* (pp. 49–70). Hillsdale, NJ: Erlbaum.
- 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, England: Blackwell.
- diSessa, A. A. (2006). A history of conceptual change research: Threads and fault lines. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 265–282). Cambridge, England: Cambridge University Press.
- diSessa, A. A., Gillespie, N., & Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28, 843–900.
- diSessa, A. A., & Sherin, B. (1998). What changes in conceptual change? *International Journal of Science Education*, 20, 1155–1191.
- diSessa, A. A., & Wagner, J. F. (2005). What coordination has to say about transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multi-disciplinary perspective* (pp. 121–154). Greenwich, CT: Information Age.
- Dufresne, R., Mestre, J., Thaden-Koch, T., Gerace, W., & Leonard, W. (2005). Knowledge representation and coordination in the transfer process. In J. Mestre (Ed.), *Transfer of learning from a modern multi-disciplinary perspective* (pp. 155–215). Greenwich, CT: Information Age.
- George, J. (1999). World view analysis of the knowledge in a rural village: Implications for science education. *Culture and Comparative Studies*, 83, 77–95.
- Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences*, 8, 371–377.
- 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.
- Gruber, H., & Voneche, J. (1977). *The essential Piaget*. New York, NY: Basic Books.
- Hammer, D., & Elby, A. (2002). On the form of a personal epistemology. In B. K. Hofer & P. R. Pintrich (Eds.), *Personal epistemology: The psychology of beliefs about knowledge and knowing* (pp. 169–190). Mahwah, NJ: Erlbaum.
- Hammer, D., & Elby, A. (2003). Tapping epistemological resources for learning physics. *Journal of the Learning Sciences*, 12, 53–91.
- Hammer, D., Elby, A., Scherr, R. E., & Redish, E. F. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 89–120). Greenwich, CT: Information Age.
- Harrison, A. G., Grayson, D. J., & Treagust, D. F. (1999). Investigating a grade 11 student's evolving conceptions of heat and temperature. *Journal of Research in Science Teaching*, 36, 55–87.
- 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). Cambridge, MA: MIT Press.

- Inagaki, K., & Hatano, G. (2002). *Young children's thinking about the biological world*. Philadelphia, PA: Psychology Press.
- Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*, 2(1), 5–62.
- Keil, F. (1994). The birth and nurturance of concepts by domains: The origins of concepts of living things. In L. Hirschfield & S. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 234–254). Cambridge, England: Cambridge University Press.
- Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). Chicago, IL: University of Chicago Press.
- Linn, M. C. (2006). The knowledge integration perspective on learning and instruction. In K. Sawyer (Ed.), *Cambridge handbook of the learning sciences* (pp. 243–264). Cambridge, England: Cambridge University Press.
- Linn, M. C., Eylon, B., & Davis, E. A. (2004). The knowledge integration perspective on learning. In M. C. Linn, E. A. Davis, & P. Bell (Eds.), *Internet environments for science education* (pp. 29–46). Mahwah, NJ: Erlbaum.
- Linn, M. C., & Hsi, S. (2000). *Computers, teachers, peers: Science learning partners*. Mahwah, NJ: Erlbaum.
- Lubben, F., Netshisaulu, T., & Campbell, B. (1999). Students' use of cultural metaphors and their scientific understandings related to heating. *Science Education*, 83, 761–774.
- McCloskey, M. (1983a, April). Intuitive physics. *Scientific American*, 122–130.
- McCloskey, M. (1983b). Naive theories of motion. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 299–323). Hillsdale, NJ: Erlbaum.
- Minstrell, J. (1982). Explaining the “at rest” condition of an object. *The Physics Teacher*, 20, 10–14.
- Minstrell, J. (1989). Teaching science for understanding. In L. Resnick & L. Klopfer (Eds.), *Toward the thinking curriculum* (pp. 129–149). Alexandria, VA: Association for Supervision and Curriculum Development.
- Minstrell, J., & Kraus, P. (2005). Guided inquiry in the science classroom. In M. S. Donovan & J. D. Bransford (Eds.), *How students learn: History, mathematics, and science in the classroom* (pp. 475–514). Washington, DC: National Academies Press.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Özdemir, G., & Clark, D. B. (2009). Coherence of Turkish students' understanding of force. *Journal of Research in Science Teaching*, 46, 570–596.
- Parnafes, O. (2007). What does “fast” mean? Understanding the physical world through computational representations. *Journal of the Learning Sciences*, 16, 415–450.
- 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., & Schank, P. (1998). Toward an integration of the social and the scientific: Observing, modeling, and promoting the explanatory coherence of reasoning. In S. Read & L. Miller (Eds.), *Connectionist models of social reasoning and social behavior* (pp. 245–274). Mahwah, NJ: Erlbaum.
- Rosenberg, S., Hammer, D., & Phelan, J. (2006). Multiple epistemological coherences in an eighth-grade discussion of the rock cycle. *Journal of the Learning Sciences*, 15, 261–292.
- Sawyer, K. (Ed.). (2006). *Cambridge handbook of the learning sciences*. Cambridge, England: Cambridge University Press.
- Schleigh, S. P., & Clark, D. B. (2010). *Format and sex in assessing the knowledge structure coherence of middle school students' understanding of the concept of force*. Manuscript submitted publication.
- Stigler, J. W., Gallimore, R., & Hiebert, J. (2000). Using video surveys to compare classrooms and teaching across cultures: Examples and lessons from the TIMSS video studies. *Educational Psychologist*, 35(2), 87–100.

- Thaden-Koch, T., Dufresne, R., & Mestre, J. (2006). Coordination of knowledge in judging animated motion. *Physics Education Research*, 2, 1–11.
- Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences*, 12, 435–466.
- Thagard, P. (2007). Coherence, truth, and the development of scientific knowledge. *Philosophy of Science*, 74, 28–47.
- Thagard, P., & Verbeurgt, K. (1998). Coherence as constraint satisfaction. *Cognitive Science*, 22, 1–24.
- Toulmin, S. (1972). *Human understanding* (Vol. 1). Oxford, England: Clarendon Press.
- van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
- 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, The Netherlands: Kluwer Academic.
- Vosniadou, S., & Ioannides, C. (1998). From conceptual development to science education: A psychological point of view. *International Journal of Science Education*, 20, 1213–1230.
- Wagner, J. F. (Chair). (2005, April). *On the nature of students' knowledge: Contrasting epistemologies in science and mathematics education research*. Symposium organized by J. Wagner including D. Clark, A. diSessa, J. Mestre, S. Vosniadou, and J. Wagner for the American Educational Research Association Annual Conference 2005, Montreal, Quebec, Canada.
- Wagner, J. F. (2006). Transfer in pieces. *Cognition and Instruction*, 24, 1–71.
- Wellman, H. M., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. Stevens (Eds.), *Mental models* (pp. 267–298). Hillsdale, NJ: Erlbaum.

APPENDIX

Annotated Example of Interview Coding Using Both Schemes

This appendix provides an annotated example of the coding of the interview with a fifth-grade Mexican student (to whom we refer with the pseudonym “Francisco”) to give readers a better sense of the data and the coding procedures. Before presenting the transcript, we first provide an overview of the process for using Ioannides and Vosniadou’s (2002) and diSessa, Gillespie, and Esterly’s (2004) coding schemes, and provide a summary of the interview and coding.

The process of coding students’ responses with I&V’s scheme involves first matching the student’s answer with question set level codes for that question (see Table 1 for the possible question set level codes for Question Set 1). Question set level codes are then mapped onto overall possible force meaning matches using a rubric for that question (see Table 2 for the rubric that matches question set level codes with force meanings for Question Set 1).

The process of coding students’ responses with DG&E’s scheme involves first identifying the coarse quantitative assertions that the student is making regarding whether or not there are forces acting on the stones in each picture (and which stones have larger forces acting on them if both stones in a question

set are identified as having forces acting on them). Then the student's explanations are examined for the presence of potential qualifiers. Finally, the student's coarse quantitative assertions and qualifiers are compared to a rubric for overall potential force meaning matches (see Table 3 for the rubric for Question Set 1).

Francisco's annotated transcript presents and summarizes his responses to each question set according to both schemes and discusses other relevant considerations. The force meaning category matches for each question set for each scheme appear in bold. Space considerations preclude presenting the full coding rubrics for I&V and DG&E here, but we make them available as supplemental materials in the publisher's online edition of *Journal of the Learning Sciences*. In addition to the full set of rubrics for both schemes, the online companion materials also include a document outlining the rules we used to resolve specific coding issues. We refer to these rules in the annotated transcript, e.g., the "gravity rules."

During this interview, Francisco responds with a few main ideas about force. His answers sometimes suggest, for example, that force depends on size or that force has something to do with motion or the ability to move something. Some of these ideas change as the context of the question changes. Sometimes there is a force if the object cannot be moved (it is heavy), and other times there is a force if the object is not moving but might move (because it is heavy). There is a force sometimes when it cannot be moved and sometimes when it is moving. However, if it can be moved, but is not moving, then there is no force. With respect to gravitational forces, Francisco never mentions gravity, gives a conceptual description of gravity, or talks about things being pulled toward the ground. As a result, although many of the responses have something to do with things being heavier or lighter, the responses do not get coded for gravity. This is explained further in our online coding rules document in the section on gravity.

According to I&V's coding scheme, Francisco codes most frequently (6 question sets out of 10) for the internal force meaning and for the internal/movement force meaning. This means that Francisco's best-match force meaning for I&V's scheme is a tie between the internal and internal/movement force meanings at 6 out of 10 question sets each. According to DG&E's coding scheme, Francisco codes most frequently (6 question sets out of 10) for the internal force meaning. His best-match force meaning for DG&E's scheme is therefore the internal force meaning with 6 out of 10 question sets. Francisco therefore does not code as fully consistent (matching for 10 out of 10 question sets for a force meaning) or consistent with allowance (8 out of 10 question sets for a force meaning) according to I&V's or DG&E's coding schemes.

TABLE A1
Annotated Coding of Francisco's Responses to Question 1

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 1: Big vs. small stone on ground	No	Can't move	<i>Question Set Level Codes</i> f. Force only on the small stone	<i>Coarse Quantitative Comparison</i> Force only on the small stone <i>Qualifiers</i> Not due to gravity
Part A I: This stone is standing on the ground. Is there a force on this stone? S: No I: Why? S: Because it cannot move				
Part B I: This stone is standing on the ground. Is there a force on this stone? S: A little bit I: Why? S: Because it is small I: And what happens if the stone is small? S: It may fall	Yes	Small & could fall	<i>Resulting Force Meaning Matches</i> Acquired/Push-Pull	<i>Resulting Force Meaning Matches</i> Acquired/Push-Pull
Additional comments	The comparison question was not asked because there was a force on only one stone.			

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A2
Annotated Coding of Francisco's Responses to Question 2

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 2: Stable vs. unstable similar stones	Yes	Heavy; could fall	<i>Question Set Level Codes</i> d. Same force on both stones (because they are similar, equally big, equally heavy)	<i>Coarse Quantitative Comparison</i> Equal force on both stones <i>Qualifiers</i> Not due to air, gravity, or ground
Part A I: This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? S: Yes I: Why? S: It is very heavy I: And what happens then? S: It may fall off the hill I: So, there is force on the stone? S: Yes				
Part B I: 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? S: Yes I: Why? S: Because it is big	Yes	Big	<i>Resulting Force Meaning Matches</i> Internal	<i>Resulting Force Meaning Matches</i> Internal Acquired/Push-Pull
Part C I: Is the force on this stone the same or different than the force on this stone? S: They are the same I: Why? S: Because they are the same [pointing to both stones in the picture] I: Do you mean that the stones are the same size? S: The same as what they have inside	Same	Same stone; same inside		
Additional comments				

Using DG&E's scheme, we did not code equal force on both stones for gravity and other because of the Gravity Rule: The student did not specifically identify gravity or give a conceptual explanation of it.

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A3
Annotated Coding of Francisco's Responses to Question 3

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 3: Unstable small vs. unstable big	No	Small, little weight	<i>Question Set Level Codes</i> a. Force only on the big stone (because it is big and/or heavy and/or it falls it can cause damage. No force on the small stone because it is small and/or light and/or it's more stable so it cannot fall)	<i>Coarse Quantitative Comparison</i> Force only on the big stone <i>Qualifiers</i> Not due to air, gravity, or ground
Part A I: This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? S: No I: Why? S: Because it is small and has almost no weight				
Part B I: This stone is standing on a hill. It is unstable. That means it could easily fall down. Is there a force on the stone? S: It is big and heavy [referring to the big stone in the picture] I: So, do you think that there is force on the stone or not? S: Yes I: Why? S: Because it is bigger	Yes	Big & heavy	<i>Resulting Force Meaning Matches</i> Internal Internal/Movement Internal/Acquired	<i>Resulting Force Meaning Matches</i> Internal Internal/Acquired
Additional comments				
The comparison question was not asked because there was a force on only one stone.				

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A4
Annotated Coding of Francisco's Responses to Question 4

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 4: Falling big vs. standing big	Yes	It's falling	<i>Question Set Level Codes</i> c. Force only on the falling stone (because it is falling. It can cause damage)	<i>Coarse Quantitative Comparison</i> Force on the falling stone only <i>Qualifiers</i> Not due to gravity
Part A I: This stone is falling. Is there a force on the stone? S: Yes I: Why? S: Because it is falling I: What do you mean by that? S: I mean that when it is falling it is very heavy I: And what happens? S: [long pause] [no answer]				
Part B I: This stone is standing on the ground. Is there a force on this stone? S: No I: Why? S: Because it is on the floor	No	On the floor (not falling)	<i>Resulting Force Meaning Matches</i> Acquired Acquired/Push-Pull	<i>Resulting Force Meaning Matches</i> Acquired Acquired/Push-Pull
Additional comments		The comparison question was not asked because there was a force on only one stone.		

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A5
Annotated Coding of Francisco's Responses to Question 5

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 5: Falling big vs. falling small	Yes	Big; falling	<i>Question Set Level Codes</i> b. Greater force on the big stone (because both stones have weight but the first stone has more weight and/or it will fall with impetus/force)	<i>Coarse Quantitative Comparison</i> Force on both stones but greater force on the big stone <i>Qualifiers</i> Not due to air, gravity, or an acquired force
Part A I: This stone is falling. Is there a force on the stone? S: Yes I: Why? S: Because it is big and it is falling with force	Yes	Big; falling		
Part B I: This stone is falling. Is there a force on the stone? S: A little bit I: Why? S: Because it is small	Yes	Small	h. Force on both stones due to motion (but greater force on the big stone because it falls with greater force or it gains greater force as it falls or it is the kinetic force)	
Part C I: Is the force on this stone the same or different than the force on this stone? S: They are different I: Why?	Different	Size		<i>Resulting Force Meaning Matches</i> Internal

(Continued)

TABLE A5
(Continued)

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
	S: Because one is big and the other is small		Internal/Acquired Acquired Acquired/Push–Pull	Internal/Acquired Acquired Acquired/Push–Pull
Additional comments	Because it is not clear whether the student thought that the forces were different or whether the student was just describing the physical properties of the stones, we coded based on the Overcoding Rule, choosing all possible responses that were equal or force on both. Only the coding for gravity and other was not included using DG&E's scheme because of the need for the student to specifically identify gravity according to our Gravity Rule.			

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A6
Annotated Coding of Francisco's Responses to Question 6

<i>Interview Section</i>	<i>Force Present</i>		<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
	<i>Is</i>	<i>Yes</i>			
Question 6: Man tries to move big stone vs. tries to move small stone	Is there a force on the stone?	Yes	Heavy; can't be moved	<i>Question Set Level Codes</i> a. Force only on the big stone (because it is big and heavy and/or the man cannot move it. No force on the small stone because it is small and/or light and/or the man can move it easily)	<i>Coarse Quantitative Comparison Qualifiers</i> Force only on the big stone Not due to air, gravity, or person
Part A I: This man is trying to move this stone. Is there a force on the stone? S: Yes I: Why? S: Because it is heavy and the man will not be able to move it					
Part B I: This man is trying to move this stone. Is there a force on the stone? S: No, because that one is small [referring to the small stone on the table] and the man is able to move it	Is	No	Can be moved	<i>Resulting Force Meaning Matches</i> Internal Internal/Movement Internal/Acquired	<i>Resulting Force Meaning Matches</i> Internal Internal/Movement Internal/Acquired
Additional comments	The comparison question was not asked because there was a force on only one stone. Although the force is connected to the man and his ability to move the stone, the rationale is more directly related to the ability to move the stone rather than to the man specifically. Therefore, coding using DG&E's scheme does not conflict with the qualifier "but not due to person." The question is also not coded using I&V's scheme as a "force from the man."				

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A7
Annotated Coding of Francisco's Responses to Question 7

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 7: Man can't move big stone vs. can't move small stone	Yes	Big; weight	<i>Question Set Level Codes</i> a. Force related to the size of the stones (greater force on the big stone because it is bigger/heavier)	<i>Coarse Quantitative Comparison</i> Force on both stones but greater force on the big stone <i>Qualifiers</i> Not due to air, gravity, or person
Part B				
I: This man is trying to move this stone and it won't move. Is there a force on the stone?	Yes	Small; weight	<i>Resulting Force Meaning Matches</i>	<i>Resulting Force Meaning Matches</i>
S: It has a little force			Internal	Internal
I: So, you are saying that there is force on the stone?			Internal/Movement	Internal/Movement
S: Yes			Internal/Acquired	Internal/Acquired
I: Why?				
S: Because it is small and weighs a little				
Part C				
I: Is the force on this stone the same or different than the force on this stone?	Different	Small stone is a little heavy		
S: They are different but [hesitation] this stone [referring to the small stone in the picture] is a little heavy				
I: So they are different but this stone [pointing to the small stone in the picture] is a little heavy. Why?				
S: Because it is small and it may have force				
Additional comments				
This was coded for "more on larger" because the smaller stone "may" have force and is "a little heavy," which seems to imply that the larger stone definitely has a force and is heavier (i.e., more force). The student does not identify the person or the person's effort as a force but does indicate that size and weight matter. Therefore, using both DG&E's and I&V's schemes, we coded only forces due to internal characteristics (size, weight) and not forces due to the person.				

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Estery (2004); I = interviewer; S = student.

TABLE A8
Annotated Coding of Francisco's Responses to Question 8

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
<p>Question 8: Part A I: This man is trying to move this stone and it won't move. Is there a force on the stone? S: It is big and it is heavy I: So, is there force on it? S: Yes</p>	Yes	Big & heavy	<p><i>Question Set Level Codes</i> a. Force related to the size of the stones (same force on both stones because the two stones are similar)</p>	<p><i>Coarse Quantitative Comparison</i> Equal force on both stones <i>Qualifiers</i> Not due to air, person, or gravity</p>
<p>Part B I: This child is trying to move this stone and it won't move. Is there a force on the stone? S: Yes, it is big and heavy I: So, you are saying that there is force S: Yes</p>	Yes	Big & heavy	<p><i>Resulting Force Meaning Matches</i> Internal Internal/Movement Internal/Acquired</p>	<p><i>Resulting Force Meaning Matches</i> Internal Internal/Movement Internal/Acquired</p>
<p>Part C I: Is the force on this stone the same or different than the force on this stone? S: They are the same I: Why? S: Because they may have the same weight</p>	Same	Weight		
<p>Additional comments Following the Gravity Rule (i.e., the student says that the bigger stone has more force but does not specifically identify gravity or give any conceptual description of gravity), the student is not coded for the gravity and other category. The student's rationale of weight does not confirm that weight is connected to gravity.</p>				

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A9
Annotated Coding of Francisco's Responses to Question 9

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
<p>Question 9: Man throwing stone vs. similar stone on ground</p> <p>Part A I: This man has thrown this stone. Is there a force on the stone? S: No, because it may not be heavy since the man was able to lift it I: So, there is no force because the man was able to lift it? S: Yes</p> <p>Part B I: This stone is standing on the ground. Is there a force on this stone? S: Yes, it is big and heavy and the man is has not strength to lift it</p> <p>Additional comments The comparison question was not asked because there was a force on only one stone.</p>	No	Not heavy; man can move it	<p><i>Question Set Level Codes</i> g. Force only on the stationary stone (no force on the moving stone because the man is able to move it)</p> <p><i>Resulting Force Meaning/Match Internal/Movement</i></p>	<p><i>Coarse Quantitative Comparison</i> Force only on the stationary stone <i>Qualifiers</i> Not due to air or gravity</p> <p><i>Resulting Force Meaning/Match Internal/Movement</i></p>

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.

TABLE A10
Annotated Coding of Francisco's Responses to Question 10

<i>Interview Section</i>	<i>Force Present</i>	<i>Student Rationale</i>	<i>I&V Coding</i>	<i>DG&E Coding</i>
Question 10: Thrown small stone vs. thrown big stone	No	Man can throw it	<i>Question Set Level Codes</i> f. No force on any stones because they have been pushed (because the man threw them they must not be so heavy)	<i>Coarse Quantitative Comparison</i> No force on any stone
Part A I: This man has thrown this stone. Is there a force on the stone? S: No, because the man was able to throw it and the stone is small so the stone does not have force I: So, you are saying that the stone does not have force because the man was able to throw it S: Yes	No	Man can throw it		
Part B I: This man has thrown this stone. Is there a force on the stone? S: No, because the man was able to throw it and the stone is big and so the stone may not be so heavy I: So, again you are saying that there is not force because the man was able to throw it [referring to the big stone in the picture] S: Yes	No	Man can throw it	<i>Resulting Force Meaning Matches</i> Internal/Movement	<i>Resulting Force Meaning Matches</i> Internal/Movement Push-Pull
Additional comments	The comparison question was not asked because there was no force on either stone.			

Note. The force meaning category matches for each question set for each scheme appear in bold. I&V = Ioannides and Vosniadou (2002); DG&E = diSessa, Gillespie, and Esterly (2004); I = interviewer; S = student.