

Exploring Sources of Variation in Studies of Knowledge Structure Coherence: Comparing Force Meanings and Force Meaning Consistency Across Two Turkish Cities

DOUGLAS B. CLARK,¹ MUHSIN MENEKSE,² GOKHAN OZDEMIR,³
CYNTHIA M. D'ANGELO,⁴ SHARON PRICE SCHLEIGH⁵

¹*Peabody College of Education, Vanderbilt University, Nashville, TN 37203, USA;* ²*LRDC, University of Pittsburgh, Pittsburgh, PA 15260, USA;* ³*Faculty of Education, Nigde University, Nigde 51240, Turkey;* ⁴*SRI International, Menlo Park, CA 94025, USA;* ⁵*Purdue University Calumet, Hammond, IN 46323, USA*

Received 9 July 2012; accepted 29 October 2013

DOI 10.1002/sce.21094

Published online 18 December 2013 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Substantial variation has been observed across an international series of studies examining the consistency of students' explanations of force and the most common meanings of force apparent in those explanations. On the surface, the variations among studies might be attributed to differences at the national level, but the studies also demonstrate differences among students from different schools in the United States. To what degree, therefore, can these variations be attributed to differences in educational systems as opposed to demographic differences or random variation? The current study compares student interviews across two cities in Turkey to provide insight into this question because Turkey, unlike the United States, has a strongly standardized national educational system. The results demonstrate no significant differences in students' consistency or meanings of force between cities. The results, however, demonstrate the expected differences across ages and majors, which suggest that the study has sufficient power. Thus, while differences

Correspondence to: Douglas Clark; e-mail: 42dc42@gmail.com

have been observed between every city and country in the previous studies, and differences are observed in the current study in terms of grade level and academic majors, no differences are observed between the cities in Turkey. The implications of these findings are discussed in terms of ongoing conceptual change research. © 2013 Wiley Periodicals, Inc. *Sci Ed* 98:143–181, 2014

INTRODUCTION

Substantial variation has been observed across an international series of studies examining the consistency of students' explanations of force and the most common meanings of force apparent in those explanations (Clark, D'Angelo, & Schleigh, 2011; diSessa, Gillespie, & Esterly, 2004; Ioannides & Vosniadou, 2002; Ozdemir & Clark, 2009; Price Schleigh, Clark, & Menekse, 2013). The two theoretical perspectives have evolved since the original studies (and thus the current study is not intended as a test of those original theoretical perspectives). Rather, the purpose of the current study is to clarify possible sources of the variation. By clarifying possible sources of variation, the current study contributes to a larger foundation for the field that can allow the field to move forward in developing finer grained and more accurate models that can account for the sources of variation.

RATIONALE FOR THE STUDY

The observed variation in findings across studies is interesting because (a) the studies employed highly similar methods and (b) subsequent analyses have suggested that coding schemes are not the source of the variation (Clark et al., 2011; Ozdemir & Clark, 2009). On the surface, the observed variation might be attributed to differences at the national level (which might include differences in educational system, language, or national culture), but the studies also demonstrated variation between students from different schools in the United States (Clark et al., 2011; diSessa et al., 2004; Price Schleigh et al., 2013). This variation in the U.S. samples suggests that language and national culture are not the primary factors because almost all of the students in the U.S. studies spoke the same native language and hailed from the same overarching national culture (although culture also clearly varies within nations).

Educational systems present themselves as a potentially important source of the observed variation because (a) educational systems vary substantially between countries and (b) the United States does not internally have a strongly standardized national educational system. Even schools within a single city in the United States often differ substantially in terms of curriculum, teacher preparation, and assessment practices. Comparing across schools in the United States therefore does not allow us to determine whether the variations we observe in terms of consistency and force meanings are a function of enacted educational systems as opposed to other local variables and/or random variation. Similarly, we cannot attribute the observed variations between countries to educational system as opposed to other differences or random variation because even more variables are involved in international comparisons than in U.S. comparisons. Comparing two cities in a country with a strongly standardized national educational system, however, would allow us to isolate many of those variables.

Turkey provides an excellent context for such a comparison. Turkey has a strongly standardized national educational system in terms of curriculum, teacher preparation, and assessment. Thus, comparing across two cities in Turkey would potentially hold the enacted educational system much more constant than a comparison of different schools in the United States. Furthermore, a comparison across two cities in Turkey would hold language and overarching national culture constant (although local cultural and demographic variations

clearly exist across cities in Turkey just as in any other country), while ensuring that the students are selected from nonoverlapping school sites, teachers, and communities.

The current study leverages these affordances of Turkey to explore the degree to which students from two cities in Turkey display different levels of consistency in their explanations of force across question contexts or different predominant force meanings in those explanations. If the results show that the outcomes are highly similar across the two cities, we might interpret the observed similarities as supporting the claim that differences between educational systems may be primary contributors to the differences observed across countries and within the United States in the previous studies. If the students in the two cities are not highly similar in the current study, then we might assume that other local variables and/or random variation trump the importance of educational system in terms of the variation observed across countries and within the United States in the previous studies.

In support of the primary comparison between cities, the current study simultaneously analyzes the same data across three other variables (grade level, gender, and academic major) to provide baselines for interpreting the magnitudes of variation observed between cities. On the basis of prior studies, we would expect students' consistency and specific force meanings to vary by grade level but not by gender (Price Schleigh et al., 2013). If we observe differences across grade level and an absence of differences across gender, we may (a) assume that our analytical approaches and student samples are representative of the analyses and data samples from the earlier studies and (b) use the magnitudes of any observed differences as baselines for interpreting the magnitude of any observed differences between cities. The additional comparison across academic major for the high school students in one of the cities provides a third baseline in terms of the variation resulting from more personal variables (e.g., individual interests, abilities, experiences, or affiliations).

The goal of the current study is therefore to explore these questions by conducting a replication study in two cities in Turkey that have very similar enacted educational systems as a function of the country's nationalized curriculum. More specifically,

1. How do students' levels of consistency and predominant expressed force meanings vary across the two cities relative to the variations observed across previous studies?
2. How does the level of variation detected across cities compare to the levels of variation detected across grade levels (where we expect substantial variation) and gender (where we would not expect significant variation)?
3. How do students' levels of consistency and predominant expressed force meanings vary across academic majors (which the current study characterizes as proxies for more personal variables including individual interests, abilities, experiences, and affiliations)?

THEORETICAL BACKGROUND

The following sections provide salient background information for the current study in terms of three critical areas. First, we provide background on the international series of conceptual change studies that the current study replicates and extends. Second, we provide background on the nationalized educational system in Turkey to support our claim that the enacted educational system across Turkey is relatively standardized in comparison to the enacted system in the United States. Third, we provide background on the system of high school academic majors in Turkey, along with previous research on the relationships of majors to academic outcomes, to clarify what the comparison across majors in the current study represents. In sum, these three sections are intended to provide key background

information to clarify the nature and implications of the comparisons made by the current study.

Foundational Conceptual Change Studies Informing the Current Study

Ioannides and Vosniadou (2002) conducted what was to become the foundational study in a subsequent string of quasireplication studies exploring the consistency of the force meanings that students of different ages express in their explanations across contexts. Ioannides and Vosniadou's goal was to explore a fundamental debate in the conceptual change literature regarding the structure and coherence of students' scientific understandings as coherent unified schemes of theory-like character (e.g., Carey, 2000; Gopnik & Schulz, 2004; McCloskey, 1983; Wellman & Gelman, 1992; Wiser & Carey, 1983) versus ecologies of quasi-independent elements (e.g., Clark, 2006; diSessa, 1983, 1993; Hammer, Redish, Elby, & Scherr, 2005; Linn, 2006; Thaden-Koch, Dufresne, & Mestre, 2006).











Vosniadou encouraged diSessa to conduct a similar study in the United States. diSessa et al. (2004) condensed and reorganized Ioannides and Vosniadou's question sets into 10 sets that each involve two initial yes/no questions and one comparison question (Figures 1a and 1b). In addition to reorganizing the question sets, diSessa et al. (2004) also revised the coding approach in a manner that they felt they could apply more reliably.

There were substantial differences between the findings of Ioannides and Vosniadou (2002) and diSessa and colleagues (2004) in terms of the levels of consistency of force meanings that students seemed to employ in their explanations. In subsequent discussions between Vosniadou and diSessa (e.g., Wagner, 2005), two promising explanations for the differences in findings focused on differences between the student populations and differences in coding methods. The discussion of differences in student populations focused on language differences. In the Greek language, the term for force (*dynamis*) is commonly used colloquially with a meaning (i.e., "power" or "strength") that parallels the scientific meaning more closely than does the colloquial meaning of "force" in the United States. The discussion of differences in coding methodology focused on the fact that 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). Thus, even though diSessa and colleagues (2004) designed their coding scheme to assign consistency more liberally, it was plausible that the differences in the coding schemes contributed to the radically different findings of Ioannides and Vosniadou (2002) and diSessa and colleagues (2004).

Ozdemir and Clark (2009) therefore conducted a third study with the same question sets and grade levels in a third country (Turkey) with a third language (Turkish) and coded the data using coding schemes based on both Ioannides and Vosniadou (2002) and diSessa and colleagues (2004). In terms of levels of consistency, the findings were intermediate to those of diSessa and colleagues and Ioannides and Vosniadou but closer to those of diSessa and colleagues. In terms of the coding schemes, the findings suggested that the differences between the findings of the original studies were not a function of the coding methods.

Clark et al. (2011) further explored the potential contributions of coding scheme differences and national differences across five countries (Mexico, China, the Philippines, the United States, and a new cohort from a different city in Turkey) using the same grade levels, consolidated question sets, and coding schemes from diSessa et al. (2004) and Ozdemir and Clark (2009). Selecting student samples from multiple countries was deemed important to explore the possibility of national differences because semantic and cultural differences have been shown to impact students' thinking about specific science concepts (Aikenhead

(a)

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?"

(b)











Question Set	Drawing A	Question A	Drawing B	Question B	Comparison Question
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 1. (a) Question Sets 1–5 that DG&E condensed from I&V's study. (b) Question Sets 6–10 that DG&E condensed from I&V's study.

& Jegede, 1999; Costa, 1995; George, 1999; Inagaki & Hatano, 2002; Lubben, Netshisaulu, & Campbell, 1999), although other differences in terms of educational systems and schools might also likely contribute differences in outcomes between student populations. Clark et al. (2011) demonstrated that differences in coding schemes seemed unlikely to account for the magnitude of differences in the findings of the original studies. The analyses did, however, demonstrate differences across countries in terms of consistency and meanings that might result from language, culture, or educational systems.

Price Schleigh et al. (2013) built on these findings by exploring the role of assessment format. The study was conducted with students at a U.S. school different from the school in diSessa et al. (2004) or the U.S. schools in Clark et al. (2011). While Price Schleigh et al. (2013) focused primarily on the role of assessment format, the study also highlighted differences in primary force meanings expressed by the U.S. students in each study across diSessa et al. (2004), Clark et al. (2011), and Price Schleigh et al. (2013).

Influenced by the ongoing series of studies and other research, current iterations of framework theory and knowledge-in-pieces perspectives have evolved to share many similarities in their explicit prediction of fragmentation and coherence (Brown & Hammer, 2008, 2013; Clark & Linn, 2013; Vosniadou, 2013; Vosniadou & Skopeliti, 2013). In her recent article, for example, Vosniadou (2013) summarizes several shared features of current framework theory and knowledge-in-pieces perspectives. The distinctions between current perspectives (and focal areas of research for both perspectives) now focus on (a) the specific nature of the knowledge elements that apply high magnitudes of influence on other elements in the conceptual ecologies and (b) the processes through which stabilities evolve and change across conceptual ecologies (e.g., Brown & Hammer, 2013; Clark & Linn, 2013; Vosniadou & Skopeliti, 2013; Vosniadou, 2013).

While the two theoretical perspectives have thus evolved since the original studies, understanding the sources of variation observed across the previous studies is critical for the field to move forward. More specifically, the evolving theoretical models need to be able to account for the variations in findings in terms of the factors and mechanisms that contribute to differences across populations. Furthermore, the researchers have been working with data sets that support different conclusions. Thus it remains important to better understand the sources of the variation observed (a) across countries across studies, (b) across countries in Clark et al. (2011), and (c) across U.S. samples across studies. By clarifying possible sources of variation, the current study therefore contributes to a larger foundation for the field that can allow the field to move forward in developing more accurate models that account for sources of variation in terms of the consistency of students' explanations and the ideas expressed in those explanations. While the current study focuses on students' conceptions of force, the presumption across the series of studies is that the patterns observed in terms of students' conceptions of force potentially extend to student understanding of a larger constellation of science ideas, the breadth of which can be explored through future research by the field building on the foundations developed through this series of studies.

Standardization Across the Turkish National Education System

Turkey's educational system focuses on the explicit goal of supporting equality and national coherence by pursuing a rigorously standardized and consistent national educational system (The Ministry of National Education [MONE], 2001). Science education has a long history as a central focus within this educational system (e.g., Ayas, Cepni, & Akdeniz, 1993). Education and instruction are planned, provided, and managed through a centralized governmental system.

MONE is responsible for pre-K through 12 education. Standardized courses are developed in terms of curriculum, instruction, and assessment for every subject at every grade level. These standardized courses are implemented in all schools across Turkey and books are written in direct support of these courses. The books used in all K-12 schools are approved by a special agency within the ministry. The books are then published and distributed to the schools at no cost. Three types of books are used in K-8 for each class at every grade level. These include a student course book, a student workbook, and a teacher guidebook. Similar materials are provided for secondary courses.

MONE staff travel to the schools to evaluate and provide feedback to the schools and teachers. Teacher professional development events, workshops, certificate learning programs, and conferences are held throughout Turkey under the control and supervision of MONE and the universities. Teachers are financially supported to attend these activities. The Council of Higher Education, which is responsible for higher education, specifies the courses offered in education departments and teacher education programs as well as the content of those courses. All teacher candidates therefore complete similar courses with similar content designed in conjunction with the standardized system of courses across grade levels and subjects in K-12 schools.

National and international assessments suggest that the Turkish system contributes to similar outcomes across the geographic regions of Turkey. Berberoglu and Kalender (2005), for example, compared students' science and mathematics scores on the Higher Education Entry Examination (formerly known as Student Selection Examination) from 1999 to 2002 across Turkey's seven geographical regions. Berberoglu and Kalender found that region did not manifest a strong main effect due to very low partial eta-square value in terms of the quantitative test scores. The region comparisons for the verbal test demonstrated the same result in this study. Similarly, although there was a statistical difference for mathematics literacy across the seven geographical regions in Turkey for the 2003 Program of International Student Assessment's (PISA) secondary data for eighth-, ninth-, and tenth-grade students, this difference was not meaningful in practice because of a low effect size (Berberoglu & Kalender, 2005). This is not to say that there are not differences between individual schools and teachers or across regions. Individual schools do vary along a number of variables between lower and higher performing schools (e.g., Aypay, Erdogan, & Sozer, 2007). Furthermore, as Titrek and Cobern (2011) explain, Turkey is becoming a more Westernized society paralleling the United States in terms of the coexistence of deeply embedded religious culture with science. Student and family attitudes about science in Turkey are influenced by religious beliefs (Ornek, 2011), which parallels findings in Egypt that teachers' personal Islamic religious beliefs inform their beliefs about the nature of science and its purpose (Mansour, 2011).

This cultural balance in Turkey differs across regions and within individual cities and communities (Dogan & Abd-El-Khalick, 2008). Dogan and Abd-El-Khalick, for example, demonstrated a trend toward more Western/"sophisticated" beliefs about the nature of science across regions of Turkey when moving from east to west that parallel increasingly secular cultural views and increasing median socioeconomic status by region. This variation across regions aligns with the findings from a comprehensive international review (Deng, Chen, Tsai, & Chai, 2011) and the findings of more focused studies in Turkey (Dogan & Abd-El-Khalick, 2008; Kalender & Berberoglu, 2009) and in other countries (e.g., Constantinou, Hadjilouca, & Papadouris, 2010; Griffiths & Barman, 1995; Karabenick & Moosa, 2005; Osman & Cobern, 2011; Sutherland & Dennick, 2002; Wen, Kuo, Tsai, & Chang, 2010) that cultural, socioeconomic, and demographic variables can impact students' beliefs about the nature of science.

Thus Turkey's educational system involves substantial standardization in terms of curriculum, teacher preparation, and assessment within a country that includes demographic and cultural diversity paralleling the demographic and cultural diversity common in many countries. We hypothesize that we should not see differences between the two cities in the current study in terms of force meanings expressed and the consistency of those force meanings if the educational system is a primary determinant of force meanings and consistency. If we were to observe substantial variation between students in the two cities, however, that would suggest that local variables or random variation overshadow the impact of a national educational system even in a country with a strongly nationalized educational system.

Similarities Across High School Academic Majors

As discussed, the current study also analyzes the data across grade level, gender, and academic major to provide baselines for interpreting the magnitudes of variation observed across cities. While the nature of the comparisons involved in terms of grade level and gender is relatively transparent, the comparison across academic majors, however, requires clarification. Essentially, the data allow us to analyze variation in the levels of consistency and force meanings of 10th- and 11th-grade students across academic majors within the same high schools in one of the cities. Whereas we would expect the strongly nationalized educational system to minimize variation between the two cities, we would potentially expect larger variation across academic majors due to differences in students' interests, abilities, experiences, and affiliations in addition to the differences in courses across majors during the 10th and 11th grades. Thus, the comparison across majors within a single city provides (a) an interesting baseline for comparing the magnitude of the variation across cities and (b) an interesting, although imperfect, lens to begin thinking about the role of personal variables including students' interests, abilities, experiences, and affiliations.

At the time of this study, all students in all majors took the very same courses from first through ninth grades.¹ These courses included a substantial focus on science culminating with two credits of physics, chemistry, and biology in ninth grade. After ninth grade, students in both general and Anatolian² high schools chose to specialize in *science and mathematics*, *Turkish and mathematics*, *social sciences*, or *foreign language*. Depending on grade and major, 12–19 hours of courses of a total of approximately 30 hours in a week were devoted to the courses in a student's selected major. Only students in the science and mathematics major took science courses after ninth grade.

In comparing across majors, it is also important to understand that high school students' decisions regarding majors impacted the Higher Education Entry Examination (YGS) examination at the time of this study because each of a student's subscale scores on the examination was computed with different coefficients depending on the student's high school major. When a student chose a major, the student's interests, abilities, achievement in ninth-grade courses, and career goals were thus considered. Teachers, school counselors, and parents provided suggestions based on this information. Students, however, ultimately chose which major they pursued.

In terms of prior comparisons of Turkish students across majors, we could find no research focusing on science learning across majors, potentially because only science and mathematics students took science courses after the ninth grade. Studies that compare

¹The Turkish system is currently undergoing a change wherein students choose their path in fifth grade, but the system described here is the one that was in effect at the time of this study.

²Anatolian high schools are public high schools in Turkey that admit students based on the national high school entrance examination.

Turkish students across majors have been conducted, however, in terms of mathematics and/or computer attitudes, learning styles and learning strategies, and multiple intelligence profiles.³ These studies suggest that (a) the Science and mathematics students tend to express the highest interest and liking of mathematics (Çelik & Ceylan, 2009; Yıldız & Turanlı, 2010); (b) the social sciences students tend to demonstrate higher linguistic skills, interests, and sophistication of study skills in related domains (Hamurcu, Günay, & Özyılmaz, 2002; Pehlivan, 2008; Sezer, 2010; Sünbül & Sarı, 2004); (c) the Turkish and mathematics students tend to be intermediate between students from the two other majors along these dimensions (Hamurcu et al., 2002); and (d) there are no differences across majors in terms of affinity and attitudes toward computers (Çelik & Ceylan, 2009). These findings parallel those of other similarly focused studies in terms of underscoring patterns of differences across majors with each major demonstrating strengths and weaknesses as opposed to any strict overall hierarchy of student quality across majors (e.g., Tunç, 2008; Demiray & Dolu, 2011; İzci, Kara, & Dalaman, 2007; Yenice & Aktamış, 2010).

Comparisons across majors have been conducted in other countries (e.g., Chai, Deng, Qian, & Wong., 2010; Miller, Montplaisir, Offerdahl, Cheng, & Ketterling, 2010). In Liu and Tsai's (2008) study, for example, nonscience undergraduate majors scored lower than science undergraduates in terms of the views on the nature of science (VNOS) dimensions focusing on the social and inventive nature of science, but the nonscience majors scored higher than the science majors on the theory-laden and cultural dimensions of the VNOS. Beyond choice of major, affective characteristics more generally have been shown to affect physics achievement in Turkey (e.g., Gungor, Eryilmaz, & Fakioglu, 2007). Kalendar and Berberoglu (2009) demonstrated that a student's personal interests can statistically increase science achievement as a function of out-of-school activities. Furthermore, personal interests can overshadow other variables. Sencar and Eryilmaz (2004), for example, showed that apparent differences in Turkish girls' and boys' thinking about electric circuits disappeared when they controlled for students' interests and prior experiences related to the topic.

In summary, while we might expect that comparing across cities focuses on the degree to which a standardized system faithfully administered should lead to similar outcomes, comparing across majors focuses on patterns in individuals' interests, abilities, experiences, and affiliations as well as the later differences in the educational systems created by the majors after ninth grade (in terms of specific courses and cohorts). The comparison across cities therefore explores the degree to which standardized educational systems can reduce variation across student outcomes, whereas the comparison across majors explores the degree to which personal interests, abilities, experiences, affiliations, and some later differences in curriculum can increase variation across student outcomes within an overarching standardized educational system (cf. Phelan, Davidson, & Thanh Cao, 1991).

METHODS

We now describe the participants, interviewers, selection of cities, interview instruments, coding schemes, and analyses involved in the current study.

³While students' major is an independent variable in the analyses of all of the studies described in the following paragraphs, each study was also analyzed in terms of other independent variables depending on the focus of the study.

TABLE 1
Comparison of City 1 and City 2 in Terms of Their Characteristics

Characteristics	City 1	City 2
Location	Central Anatolia Region in Turkey	Southeastern Anatolian Region in Turkey
Population size	One of the largest cities in Turkey	One of the largest cities in Turkey but three times smaller than City 1
Population type	Cosmopolitan	Less cosmopolitan
Cultural structure	More western	More eastern
Employment	Heavy emphasis on government institutions, universities, and research institutes	Heavy emphasis on industrial and agricultural centers
Socioeconomic status	High	High
Language	Turkish	Turkish
Religion	Primarily Muslim: Highly secular	Primarily Muslim: Less highly secular
Average scores on the Higher Education Entry Examination in 2010	240 ranked sixth of 81 cities	229 ranked 52nd of 81 cities

Participants

This study involves 78 students from two cities in Turkey (32 from City 1 and 46 from City 2). The students were selected from four grade levels (pre-K, elementary school, middle school, and high school). The mean student ages were 5, 10, 13, and 16 years, respectively. Approximately half of the students were girls and half were boys. All of the schools were public and funded by the government. Approximately eight pre-K, eight elementary, and eight middle school students were interviewed in each city. In City 1, all pre-K, elementary, and middle school students were selected from one school at each grade level. In City 2, no more than three pre-K, elementary, and middle school students were selected from any one school. For high school, eight students were interviewed in City 1 from one school (of mixed majors) and 23 students were interviewed in the City 2 across multiple high schools (including eight science and mathematics, eight Turkish and mathematics, and seven social sciences students across the schools).

The students in City 1 are the students from Ozdemir and Clark (2009) and the students from City 2 are the Turkish students from Clark et al. (2011). The coders from Clark et al. (2011) recoded the Ozdemir and Clark (2009) student interviews to ensure that the interviews in both samples were coded by the same coders in the same way.

Description of Cities

The two cities were chosen based on access by the research team, but the two cities provide a representative level of distinctness to support a fair comparison of cultural and socioeconomic differences at the city level in terms of student outcomes across Turkey (Table 1).

In terms of cultural/resource comparisons, City 1 is among the very largest cities in Turkey in terms of population and is located in the Central Anatolia region in the middle of

Turkey. Most of the government institutions are located in City 1, and a large percentage of the population is employed in those state institutions. City 2 is located in the Southeastern Anatolia region of southeastern Turkey. City 2 is also among the largest cities in Turkey, but it is only one third the size of City 1 in terms of population. City 2 is an important industrial and agricultural center of Turkey. Compared to City 2, City 1 is a more cosmopolitan city with many universities, research institutes, foreign embassies, and top state institutions. City 2 by comparison is more entrepreneurial and industrial. Both cities were ranked by the State Planning Organization (SPO) as “high SES” (socioeconomic status), but the Central Anatolia region is “medium SES,” whereas Southeastern Anatolia is “low SES” as a region (SPO, 2003). Turkish is the primary language in both cities, and religious practices are relatively similar in both cities, although City 2 is somewhat more culturally eastern and less secular than City 1 as a function of its location as discussed by Dogan and Abd-El-Khalick (2008).

In terms of the national standardized test scores on the Higher Education Entry Examination at the end of high school, high school graduates in City 1 score somewhat higher on average than graduates from City 2. For the 2010 examination, for example, City 1 ranked sixth of the 81 ranked cities in Turkey with a score of 240, whereas City 2 ranked 52nd in terms of the average Higher Education Entry Examination scores with a mean score of 229. As discussed earlier, Berberoglu and Kalender (2005) found that (a) comparing students’ science and mathematics scores on the Higher Education Entry Examination across the seven regions did not manifest a strong main effect due to very low partial eta-squared values and (b) comparing across the 2003 PISA secondary data for eighth-, ninth-, and tenth grade students was similarly not meaningful in practice because of a low effect size. Thus, these two cities provide a representative level of distinctness in terms of socioeconomic and cultural variables to support the current study’s comparison of differences in student outcomes across cities in Turkey.

Interviews and Interviewers


Students were interviewed in terms of the same 10 sets of questions used by diSessa et al. (2004), Ozdemir and Clark (2009), Clark et al. (2011), and Price Schleigh et al. (2013) that diSessa et al. consolidated from Ioannides and Vosniadou (2002). As described earlier, each set involves two initial yes/no questions and one comparison question (Figures 1a and 1b). The initial questions directly ask whether or not there is a force on a specified stone in each of the pictures. The comparison question asks the student to compare the forces between the two pictures. The comparison question provided more information related to the student’s understanding of force in terms of relative strength and contextual differences.

The two interviewers were Turkish native speakers who were doctoral students in science education in the United States at the time of the interviews. All students were interviewed individually for about 20–25 minutes. Each student was asked all questions during one session. All interviews were videotaped, transcribed, and translated into English by the interviewers prior to coding.

Coding Schemes

The current study separately applied the coding schemes of both Ioannides and Vosniadou (2002) and diSessa et al. (2004) to each student’s responses as implemented in Clark et al. (2011) and Price Schleigh et al. (2013). Full details about the schemes and procedures are included in those papers.

TABLE 2
Rubric for Assigning Categories of Responses Based on the I&V Coding Scheme for Question Set 1

Set 1	Big Versus 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?
Response Categories	
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 ^a interpretation of the force of gravity	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 this category from the previous category of “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 it is an alternative conception (i.e., a non-normative or naïve conception) of gravity.

Ioannides and Vosniadou (2002) Coding Scheme. Ioannides and Vosniadou (2002) coded students first at the “question set” level and then at the “overall” level. We henceforth refer to their coding scheme as the “I&V coding scheme” for brevity. Ioannides and Vosniadou’s question set rubrics coded each set of questions using a set of response categories. Ozdemir and Clark (2009) and Clark et al. (2011) transferred Ioannides and Vosniadou’s coding rubrics to the revised organization of question sets consolidated by diSessa et al. (2004) using the same scoring categories employed by Ioannides and Vosniadou in their rubrics. Each question was first scored in terms of categories of answers specific to that question (which Ioannides and Vosniadou referred to as the “question set level”). This rubric for Question Set 1 is presented in Table 2. The seven categories of force meanings are outlined in Table 3.

After scoring all questions at the question set level for the student’s specific responses, Ioannides and Vosniadou used an “overall level” rubric for the question set to code the student’s responses in terms of potential matches with the seven force meaning categories (e.g., *internal*, *push-pull*, *gravity* and *other*). As with the question set level rubrics, we

TABLE 3**The Seven Categories of Force Meanings Outlined by Ioannides and Vosniadou (2002)**

-
1. *Internal force.* Students are coded for this meaning of force if they provide explanations 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 are coded for this meaning of force if they provide explanations indicating that force is due to size/weight of object, but also if moving objects and objects that are likely to fall have less internal force than stationary objects.
 3. *Internal and acquired.* Students are coded for this meaning of force if they indicate that there is a force on or in stationary objects due to size or weight and if they indicate that these objects acquire an additional force when they are set in motion. Ioannides and Vosniadou included students in this meaning who were ambivalent about unstable objects and interpreted unstable objects as either lacking internal force or as being likely to acquire additional force.
 4. *Acquired.* Students are coded for this meaning if their explanations indicate that force is a property of objects that explains motion and potentially acts on other objects. These students answer that there is no force on stationary objects and that the force on a moving object disappears when the object stops moving. Ioannides and Vosniadou also included students who explained that force is only acquired by heavy, moving objects. Ioannides and Vosniadou claim that this response indicates that these students relate the acquired force to both the weight and the motion of the object. In addition, Ioannides and Vosniadou included students in this meaning who thought that unstable stones have more force because they can be set in motion more easily as well as students who explain that all stones (stable and unstable) can be set in motion easily.
 5. *Acquired and force of push-pull.* Students are coded for this meaning if they provide explanations meeting the criteria described above for the acquired meaning of force, but also answered that there is a force on an object when acted on by an agent regardless of whether or not it moves.
 6. *Force of push-pull.* Students are coded for this meaning if they indicate that a force is exerted only on objects being pushed by an agent whether or not the object moves.
 7. *Force of gravity and others.* Students are coded for to this meaning if they mention gravity or gravity and other forces in their explanations. Ioannides and Vosniadou allowed for students to be considered consistent with this *gravity and others* meaning for Question Sets 7 and 8 even if the students do not mention the word gravity in these sets.
-

transferred Ioannides and Vosniadou's rubric categories and criteria to the question sets consolidated by diSessa et al. (2004). Table 4 presents the overall level scoring rubric for Question Set 1.

diSessa, Gillespie, and Esterly (2004) Coding Scheme. diSessa et al. (2004) were concerned about the reliability with which they could apply the I&V coding scheme to their interviews. diSessa et al. therefore adapted the I&V coding scheme into a "coarse quantitative" format. We henceforth refer to this coding scheme as the "DG&E coding scheme" for brevity. diSessa et al. designed their rubrics to code for potential force meanings more liberally than Ioannides and Vosniadou's rubrics, but did not formally test their rubrics against Ioannides and Vosniadou's rubrics. Rather than creating a rubric of categories of qualitative meanings of force for each question set, diSessa et al. (2004) instead developed a

TABLE 4
Rubric for Assigning Force Meanings at the Overall Level Based on the I&V Coding Scheme for Question Set 1

Meaning of Force	Internal	Internal/ Move	Internal/ Acquired	Acquired	Acquired/ Push-Pull	Push-Pull	Gravity and Other
Set 1: Big versus 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.	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. H: Force from the air.	G: No force on any stones because no one pushes them.	C: Force of gravity on both stones. D: Greater force of gravity on big stone.

“model mapping” rubric that includes all Ioannides and Vosniadou’s meanings and specific codes. To do this, diSessa et al. (2004) compared students’ responses to expected patterns for the seven force meanings outlined in Ioannides and Vosniadou (2002) at the “coarse quantitative” level by comparing combinations of the existence, absence, and relative sizes of forces on each object. They also included potential exemptions based on the inclusion of specific sources of force in students’ explanations. diSessa et al.’s rubric for Question Set 1 is presented in Table 5.

The current study uses diSessa et al.’s (2004) set of consolidated questions and therefore employs the same rubrics from diSessa and colleagues (2004). One point of clarification, however, involves the gravity and other category. diSessa et al. (2004) expressed specific concerns about the lack of specificity of the category in their study. In their study, a student was automatically precluded from being coded into any other meaning than gravity and other if the student mentioned gravity as a force on the object. Furthermore, students 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.

Students in the current study, Ozdemir and Clark (2009), Clark et al. (2011), and Price Schleigh et al. (2013), often outlined multiple independent force explanations within a single question set (as was also noted by diSessa et al.). Clark et al. (2011), Price Schleigh et al. (2013), and the current study therefore modified the DG&E coding scheme to code each independent force separately if there were multiple forces. Separate meanings within a question set that were not specifically connected to gravity in the student’s explanations were coded for those meanings. In addition, we did not assign a student into the gravity and other category for the DG&E coding scheme unless the student explicitly referred to gravity or an attractive force from the earth. Thus, while the application of the

TABLE 5**Rubric for Assigning Force Meanings From the DG&E Coding Scheme for Question Set 1**

Meaning of Force	Internal	Internal/ Move	Internal/ Acquired	Acquired	Acquired/ Push-Pull	Push-Pull	Gravity and Other
Set 1: Big versus 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.	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.		Force only on the small stone but not due to gravity		Force on both stones but greater force on the big stone.

DG&E coding scheme in the current study was generally symmetrical with its application in diSessa et al. (2004), we did adjust the coding scheme to increase the specificity of the gravity and other category in alignment with the concerns expressed by diSessa et al. (2004), Ozdemir and Clark (2009), and Clark et al. (2011). This adjustment mitigates the possibility of the gravity and other category masking expression of the other force meaning categories. This adjustment does not, however, adjust for a limitation discussed further in the Results and Limitations sections regarding the potential of the gravity and other category to mask fragmentation within a student's explanations across question contexts if a student expressed multiple meanings across contexts that fell within the hybrid structure of the gravity and other category. The purpose of the current study involved replicating the prior series of studies, and thus new categories were not introduced, but this caveat needs to be considered when interpreting the results of the current study.

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 that corresponded to each of the seven force meanings using each coding scheme. This corresponds to a total number of 140 cells for each student (i.e., 10 question sets multiplied by seven force meaning categories multiplied by two coding schemes). Any differences were discussed. Final codes were tabulated across the question sets to determine how many times each student matched each force meaning category according to each scheme. The overall interrater reliability between the two coders before discussion was 93%, calculated by the percentage of matched cells in the coding schemes.

Each student's consistency across the 10 question sets was then checked for each of the seven possible force meanings. Students were first checked for the meanings that they applied consistently across all 10 question sets. If a student matched for a force meaning

TABLE 6
Summary of Dependent Variables Used in Analyses

<i>Consistent-With-Allowance Code:</i> Consistent with allowance is a measure of consistency specifying whether or not a student matched for the same force meaning on at least 8 of the 10 questions sets (e.g., if a student matches for 8, 9, or 10 of the question sets with the same force meaning, that student was coded as “consistent with allowance”). Note that the specific force meaning does not matter—only that the student matched the same meaning for 8, 9, or 10 of the question sets matters.
<i>Best-Match Score:</i> Best-match score is a measure of consistency that specifies the number of question sets the student matched for her/his best-match meaning out of the ten possible question sets (e.g., “7”).
<i>Best-Match Meaning:</i> A measure of force meanings expressed. Best-match meaning specifies the force meaning category for which student matched for the most question sets (e.g., “push-pull” or “gravity and other”). Note that the best-match meaning for a given student may have a best-match score that is lower than the threshold for being coded as consistent with allowance.
<i>Force Meaning Scores:</i> A measure of force meanings expressed. The force meaning scores for a student specify the numbers of question sets (between 0 and 10) for which a student matched for each of the seven force meanings (e.g., “7, 8, 4, 2, 2, 1, 6”). This array of seven numbers therefore represents the student’s number of matches for each of the seven force meanings and is used for analyses comparing students’ matches across the seven meanings.

across all 10 question sets, the student was classified as consistent for that force meaning. Students were classified as “consistent with allowance” if the student matched for at least eight of 10 question sets.

Analyses of Consistency and Force Meanings

The current study focuses on the analysis of four dependent variables. This section outlines the analyses and rationales for the analyses for each dependent variable. Table 6 summarizes the dependent variables involved in each of these analyses.

Consistency: Two-Way Contingency Table Analyses of Consistency Codes. The analysis in terms of consistency codes focuses on the consistency with which an individual student expressed the same force meaning across question sets. Ioannides and Vosniadou (2002) argued that consistency of meanings applied across contexts (i.e., question sets) provided strong evidence of coherent theory-like understandings. For this reason, Ioannides and Vosniadou (2002), diSessa et al. (2004), Ozdemir and Clark (2009), Clark et al. (2011), and Price Schleigh et al. (2013) focused primarily on levels of students’ consistency in their explanations across question contexts. While Ioannides and Vosniadou focused on full consistency (where a student answered all questions using the same force meaning), diSessa et al. added a second category for students who matched at least eight of 10 question sets for a single force meaning to account for any anomalies in their interview or coding process. We refer to these respectively as “fully consistent” and “consistent-with-allowance” codes.

Approximately, 10% of students in the current study overall are coded as “fully consistent” and approximately 55% are coded as “consistent with allowance” according to the DG&E coding scheme as well as the I&V coding scheme. The analyses comparing across cities, grade level, gender, and academic major in the current study are therefore based

on the “consistent-with-allowance” codes because too few students are coded as “fully consistent” in some cells to meet statistical assumptions for two-way contingency table analyses. The analyses comparing across cities, grade level, gender, and academic major were run separately for each coding scheme.

Consistency: One-Way Analyses of Variance of Best-Match Scores. We also analyzed students’ consistency using the best-match scores underlying the “fully consistent” and “consistent-with-allowance” codes. A student’s best-match score is the number of question sets for which the student matched for his or her best-match meaning (which is the force meaning for which the student matched on the most question sets). We conducted one-way analyses of variance (ANOVAs) to evaluate whether statistical relationships exist between students’ best-match scores and (1) cities, (2) grade levels, (3) gender, and (4) academic major. All analyses were run separately for each coding scheme.

Force Meanings: Two-Way Contingency Table Analyses of Best-Match Meanings. In terms of the force meanings most frequently expressed by a student, we focused on students’ best-match meanings (i.e., internal, internal/ move, internal/acquired, acquired, acquired/push-pull, push-pull, or gravity and other). A student’s best-match meaning is the force meaning expressed across the largest number of question sets. We conducted two-way contingency table analyses to evaluate whether statistical relationships exist between best-match meanings and (1) cities, (2) high school majors, (3) grade level, and (4) gender. All analyses were run separately for each coding scheme.

Force Meanings: One-Way Multivariate Analyses of Variance of Force Meaning Scores. Finally, we examined patterns in the force meaning scores for each student. The force meaning scores for a student are the numbers of question sets (between 0 and 10) for which a student matched for each of the seven force meanings expressed as an array (e.g., “7, 8, 4, 2, 2, 1, 6”). This array of seven numbers thereby represents the student’s number of matches for each of the seven force meanings. Thus, this is the array of numbers that underlies the determination of best-match meanings. One-way multivariate analyses of variance (MANOVAs) were conducted with seven dependent variables of students’ force meaning scores. All analyses were run separately for each coding scheme.

RESULTS AND DISCUSSION

We now present and discuss the results from the four analyses described above for each of the four main comparisons: (1) city, (2) grade level, (3) gender, and (4) academic major. Follow-up tests are reported for each analysis as appropriate. Please refer to the Methods section and Table 6 for overviews of the analyses and dependent variables.

Results for Comparisons Across Cities

We first present the focal comparisons across cities in terms of the consistency of students’ explanations and the specific meanings of force expressed in those explanations.

Consistency: Consistent-with-Allowance Codes Across Cities. Contingency table analyses showed no significant differences overall across cities in terms of consistent-with-allowance codes based on either coding scheme. At a finer grain size, we conducted

a multinomial logistic regression to evaluate the effect of city on the consistent-with-allowance codes at the elementary, middle, and high school levels. Pre-K students could not be included in these analyses due to small cell size. The likelihood ratio tests with chi-squared statistics indicate that the contribution of independent variables (*grade level* and *city*) does not contribute to the model for the interaction of city with these other three grade levels. In addition, the goodness-of-fit indices show the interaction of the independent variable (i.e., city) does not affect the *consistent-with-allowance scores* in terms of city and grade level for either coding scheme.

Consistency: Best-Match Scores Across Cities. One-way ANOVAs showed no significant differences on best-match score across cities based on either coding scheme. At a finer grain size, a 4×2 ANOVA was conducted for each coding scheme to evaluate differences between the two cities by grade level in terms of best-match score. All four grade levels were included in this analysis. There was no significant interaction between city and grade level based on either coding scheme.

Meanings: Best-Match Meanings Across Cities. Two-way contingency table analyses showed no significant differences across cities in terms of students' best-match meanings for either coding scheme.

Meanings: Force Meaning Scores Across Cities. One-way MANOVA indicated no significant differences between students from City 1 and City 2 in terms of force meaning scores based on either coding scheme.

City Summary. There are no significant differences between cities overall or in interaction with grade level in terms of consistency of explanations or proportions of force meanings expressed based on either coding scheme.

Results for Comparisons Across Grade Level

As discussed, the grade-level comparisons were conducted to provide a baseline for the upper threshold of variation based on the high degree of variation observed across grade levels in previous studies.

Consistency: Consistent-With-Allowance Codes Across Grade Levels. The students from the two cities were combined for these analyses to increase power for the comparisons. It should be noted, however, that this still resulted in smaller sample sizes in each group than the sample sizes in the city comparisons because there are four grade levels as opposed to only two cities. Two-way contingency table analyses showed a significant relationship between students' grade level and consistent-with-allowance codes for the I&V coding scheme, Pearson $\chi^2(3, N = 78) = 10.67, p = .01$, Cramer's $V = 0.37$. The percentages of pre-K, elementary, middle, and high school students coded as consistent with allowance were 27%, 47%, 67%, and 74%, respectively, for the I&V scheme. A similar, though nonsignificant, pattern was observed between grade levels and consistent-with-allowance codes for the DG&E scheme (with proportions of 27%, 59%, 53%, and 61%, respectively). Follow-up pairwise comparisons were conducted based on the I&V scheme. Table 7 presents the results for these analyses. The Holm's sequential Bonferroni method was used to control

TABLE 7**Results for the Pairwise Comparisons of Consistency Codes Across Grade Levels by Using the Holm's Sequential Bonferroni Method**

Comparison	Pearson Chi-Square	<i>p</i> (Alpha)	Cramer's <i>V</i>
Pre-K versus high school	9.42 ^a	.002 (.008)	0.45
Pre-K versus middle school	4.82	.028 (.001)	0.40
Elementary school versus high school	3.53	.060 (.012)	0.27
Pre-K versus elementary school	1.41	.230 (.017)	0.21
Elementary school versus middle school	1.25	.265 (.025)	0.20
Middle school versus high school	0.28	.595 (.050)	0.08

^a $p \leq \alpha$.**TABLE 8****Means and Standard Deviations for the Force Meaning Scores Based on the I&V Coding Scheme Across Grade Levels**

Grade Level	<i>M</i>	<i>SD</i>
Pre-K	6.80	1.74
Elementary	7.47	1.62
Middle	7.80	1.42
High	8.39	1.54

for Type 1 error at the .05 level across all six comparisons. The only pairwise difference that was significant was between the pre-K and high school students. The probability of a high school student being coded as "consistent with allowance" was about 2.74 times (74%/27%) more likely compared to a pre-K student.

Consistency: Best-Match Scores Across Grade Level. The one-way ANOVA was significant for the I&V coding scheme, $F(3, 74) = 3.67$, $p = .016$. The η^2 of 0.13 indicates a medium to large effect size regarding the relationship between students' best-match scores across grade levels. A similar, though nonsignificant, pattern was observed between grade levels and best-match scores for the DG&E scheme.

Follow-up pairwise comparisons were conducted. The Holm's sequential Bonferroni method was used to control Type 1 error at the .05 level across all six comparisons. The only significant comparison was between pre-K and high school students' best-match scores. Table 8 shows means and standard deviations for students' best-match scores across grade levels.

Meanings: Best-Match Meanings Across Grade Level. The two-way contingency table analyses demonstrated a significant relationship between best-match meanings and grade level based on both the DG&E and I&V schemes, Pearson $\chi^2 (18, N = 98) = 46.081$, $p = .00$, Cramer's $V = 0.40$ and Pearson $\chi^2 (18, N = 89) = 51.44$, $p = .00$, Cramer's $V = 0.44$, respectively. Figures 2a and 2b present the patterns of best-match meanings across grade levels.

Follow-up pairwise comparisons were conducted. Table 9 presents the results of these analyses based on both coding schemes. The Holm's sequential Bonferroni method was used to control Type 1 error at the .05 level across all six comparisons. The pairwise comparisons

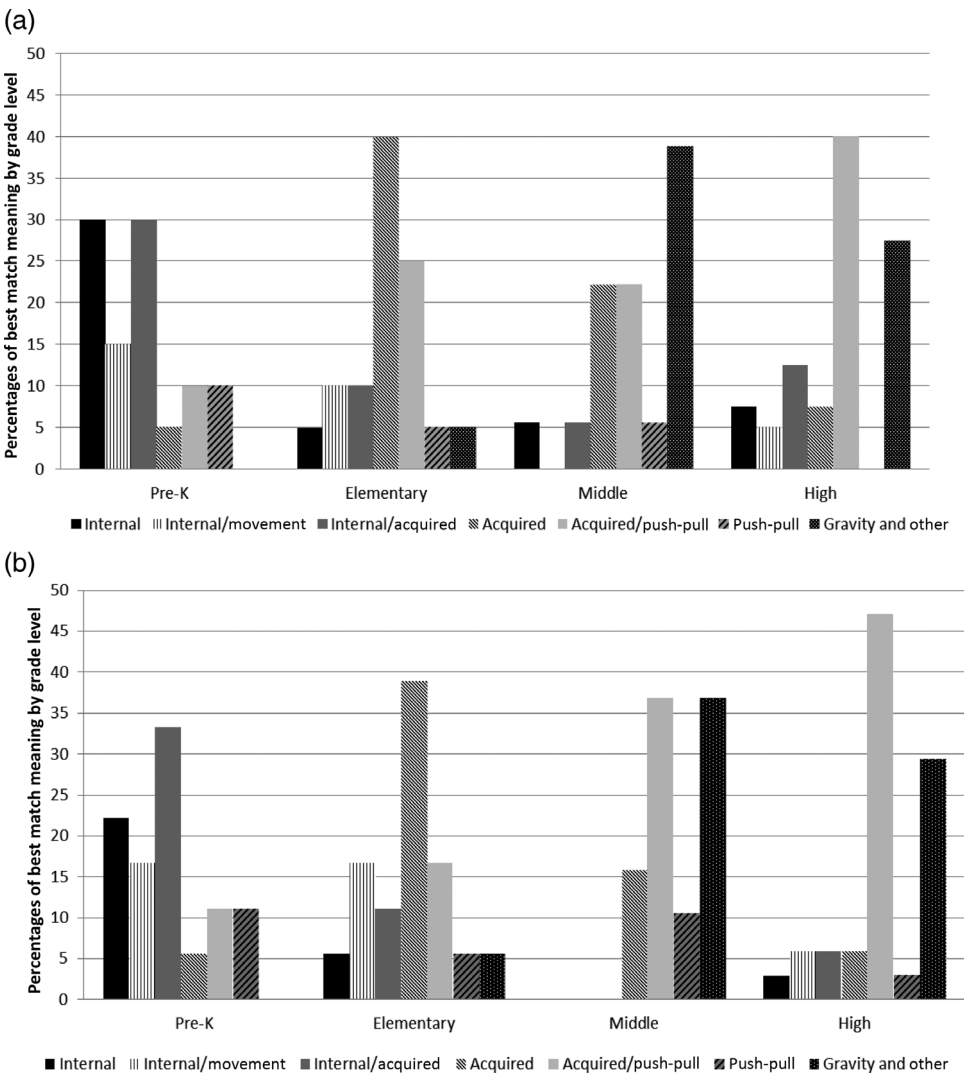


Figure 2. Best-match meanings by grade levels based on (a) the DG&E coding scheme and (b) the I&V coding scheme.

between pre-K and high school students and pre-K and middle school students were significant for both coding schemes. Essentially, as in earlier studies, pre-K students tend to express internal-related best-match meanings whereas middle school students and high school students tend to express more acquired/push-pull and gravity and other meanings. While the differences between the elementary students and the other students were not significant in these analyses, elementary students primarily expressed the acquired meaning.

Meanings: Force Meaning Scores Across Grade Level. The pre-K students' force meaning scores are highest on average for internal-related force meanings. The elementary, middle, and high school students' force meaning scores are highest on average for acquired-related meanings. The force meaning scores for gravity and other are higher for middle and high school students than the elementary school students' score. Figures 3a and 3b show

TABLE 9**Results for the Pairwise Comparisons of Best-Match Meanings Across Grade Levels by Using the Holm's Sequential Bonferroni Method**

Scores Based on Comparison	DG&E Coding Scheme			I&V Coding Scheme		
	Pearson Chi-Square	<i>p</i> (Alpha)	Cramer's <i>V</i>	Pearson Chi-Square	<i>p</i> (Alpha)	Cramer's <i>V</i>
Pre-K versus high school	21.95 ^a	.001 (.008)	0.61	22.79 ^a	.001 (.008)	0.66
Pre-K versus middle school	19.89 ^a	.003 (.01)	0.72	23.77 ^a	.001 (.01)	0.80
Elementary school versus high school	14.61	.024 (.012)	0.49	15.81	.015 (.012)	0.55
Pre-K versus elementary school	13.84	.032 (.017)	0.59	9.83	.13 (.017)	0.52
Elementary school versus middle school	8.16	.224 (.025)	0.46	14.08	.029 (.025)	0.62
Middle school versus high school	7.66	.264 (.050)	0.36	5.80	.45 (.050)	0.33

^a $p \leq \alpha$.

these relationships. One-way MANOVAs demonstrated significant differences among the grade levels on force meaning scores across the seven force meaning categories for both coding schemes [Wilks' lambda $\Lambda = 0.36$, $F(21, 195) = 3.90$, $p < .01$ for the DG&E coding scheme and Wilks' lambda $\Lambda = 0.32$, $F(21, 195) = 3.90$, $p < .01$ for the I&V coding scheme]. The multivariate η^2 based on Wilks' lambda were strong, at 0.29 and 0.32, respectively.

ANOVA on the dependent variables were conducted as follow-up tests to the MANOVA. To control Type I error, the Dunnett's *C* method was employed. For the DG&E scheme, the ANOVAs on the acquired, acquired/push-pull, and gravity and other force meaning scores were significant. For the I&V scheme, the ANOVAs on the internal/movement, acquired, acquired/push-pull, and gravity and other force meaning scores were significant.

Using the DG&E scheme, post hoc analyses to the univariate ANOVA for the acquired, acquired/push-pull, and gravity and other force meaning scores consisted of conducting pairwise comparisons to the grade levels for which these force meaning scores were significantly different. The pre-K students have significantly lower acquired/push-pull scores on average than the elementary school, middle school, and high school students. The elementary school students have higher acquired scores on average than the pre-K and high school students. Finally, high and middle school students have significantly higher gravity and other scores than the elementary and pre-K students, but the middle and high school students are not significantly different from each other in terms of gravity and other scores.

Using the I&V scheme, the pre-K students have significantly higher internal/movement scores on average than the middle and high school students. The pre-K students have significantly lower acquired/push-pull scores on average than the elementary school, middle school, and high school students. The elementary school students have higher acquired scores on average than the pre-K students. Finally, high school and middle school students have significantly higher gravity and other scores than the elementary and pre-K students,

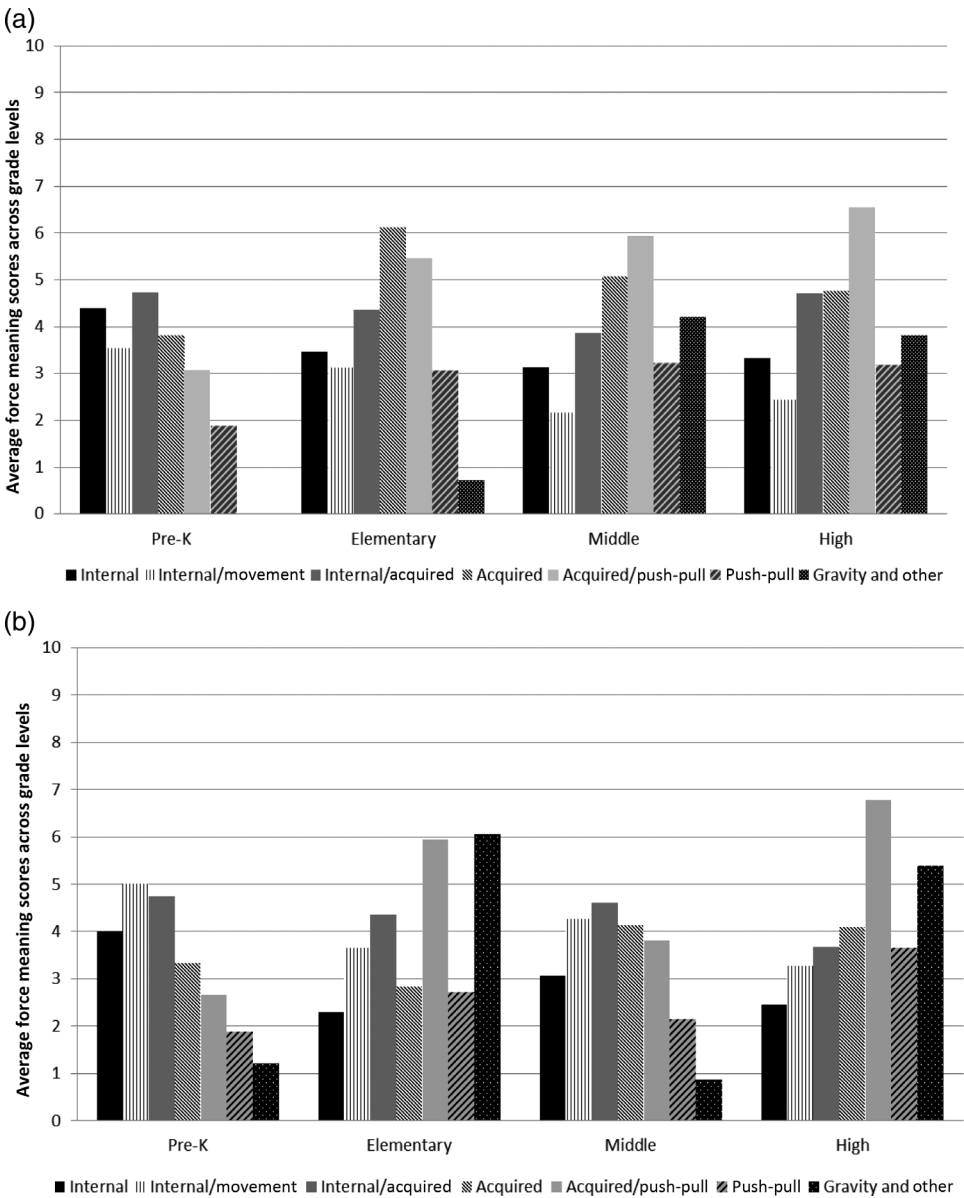


Figure 3. Average force meaning scores across grade levels based on (a) the DG&E coding scheme and (b) the I&V coding scheme.

but the middle school and high school students are not significantly different from each other in terms of gravity and other scores.

Grade-Level Summary. There are significant differences across grade levels in terms of both consistency and force meanings. In terms of consistency, there is a significant difference across grade levels with high school students demonstrating a higher probability of being consistent with allowance and having higher best-match scores than pre-K students according to the I&V scheme. Similar nonsignificant patterns are observed for the DG&E

scheme. That said, as discussed later in the Limitations section, the gravity and other category (common to many older students) likely hides a fair amount of fragmentation within itself even though the current study enhanced the specificity of the category as discussed in the Methods and Limitations sections. The findings related to the consistency of the high school students and to some degree the middle school students should therefore be considered with that caveat.

In terms of force meanings, grade level is significantly related to best-match meanings and force meaning scores based on both coding schemes following patterns observed in previous studies. Essentially, pre-K students tend to express internal-related meanings, elementary students primarily express the acquired meaning, and middle and high school students tend to express acquired/push-pull and gravity and other meanings.

Results for Comparisons Across Gender

As discussed, the gender comparisons are intended to provide a baseline for the lower threshold of variation because the prior studies in the series have demonstrated no significant differences across gender. It should be noted that the analyses across gender involve larger sample sizes than the grade-level comparisons because there are only two cells (i.e., male and female) whereas the grade-level analyses involve four cells (i.e., pre-K, elementary, middle school, and high school).

Consistency: Consistent-With-Allowance Codes Across Gender. The two-way contingency table analyses demonstrated that gender and consistent-with-allowance codes are not significantly related for either coding scheme.

Consistency: Best-Match Scores Across Gender. The one-way ANOVAs demonstrated that gender and best match scores are not significantly related for either coding scheme.

Meanings: Best-Match Meanings Across Gender. The two-way contingency table analyses demonstrated that gender and best-match meanings are not significantly related for either coding scheme.

Meanings: Force Meaning Scores Across Cities. The one-way MANOVAs demonstrated that gender and force meaning scores are not significantly related for either coding scheme.

Gender Summary. As seen across the prior studies, this study demonstrates no significant differences by gender in terms of either consistency or force meanings.

Results for Comparisons Across Academic Major

Finally, the comparison across academic major for the high school students in one of the cities provides a third baseline in terms of the variation resulting from more personal variables (e.g., individual interests, abilities, experiences, and affiliations).

Consistency: Consistent-With-Allowance Codes Across Major. There are pronounced differences across high school majors in terms of consistent-with-allowance codes

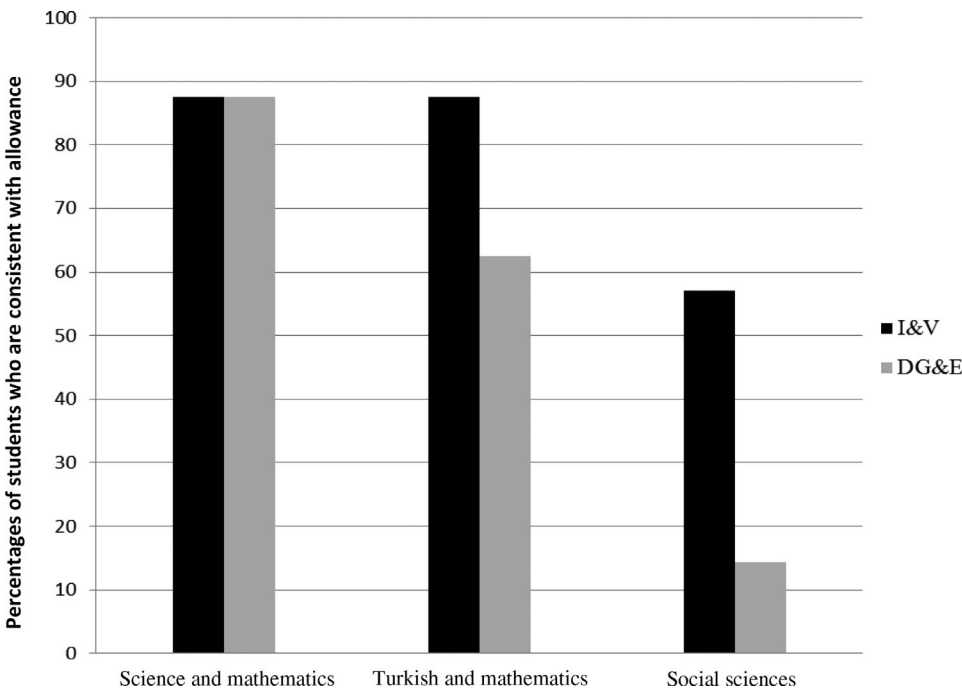


Figure 4. Percentages of students across high school majors coding as consistent using with “consistent-with-allowance” criterion for I&V and DG&E schemes.

TABLE 10
Results for the Pairwise Comparisons of Consistency Codes Across High School Majors by Using the Holm’s Sequential Bonferroni Method

Comparison	Pearson Chi-Square	<i>p</i> (Alpha)	Cramer’s <i>V</i>
Science and mathematics versus social science	8.04 ^a	.005 (.017)	0.73
Science and mathematics versus Turkish and mathematics	1.33	.248 (.025)	0.29
Turkish and mathematics versus social science	3.62	.057 (.050)	0.49

^a*p* ≤ alpha.

(Figure 4). Based on the DG&E coding scheme, only 14% of social sciences students are “consistent with allowance” although 88% of science and mathematics students and 63% of Turkish and mathematics students are “consistent with allowance.” Similarly, based on the I&V scheme, only 57% of social sciences students are “consistent with allowance,” whereas nearly all (88%) of the students in the two other majors are “consistent with allowance” based on the I&V scheme. Two-way contingency table analyses demonstrate that majors and consistency are significantly related for the DG&E coding scheme, Pearson χ^2 (2, *N* = 23) = 8.32, *p* = .02, Cramer’s *V* = 0.60. Analyses for the I&V scheme followed a similar though nonsignificant pattern.

Follow-up pairwise comparisons were conducted with “consistent-with-allowance” codes based on the DG&E coding scheme. Table 10 shows the results for these analyses. The Holm’s sequential Bonferroni method was used to control for Type 1 error at

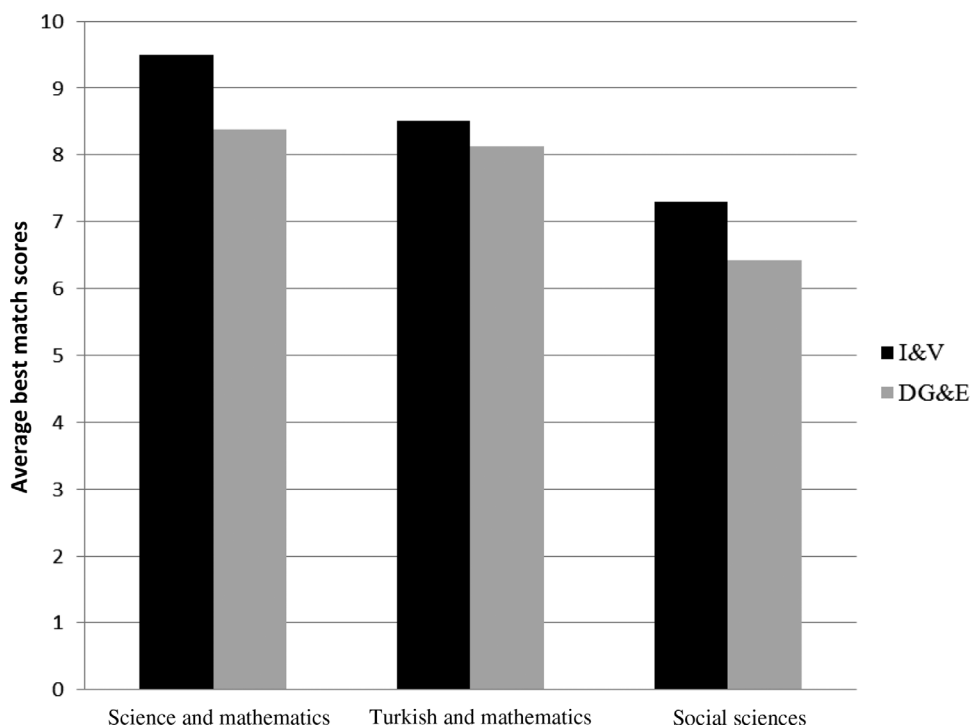


Figure 5. Average best-match scores based on both I&V's and DG&E's coding schemes across high school majors.

the .05 level across all three comparisons. The only pairwise difference that is significant is between the science and mathematics and social sciences majors. The probability of a student being "consistent with allowance" is about 6.2 times ($0.88/0.14$) more likely when the student is in the science and mathematics major than in the social sciences major.

Consistency: Best-Match Scores Across Major. The one-way ANOVAs showed differences in best-match scores across majors on both the DG&E and the I&V schemes, $F(2, 20) = 5.05$, $p = .017$, $F(2, 20) = 4.64$, $p = .022$, respectively. The η^2 of 0.34 and 0.32 indicated strong relationships between students' best-match scores and majors. Figure 5 presents the patterns in best-match scores across majors.

Follow-up pairwise comparisons were conducted. The Holm's sequential Bonferroni method was used to control Type 1 error at the .05 level across all three comparisons. There are significant differences between the social sciences students and the other two majors but not between science and mathematics students and Turkish and mathematics students. Table 11 shows means and standard deviations for students' best-match scores for each coding scheme.

Meanings: Best-Match Meanings Across Major. Two-way contingency table analyses found no significant relationships across majors and best-match meanings for either coding scheme, but science and mathematics students most frequently express gravity and other as their best-match meaning, whereas the Turkish and mathematics and social sciences students most frequently express acquired/push-pull as their best-match meaning (Figures 6a and 6b).

TABLE 11
Means and Standard Deviations for the Best-Match Scores Based on Both DG&E and I&V Coding Schemes Across High School Tracks

High School Tracks	DG&E's Coding Scheme		I&V's Coding Scheme	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Science and mathematics	8.38	0.92	9.50	1.07
Turkish and mathematics	8.13	1.46	8.50	1.07
Social sciences	6.43	1.40	7.29	1.98

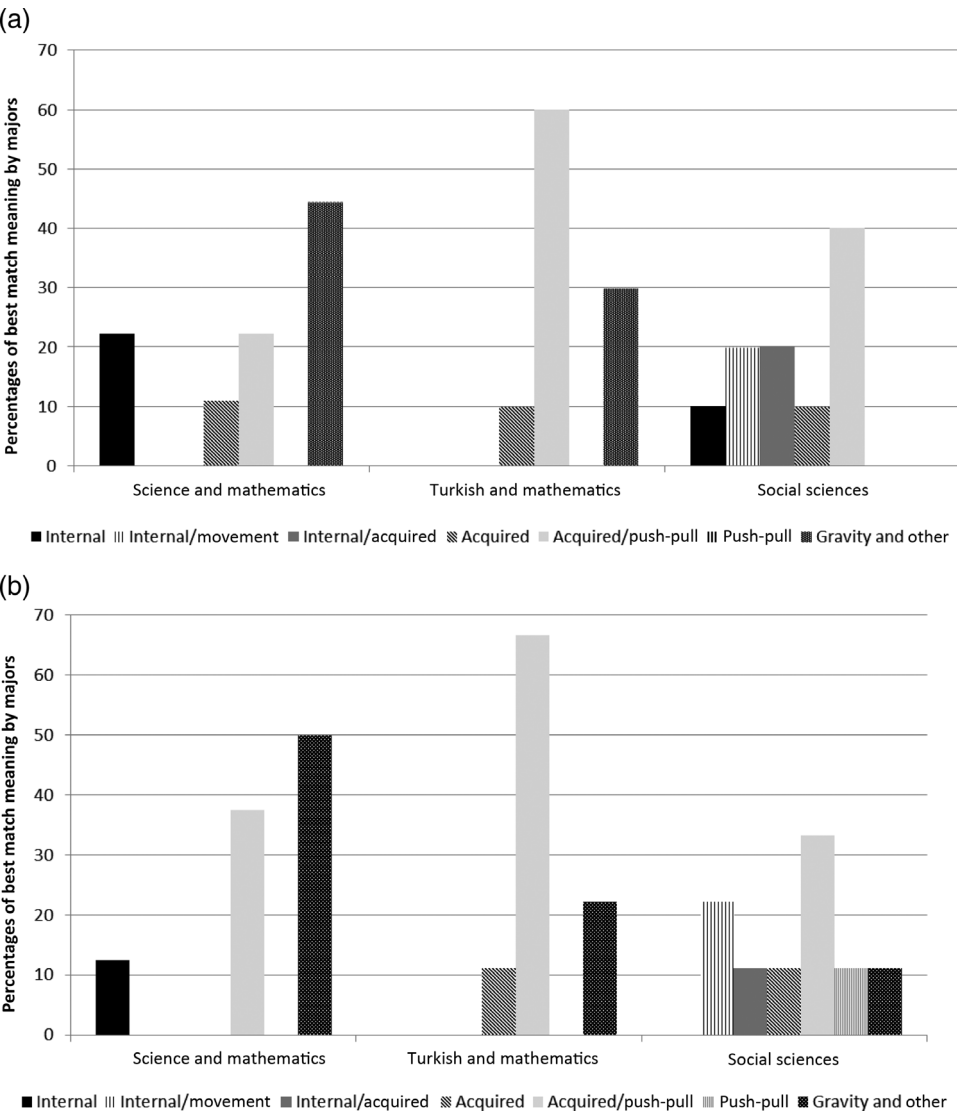


Figure 6. Best-match meanings by majors based on (a) the DG&E coding scheme and (b) the I&V coding scheme.

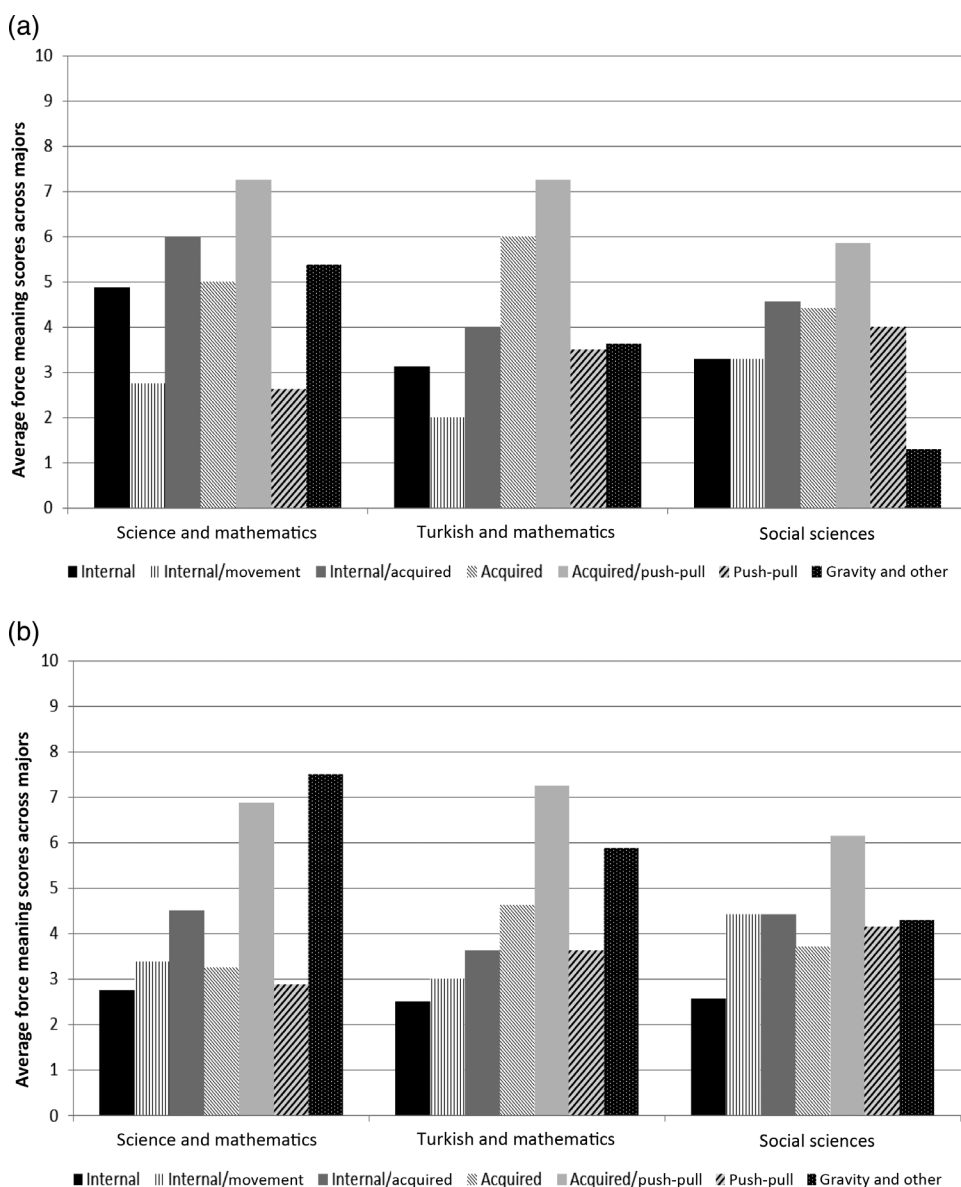


Figure 7. Average force meaning scores across high school majors based on (a) the DG&E coding scheme and (b) the I&V coding scheme.

Meanings: Force Meaning Scores Across Major. The one-way MANOVA demonstrated significant differences among the students from different majors on force meaning scores across the force meaning categories based on the DG&E scheme, Wilks' lambda $\Lambda = 0.16$, $F(14, 28) = 2.97$, $p < .01$. The multivariate η^2 based on Wilks' lambda was quite strong, 0.59. The multivariate partial eta-squared indicates 59% of multivariate variance of the dependent variables is associated with the major factor. Patterns were similar but not significant based on the I&V scheme. Figures 7a and 7b present force meaning scores by major for both schemes. Table 12 organizes the means and standard deviations on the dependent variables by high school majors based on the DG&E coding scheme.

TABLE 12
Means and Standard Deviations for the Force Meanings Scores by High School Majors

Force Meanings	Mathematics and Science		Mathematics and Literature		Social Sciences	
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>
Internal	4.88	2.29	3.13	1.55	3.29	2.06
Internal/movement	2.75	1.65	2.00	1.07	3.29	2.14
Internal/acquired	6.00	1.07	4.00	1.77	4.57	1.72
Acquired	5.00	1.51	6.00	2.27	4.43	0.98
Acquired/push-pull	7.25	1.16	7.25	1.67	5.86	1.86
Push-pull	2.63	0.52	3.50	1.77	4.00	0.82
Gravity and other	5.37	4.14	3.63	3.50	1.29	1.25

Separate ANOVAs were conducted on the dependent variables as follow-up tests to the MANOVA. To control for Type I error across multiple ANOVAs, Holm’s sequential Bonferroni method was used. The ANOVA analyses on internal/acquired, push-pull, and gravity and other categories were significant. Post hoc analyses to the univariate ANOVAs for the scores from these three categories were conducted to find which majors differed significantly in terms of force meaning scores. The science and mathematics students demonstrate significantly higher gravity and other scores than the social sciences students, $F(1, 20) = 5.80, p = .03$. Conversely, social sciences students demonstrate significantly higher push-pull scores than the science and mathematics students, $F(1, 20) = 5.06, p = .04$. Finally, the science and mathematics students have higher internal/acquired scores than the Turkish and mathematics students, $F(1, 20) = 6.71, p = .02$.

Academic Major Summary. In terms of consistency, there are pronounced differences across high school majors in terms of consistent-with-allowance codes and best-match scores. Essentially, students in the social sciences majors are less likely to be coded as consistent with allowance than students in the science and mathematics and Turkish and mathematics majors. This pattern is significant for the DG&E scheme and follows a similar nonsignificant pattern for the I&V scheme. In terms of best-match scores, there are significant and strong differences for both coding schemes with social sciences students having significantly lower scores than either Science and mathematics or Turkish and mathematics students. Whereas the caveat about the hybrid nature of the gravity and other category potentially concealing fragmentation applies in terms of the science and mathematics students in comparison to the social sciences students, the caveat seems less applicable to the comparison between the Turkish and mathematics and social sciences students because the students in both of these majors most frequently expressed acquired/push-pull as their best-match meaning.

In terms of force meanings, no significant relationships were found across majors in terms of best-match meanings based on either coding scheme, but science and mathematics students predominantly express gravity and other as their best-match meaning while Turkish and mathematics and social sciences students predominantly express acquired/push-pull as their best-match meaning. There are significant differences across majors, however, on force meaning scores based on the DG&E coding scheme. The science and mathematics students have significantly higher gravity and other scores than the social sciences students. Conversely, social sciences students have significantly higher push-pull scores than the

science and mathematics students. Finally, the science and mathematics students have higher internal/acquired scores than the Turkish and mathematics students.

LIMITATIONS OF THE CURRENT STUDY

The limitations of the current study are similar to those discussed by Ioannides and Vosniadou (2002), diSessa et al. (2004), Ozdemir and Clark (2009), Clark et al. (2011), and Price Schleigh et al. (2013). Essentially, our replication of the methodologies of the previous studies facilitates comparisons but also inherits limitations in terms of the sample size, the framing of the questions, and the nature of the gravity and other category.

Number of Students

Individually interviewing students requires substantial resources for data collection. Thus data collection can limit the number of students involved in conceptual change studies involving interviews, which in turn impacts claims of generalizability of the findings. Ioannides and Vosniadou (2002) analyzed interviews for 105 students. diSessa et al. (2004) analyzed interviews for 30 students. Ozdemir and Clark (2009) analyzed interviews for 32 students. Clark et al. (2011) analyzed interviews for 201 students. The current study analyzed interviews for 78 students.

Unfortunately, none of these studies was large enough to involve truly nationally representative samples. In addition, the sampling methods involved cluster sampling in which groups/clusters (schools in this case) were randomly selected rather than individual students from the population since it was often impossible to reach individual students from different schools.

Larger numbers of students from wider populations of schools and geographic areas would strengthen claims and generalizability. We would like to draw from a broader population of students in future work. Toward this goal, we are exploring the potential of a written instrument to collect similar data to facilitate data collection across larger groups of students (Price Schleigh et al., 2013). By increasing sample size in terms of specific variables of interest, we could more carefully explore interactions between factors, such as potential interactions between city and grade level in the current study.

That said, the current study has a sufficient sample size to address the questions raised. The sample size of 78 provides adequate power (assuming a medium effect size) for the analyses regarding the differences between two cities, the differences between male and female students, and the differences between grade levels. The sample size of 23 students in three groups for the academic major comparison has sufficient power assuming a very large effect size.

More specifically, we conducted statistical power analyses by using *G*Power 3* statistical software (Erdfelder, Faul, & Buchner, 1996). The power analysis for the one-way ANOVA tests revealed the required sample size of 52 for two groups, 66 for three groups, and 76 for four groups by using the significance criterion of 0.05, an effect size of 0.40, and a power of 0.80. To have power of 0.80 at the .05 alpha level for the academic major comparison of 23 students in three groups, the effect size would need to be 0.70 (very large).

Similarly, the power analysis for the MANOVA tests revealed that with seven dependent variables (seven force meanings), the required sample size is 54 for two groups, 69 for three groups, and 80 for four groups by using the significance criterion of 0.05, an effect size of 0.30 and a power of 0.80. To have power of 0.80 at the .05 alpha level for the academic major comparison of 23 students in three groups, the effect size would need to be 0.71 (very large).

Finally, for the contingency table analyses, a power analysis revealed sample sizes of 61 for the city and gender comparisons (i.e., two degrees of freedom) and 75 for the grade-level comparisons (i.e., four degrees of freedom) with significance criterion of 0.05, an effect size of 0.40, and a power of 0.80. To have power of 0.80 at the .05 alpha level for the academic major comparison of 23 students (i.e., three degrees of freedom), the effect size would need to be 0.67 (very large).

Thus, we have sufficient power for our comparisons of city, gender, and grade level assuming medium effect sizes (and we have more power for the city and gender comparisons than the grade-level comparisons because the city and gender comparisons involve only two groups, whereas the grade-level comparisons involve four groups). For the academic major comparisons, we have sufficient power to detect differences involving very large effect sizes. The current study detected significant differences in terms of grade level and academic major (where we had the least power) but not for gender or city (where we had the most power). The findings of the current study thus suggest that differences across city and gender are of much lower magnitude than those observed across grade level and academic major.

Framing of the Questions

This study adopted the questions from the earlier studies for the purpose of comparison and replication. As discussed in Clark et al. (2011), however, the framing and phrasing of the questions focus not only on the student's thoughts about the underlying physical mechanisms but also on the student's understanding of the word for "force." 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 ask students to predict and explain what will happen next in an everyday context (e.g., "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?"). This approach would allow the interviewer to investigate how students think and provide explanations about the mechanisms involved without requiring the interviewer to introduce terminology. The current format of questions has facilitated this ongoing set of studies, but this alternative approach seems to offer additional affordances, particularly in combination with diSessa et al. (2004)'s "plausible set of requirements for specifying important aspects of the content of a concept that is a physical quantity, such as force" (p. 854) that extend beyond existential and coarse quantitative aspects.

Problematic Issues With the Gravity and Other Meaning

As discussed by diSessa et al. (2004), Ozdemir and Clark (2009), and Clark et al. (2011), the gravity and other category is a hybrid category that can include many other ideas in addition to gravity. Ioannides and Vosniadou (2002) had originally anticipated a strict gravity meaning (at least among their high school students), but no students in their cohort (including their high school students) consistently assigned only gravitational forces in the question sets. Ioannides and Vosniadou (2002) 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.

The current study increased the specificity of the coding for the gravity and other category to account for this issue and related issues as described in detail in the Methods section. Further work and specification of gravity and other into subcategories representing specific meanings, however, 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. One specific limitation of the hybrid nature of this category is that it could mask fragmentation within a student's explanations across contexts if a student expressed multiple meanings across question contexts that still fell within the hybrid structure of the gravity and other category. Thus comparing levels of consistency between groups may artificially attribute higher levels of consistency to the group whose explanations were more frequently coded as gravity and other (i.e., middle school and high school students in comparison to pre-K and elementary students and science and mathematics majors in comparison to social sciences and Turkish and mathematics majors). We have discussed this possibility and its implications in the Results section, and we discuss it further in the Implications and Conclusions section.

IMPLICATIONS AND CONCLUSIONS

We first synthesize and discuss the implications at a general level and then by city, grade level, gender, and high school major within the current study and across the previous studies. We close by considering the overall implications and next steps.

General Implications of Overall Consistency Levels and Force Meaning Variations

At the most general level, approximately only 10% of students coded as “fully consistent.” This number increases to 55% when applying the looser “consistent-with-allowance” criterion (eight of 10 question sets). These findings underscore the existence of some systematicities in students' explanations across contexts, but also highlight the presence of substantial conflicts and inconsistencies in students' explanations across all four grade levels. Related to this, the low levels of consistency at the pre-K level suggest that models of conceptual change that predict high levels of consistency for children need to shift their predictions of increased consistency earlier in the developmental timeline somewhere nearer to the early stages of development where compelling arguments have been made for coherence in babies' expectations about force and motion (e.g., Spelke, Katz, Purcell, Ehrlich, & Breinlinger, 1994). The findings across all grade levels therefore suggest that even elements that wield high magnitudes of influence in a student's conceptual ecology are constantly in competition with other elements depending on context (Clark & Linn, 2013). Taken together, these general findings further underscore the importance of better understanding the processes through which stabilities evolve and subside because flux appears to be the constant state rather than an interim state. This lends support to the potential power of Brown and Hammer's (2008, 2013) focus on local stabilities emerging from complex dynamic system interactions within students' conceptual ecologies and Vosniadou and Skopeliti's (2013) perspective of a framework theory as “a very skeletal structure that imposes only some general constraints on students' explanation of physical phenomenon leaving a degree of freedom for variation and for adjusting to different contextual situations” (p. 7).

City-Specific Conceptual Change Conclusions and Implications

Consistency and Force Meanings Findings. There are no significant differences between cities overall or in interaction with grade level in terms of consistency of explanations or proportions of force meanings expressed for either coding scheme.

Conclusions and Implications. While consistency of explanations has been similar across some countries and cities in the previous studies, best-match force meanings and force meaning scores have varied substantially across previous studies. More specifically, the absence of differences in the current study in terms of best-match force meanings and force meaning scores between the two Turkish cities along with the tight parallels in levels of consistency stand in stark contrast to the differences noted across schools and cities in the United States in Clark et al. (2011), diSessa et al. (2004), and Price Schleigh et al. (2013).

This is especially interesting in light of the fact that the same coding schemes and questions were employed in all of those studies. Furthermore, the very same researchers coded the data in Clark et al. (2011), Price Schleigh et al. (2013), and the current study. This suggests that students' best-match force meanings are more similar across schools and cities in Turkey than they are across schools in the United States or across countries based on the findings of the previous studies. Turkey has a highly standardized national educational system, whereas the United States does not have a highly standardized educational system (even at the level of individual schools or school districts). The differences among educational systems internationally are even greater. These findings therefore suggest that the differences noted across schools and cities in the United States in the ongoing series of studies (a) are less likely a function of random variation and noise in the data collection and analysis process and (b) are more likely a function of differences in the enacted educational systems.

Grade-Level Specific Conceptual Change Conclusions and Implications

Consistency Findings. In terms of consistency, the high school students demonstrate a higher probability of being consistent with allowance and having higher best-match scores than pre-K students according to the I&V scheme. Similar nonsignificant patterns are observed for the DG&E scheme.

Force Meanings Findings. In terms of force meanings, grade levels are significantly related to best-match meanings and force meaning scores based on both coding schemes. Essentially, pre-K students tend to express internal-related meanings, elementary students primarily express the acquired meaning, and middle and high school students tend to express acquired/push-pull and gravity and other meanings.

Conclusions and Implications. In terms of consistency, the older students in the current study are more likely to be consistent with allowance and to have higher best-match scores than younger students. These findings follow the patterns of U.S. students for diSessa et al. (2004) and Clark et al. (2011), but these findings match less well with Clark et al.'s (2011) students from Mexico, China, and the Philippines, who exhibit more similar levels of consistency across grade levels. Generally speaking, the percentages of consistent-with-allowance students are similar for older students across studies independent of the ratio of gravity and other versus push-pull variants observed for the older students in each country and study.⁴ Thus, the differences in levels of consistency across studies and countries are largely a function of differences in levels of consistency among the younger students.

⁴Ioannides and Vosniadou's (2002) study in Greece is an exception to this pattern of similarity. The differences in terms of the levels of consistency observed in Ioannides and Vosniadou's study and the other studies is discussed in great detail in Clark et al. (2011).

One interesting question to explore focuses on why the younger students in the U.S. and Turkey seem to show lower levels of consistency with allowance and best-match scores than the younger students in Mexico, China, and the Philippines. In considering this question, it is important to remember that the coding scheme and question sets are essentially the same in diSessa et al. (2004), Clark et al. (2011), and the current study (and that the latter two studies were coded by the same researchers). Because the differences in consistency manifest primarily at the younger grade level before the educational system has had much opportunity to impact students, and because the outcomes in terms of consistency increase across grade levels, it seems likely that some combination of language and culture provides the dominant influence in terms of the consistency levels at younger grade levels, whereas educational systems bring students into closer alignment over time within and across countries. This perspective is supported by the findings across the two cities in the current study because the biggest differences between the two cities in terms of consistency levels occur at the pre-K level and these differences progressively disappear across the other three grade levels (although, again, the hybrid nature of the gravity and other category adds caveats to these interpretations for the high school students).

In terms of the implications of the force meaning findings, we observe less variation across schools and cities in the current study at each grade level than we observe across countries at each grade level in Clark et al. (2011) and across the U.S. student populations at each grade level in Clark et al. (2011), diSessa et al. (2004), and Price Schleigh et al. (2013).

Four observations seem relevant to explaining this distinction in patterns of force meanings. First, the students across the two Turkish cities in the current study tend to be distributed across the same best-match force meanings at each grade level (e.g., the middle school students in City 1 are distributed across the same force meanings as the middle school students in City 2), whereas the U.S. students at each grade level tend to be distributed across different force meanings from one study to the next (e.g., the middle school U.S. students in one study are distributed across different force meanings than the U.S. middle school students in other studies). This suggests less variation across schools in Turkey than in the U.S. overall.

Second, younger students within each study and country are dispersed across a larger number of best-match meanings than older students. While this difference is largest in comparing high school to pre-K, the pattern is displayed across the grade levels. Although the specter of the hybrid nature of the gravity and other category is once again evoked, this difference potentially suggests that (a) differences in cultural, demographic, and family backgrounds have the most prominent effect at the younger grades and (b) school instruction increases homogeneity over time across the grades.

Third, U.S. students at each grade level in Clark et al. (2011) tend to be dispersed across a larger number of best-match meanings than students in Turkey and the other countries. This highlights variation among schools in the U.S., which makes sense in light of the lack of standardization of educational systems in the U.S. in comparison to Turkey and many other countries.

Fourth, U.S. students in Clark et al. (2011) tend to be dispersed across a larger number of best-match meanings than the U.S. students in diSessa et al. (2004) and Price Schleigh et al. (2013). The U.S. students in Clark et al. (2011) were drawn from at least three different schools at each grade level, whereas all of the students in diSessa et al. (2004) and Price Schleigh et al. (2013) were drawn from a single school in each study for each grade level. This supports the premise of the impact of school system or other school-level variables on outcomes in the sense that Clark et al. (2011) drew U.S. students from a larger number of schools at each grade level than the other two studies and demonstrates greater dispersion of

students across force meanings at each grade level than the other two studies. This variation between schools in the U.S. contrasts starkly with the similarities observed across the two Turkish cities in the current study.

In summary, these findings support the conjecture that the influence of culture and language are strongest at younger ages but that educational system also provides an increasingly significant influence over time. This comparison also supports the conjecture that the differences noted across previous studies are driven by the same combination of influences.

Gender-Specific Conceptual Change Conclusions and Implications

Consistency and Force Meaning Findings. As seen across the prior studies, this study demonstrates no significant differences by gender in terms of either consistency or force meanings.

Conclusions and Implications. The absence of differences across gender (where we would not expect to observe differences based on the findings across all of the previous studies) suggests that the variations we see across majors and grade level are not merely noise in the data. Furthermore, while not the central focus of the current study, these findings also contribute to the larger debate about gender differences in cognitive skills and abilities in science and other domains (e.g., Ceci & Williams, 2007; Fine, 2011; Hines, 2010). More specifically, these findings clarify the similarities across genders in terms of the core science thinking skills at the heart of the current study. These findings parallel (a) findings by Sencar and Eryilmaz (2004) that Turkish students' thinking about electromagnetism does not differ by gender when controlling for interests and (b) findings by Dogan and Abd-El-Khalick (2008) that Turkish students' beliefs about the nature of science do not differ by gender.

Academic-Major-Specific Conceptual Change Conclusions and Implications

Consistency Findings. There are pronounced differences across high school majors in terms of consistent-with-allowance codes and best-match scores. Essentially, social sciences students are less likely to be consistent with-allowance than science and mathematics and Turkish and mathematics students. This pattern is significant for the DG&E scheme and follows a similar nonsignificant pattern for the I&V scheme. Similarly, social sciences students have significantly lower best-match scores than either science and mathematics or Turkish and mathematics students for both coding schemes.

Force Meanings Findings. No significant relationships were found across majors in terms of best-match meanings based on either coding scheme, but science and mathematics students predominantly express gravity and other as their best-match meaning whereas Turkish and mathematics and social sciences students predominantly express acquired/push-pull as their best-match meaning. There are significant differences across majors, however, on force meaning scores based on the DG&E coding scheme. The science and mathematics students have significantly higher gravity and other scores than the social sciences students. Conversely, social sciences students have significantly higher push-pull scores than the science and mathematics students. Finally, the science and mathematics students have higher internal/acquired scores than the Turkish and mathematics students.

Conclusions and Implications. There are significant differences across majors within a single city even though the comparisons across majors involved much lower statistical power than the comparisons across cities (i.e., the power to detect only very large effect sizes versus the power to detect medium effect sizes). The differences between majors can therefore be assumed to be substantial. Interpreting the sources of the differences between students in the different majors is challenging to interpret, however, because (a) all students have the same course work for first through ninth grade, (b) the students in the current study are in 10th or 11th grade but only science and mathematics students take science courses in 10th and 11th grades, and (c) students choose their majors based on personal preferences, strengths, goals, and academic histories that have developed through their lived experiences under the influences of their home, school, and peer communities and resources. It would be easier to interpret if the study had focused on ninth graders and categorized the students post hoc based on their choices at the end of ninth grade, but 10th- and 11th-grade students were chosen to align with the grade level of the preceding studies for the purposes of comparison.

Interestingly, the Turkish and mathematics students are significantly more consistent than the social sciences students even though both groups most frequently express acquired/push-pull meanings. The significantly higher consistency of the science and mathematics students in comparison might be attributed to the hybrid nature of the gravity and other category (which was the predominantly expressed meaning for the science and mathematics students), but the higher consistency of the Turkish and mathematics students in comparison to the social sciences students does not include that caveat. The social sciences and Turkish and mathematics students therefore demonstrate larger differences than seen across the two cities even though neither group participates in any additional science courses after ninth grade and both groups have taken the same classes from first through ninth grade in the same city and school systems.

Students' personal preferences, strengths, goals, and academic histories therefore contribute substantially more variation in outcomes than is contributed by city. At a more general level, these findings support earlier research focusing on the interaction of cultural context with individual cognition and conceptual change (e.g., Hatano, 1994; Hatano & Inagaki, 2003). These findings also provide additional evidence for the existence of multiple paths of conceptual change that students pursue even within a standardized educational system. Clark (2006) documented the diverse and multiple longitudinal paths and trajectories of four students making sense of thermodynamics ideas across eighth grade and high school. The current study builds on those findings by demonstrating that students in different academic majors demonstrate differing levels of consistency and predominant force meanings despite highly standardized educational experiences.

Overall Conclusions, Synthesis, and Final Thoughts

The current study provides insight into the degree to which the variation in findings across the studies might be attributed to random variation, cultural differences, or differences in educational system. While the limited scale of the current study clearly precludes conclusive claims, the findings of the current study in comparison to the findings of the previous studies suggest that (a) individuals' interests, abilities, experiences, and affiliations contribute to variation (particularly at younger grades but also in older grades as highlighted by the comparisons across majors for high school students in the current study), (b) instruction provides an increasingly significant influence over time (with the potential caveat in terms of claims of increased consistency in light of the hybrid nature of the gravity and other category that is more prevalent for older students), (c) educational systems can support

homogenization of outcomes across schools and cities, (d) differences in educational systems likely contribute substantially to differences across U.S. samples in prior studies, and (e) international differences in educational systems and cultures likely contribute substantially to differences across countries in the prior studies. The findings of the current study also clarify (a) that processes of flux and change are constant rather than interim states and (b) that conceptual change proceeds through multiple possible paths that are influenced by differences in individuals' interests, abilities, experiences, and affiliations (which to some degree might be considered imperfect surrogates for culture).

In terms of methodological implications and future research, the current study's findings highlight the value of the current approach and data grain size but suggest methodological refocusing for future work. In particular, the findings of the current study highlight the importance of focusing future research on longitudinal and microgenetic studies rather than cross-sectional studies to more clearly map (a) the specific nature of the knowledge elements that apply high magnitudes of influence on other elements in the conceptual ecologies, (b) the processes through which stabilities evolve and change, and (c) the relationships between instructional variables and individual variables including interests, abilities, experiences, and affiliations in these processes.

In addition to implications for theory and future research, the findings of the current study carry important implications for instruction. On an optimistic note, while individuals' interests, abilities, experiences, and affiliations clearly matter, the findings across cities suggest that instruction clearly matters. From this perspective, future research characterizing students' conceptual ecologies could beneficially leverage Brown and Hammer's (2013) calls for research identifying productive "conceptual attractors" to foster productive stabilities across the longitudinal curriculum. The current study highlights, however, the sheer number of influences and sources of ideas in students' conceptual ecologies beyond instruction. The diverse range of influences guarantees that learners will follow many different paths as they make sense of abstract school science ideas. This underscores (a) the emphasis of current knowledge-in-pieces and framework theory perspectives that every student's personal learning history is unique and (b) the importance of flexible curricula that support multiple paths of conceptual change based on individuals' interests, abilities, experiences, and affiliations.

Finally, while not the focus of the current study, the findings also suggest policy implications. The comparison across cities suggests that the Turkish educational system appears relatively successful in supporting the espoused Turkish goal of promoting more equal and standardized educational outcomes through a strongly standardized national educational system. While the scope and implementation of the science education curriculum in Turkey are certainly subject to improvement in the sense of overall test scores on *Trends in International Mathematics and Science Study* (TIMSS) and PISA as well as in terms of emphasis on inquiry and the processes of science (cf., Dogan & Abd-El-Khalick, 2008; Irez, 2006), the consistency across cities in the current study underscores the potential of a strongly nationalized educational system to affect learning outcomes across a country. By comparison, the ability to impact and influence students' understanding of science across grades has historically eluded reform calls and efforts in the United States over the past century. The similarity of findings across the two cities in Turkey could therefore be construed as supporting claims that the educational system (or lack of a strong national educational system) in the United States represents a powerful missed opportunity to influence science learning outcomes. The current study's findings therefore suggest implications and provide foundations for future research into policy as well as theory.

This material is based on work supported by a National Academy of Education/Spencer Postdoctoral Fellowship awarded to Douglas Clark. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Academy of Education or the Spencer Foundation.

REFERENCES

- Aikenhead, G., & Jegede, O. J. (1999). Cross-cultural science education: A cognitive explanation of a cultural phenomenon. *Journal of Research in Science Teaching*, 36(3), 269–287.
- Ayas, A., Cepni, S., & Akdeniz, A. R. (1993). Development of the Turkish secondary science curriculum. *Science Education*, 77(4), 422–440.
- Aypay, A., Erdogan, M., & Sozer, M. A. (2007). Variation among schools on classroom practices in science. *Journal of Research in Science Teaching*, 44(10), 1417–1435.
- Berberoglu, G., & Kalender, I. (2005). Investigation of student achievement across years, school types and regions: The SSE and PISA analyses. *Educational Sciences and Practice*, 4(7), 21–35.
- Brown, D., & Hammer, D. (2008). Conceptual change in physics. In S. Vosniadou (Ed.), *The international handbook of research on conceptual change* (pp. 127–154). New York: Routledge.
- Brown, D., & Hammer, D. (2013). Conceptual change in physics. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (2nd ed., pp. 121–137). New York: Routledge.
- 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. (2000). Science education as conceptual change. *Journal of Applied Developmental Psychology*, 21(1), 13–19.
- Ceci, S. J., & Williams, W. M. (Eds.). (2007). *Why aren't more women in science? Top researchers debate the evidence*. Washington, DC: American Psychological Association.
- Çelik, C. H., & Ceylan, H. (2009). The comparison of high school students' mathematics and computer attitudes according to various variables. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 26, 92–101.
- Chai, C. C., Deng, F., Qian, Y. Y., & Wong, B. (2010). South China education major's epistemological beliefs and their conceptions of nature of science. *The Asia-Pacific Education Researcher*, 19(1), 111–125.
- 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(4), 467–563.
- Clark, D. B., D'Angelo, C., & Schleigh S. (2011). Multinational comparison of students' knowledge structure coherence. *Journal of the Learning Sciences*, 20(20), 207–261.
- Clark, D. B., & Linn, M. C. (2013). The knowledge integration perspective: Connections across research and education. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (2nd ed., pp. 520–538). New York: Routledge.
- Constantinou, C., Hadjilouca, R., & Papadouris, N. (2010). Students' epistemological awareness concerning the distinction between science and technology. *International Journal of Science Education*, 32(2), 143–172.
- Costa, V. (1995). When science is "another world": Relationships between worlds of family, friends, school, and science. *Science Education*, 79, 313–333.
- Demiray, G., & Dolu, N. (2011). Evaluation of multiple intelligence in the students preparing university exam. *Journal of Health Sciences*, 20(1), 29–38.
- Deng, F., Chen, D.-T., Tsai, C.-C., & Chai, C. S. (2011). Students' views of the nature of science: A critical review of research. *Science Education*, 95(6), 961–999.
- diSessa, A. A. (1983). Phenomenology and the evolution of physics. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 5–33). Hillsdale, NJ: Erlbaum.
- diSessa, A. A. (1993). Toward an epistemology of physics. *Cognition and Instruction*, 10(2/3), 105–225.
- diSessa, A. A., Gillespie, N., & Esterly, J. (2004). Coherence versus fragmentation in the development of the concept of force. *Cognitive Science*, 28, 843–900.
- Dogan, N., & Abd-El-Khalick, F. (2008). Turkish grade 10 students' and science teachers' conceptions of nature of science: A national study. *Journal of Research in Science Teaching*, 45(10), 1083–1112.
- Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, 28, 1–11.
- Fine, C. (2011). Explaining, or sustaining, the status quo? The potentially self-fulfilling effects of "hardwired" accounts of sex differences. *Neuroethics*, Online First™, 21 June 2011.
- 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(8), 371–377.
- Griffiths, A. K., & Barman, C. R. (1995). High school students' views about the nature of science: Results from three countries. *School Science and Mathematics*, 95(5), 248–255.
- Gungor, A., Eryılmaz, A., & Fakioglu, T. (2007). The relationship of freshmen's physics achievement and their related affective characteristics. *Journal of Research in Science Teaching*, 44(8), 1036–1056.
- Hammer, D., Redish, E. F., Elby, A., & Scherr, R. E. (2005). Resources, framing, and transfer. In J. Mestre (Ed.), *Transfer of learning: Research and perspectives* (pp. 89–120). Greenwich, CT: Information Age.
- Hamurcu, H., Günay, Y., & Özyılmaz, G. (2002). Buca eğitim fakültesi fen bilgisi ve sınıf öğretmenliği bölümü öğrencilerinin çoklu zekâ kuram'ına dayalı profilleri. *Proceedings of V. Ulusal Fen Bilimleri ve Matematik Eğitimi Kongresi* (Vol. 1, pp. 415–421). Ankara, Turkey : ODTÜ.
- Hatano, G. (1994). Conceptual change—Japanese perspectives. *Introduction. Human Development*, 37(4), 189–197.
- Hatano, G. & Inagaki, K. (2003). When is conceptual change intended? A cognitive-sociocultural view. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change*. (pp. 407–427). Mahwah, NJ: Erlbaum.
- Hines, M. (2010). Sex-related variation in human behavior and the brain. *Trends in Cognitive Sciences*, 14(10), 448–456.
- Inagaki, K., & Hatano, G. (2002). *Young children's naive thinking about the biological world*. New York: Psychology Press.
- Ioannides, C., & Vosniadou, S. (2002). The changing meanings of force. *Cognitive Science Quarterly*, 2(1), 5–62.
- Irez, S. (2006). Are we prepared?: An assessment of preservice science teacher educators' beliefs about nature of science. *Science Education*, 90(6), 1113–1143.
- İzci, E., Kara, A., & Dalaman, F. (2007). The analysis of private courses teaching's students via the theory of multiple intelligence. *Pamukkale Üniversitesi Eğitim Fakültesi Dergisi*, 21, 1–14.
- Kalender, I., & Berberoglu, G. (2009). An assessment of factors related to science achievement of Turkish students. *International Journal of Science Education*, 31(10), 1379–1394.
- Karabenick, S. A., & Moosa, S. (2005). Culture and personal epistemology: U.S. and Middle Eastern students' beliefs about scientific knowledge and knowing. *Social Psychology of Education*, 8, 375–393.
- Linn, M. C. (2006). The knowledge integration perspective on learning and instruction. *The Cambridge handbook of the learning sciences* (pp. 243–264). New York: Cambridge University Press.
- Liu, S.-Y., & Tsai, C.-C. (2008). Differences in the scientific epistemological views of undergraduate students. *International Journal of Science Education*, 30(8), 1055–1073.
- Lubben, F., Netshisaulu, T., & Campbell, B. (1999). Student's use of cultural metaphors and their scientific understandings related to heating. *Science Education*, 83, 761–774.
- Mansour, N. (2011). Science teachers' views of science and religion vs. the Islamic perspective: Conflicting or compatible? *Science Education*, 95(2), 281–309.
- McCloskey, M. (1983). Intuitive physics. *Scientific American*, 248(4), 122–130.
- Miller, M. C. D., Montplaisir, L. M., Offerdahl, E. G., Cheng, F.-C., & Ketterling, G. L. (2010). Comparison of views of the nature of science between natural science and nonscience majors. *CBE—Life Sciences Education*, 9, 45–54.
- Nisbett, R. E., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.
- Ornek, F. (2011). Cultural influence on attitudes towards science. In I. M. Saleh & M. S. Khine (Eds.), *Attitude research in science education: Classic and contemporary measurements*. Charlotte, NC: Information Age.
- Osman, T., & Cobern, W. W. (2011). Valuing science: A Turkish-American comparison. *International Journal of Science Education*, 33(3), 401–421.
- Ozdemir, G., & Clark, D. B. (2009). Knowledge structure coherence of Turkish students' understanding of force. *Journal of Research on Science Teaching*, 46(5), 570–596.
- Pehlivan, M. (2008). The correlation between multiple intelligence profiles and the fields and the scores according to which the students are placed in student selection exam (ÖSS). Unpublished master's thesis, Zonguldak Karaelmas Üniversitesi, Zonguldak, Turkey.
- Phelan, P., Davidson, A., & Thanh Cao, H. (1991). Students' multiple worlds: Negotiating the boundaries of family, peer, and school cultures. *Anthropology and Education Quarterly*, 22, 224–250.
- Price Schleigh, S., Clark, D. B., & Menekse, M. (2013). Constructed-response as an alternative to interviews in conceptual change studies: Students' explanations of force. Manuscript submitted for publication.
- Sencar, S., & Eryılmaz, A. (2004). Factors mediating the effect of gender on ninth-grade Turkish students' misconceptions concerning electric circuits. *Journal of Research in Science Teaching*, 41(6), 603–616.
- Sezer, A. (2010). Defining the realization level of learning experiences in high school Geography course. *Eastern Geographical Review*, 24, 211–236.

- Spelke, E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1994). Early knowledge of object motion: Continuity and inertia. *Cognition*, 51(2), 131–76.
- State Planning Organization. (2003). Survey on the ranking of provinces and regions by social–economic development levels. SPO publication no: 2671. Ankara, Turkey: General Directorate of Regional Development and Structural Adjustment (in Turkish).
- 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.
- Sünbül, M. A., & Sari, H. (2004). An analysis of high school students' learning strategies and styles in Turkey. Proceedings of the International Conference on Quality in Education in the Balkan Countries, Bulgaria (Vol. 1, pp. 530–545). Thessaloniki, Greece: Kyriakidis Brothers s.a. Publishing House.
- Sutherland, D., & Dennick, R. (2002). Exploring culture, language and the perception of the nature of science. *International Journal of Science Education*, 24(1), 1–25.
- Thaden-Koch, T. C., Dufresne, R. J., & Mestre, J. P. (2006). Coordination of knowledge in judging animated motion. *Physical Review Special Topics—Physics Education Research*, 2(2), 020107.
- The Ministry of National Education. (2001). The Turkish Education System and developments in education. Retrieved November 20, 2013, from <http://www.ibe.unesco.org/International/ICE/natrap/Turkey.pdf>.
- Titrek, O., & Cobern, W. W. (2011). Valuing science: A Turkish-American comparison. *International Journal of Science Education*, 33(3), 401–421.
- Tunç, E. (2008). The relationship between the multiple intelligence field of eleven grade of high school students and type of high-school they attend, their education and sexuality. *Ataturk Üniversitesi Kazım Karabekir Eğitim Fakültesi Dergisi*, 17, 108–130.
- van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
- Vosniadou, S. (2013). Conceptual change in learning and instruction: The framework theory approach. In S. Vosniadou (Ed.), *International handbook of research on conceptual change* (2nd ed., pp. 539–559). New York: Routledge.
- Vosniadou, S., & Skopeliti, I. (2013). Conceptual change from the framework theory side of the fence. *Science & Education*. Published Online August 2013 (pp. 1–19).
- 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.
- Wellman, H. M., & Gelman, S. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology*, 43, 337–375.
- Wen, M. L., Kuo, P.-C., Tsai, C.-C., & Chang, C.-Y. (2010). Exploring high school students' views regarding the nature of scientific theory. *The Asia-Pacific Education Researcher*, 19(1), 161–177.
- Wiser, M., & Carey, S. (1983). When heat and temperature were one. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 267–298). Hillsdale, NJ: Erlbaum.
- Yenice, N., & Aktamış, H. (2010). Sınıf öğretmenleri adaylarının çoklu zekâ alanlarının demografik özelliklere göre incelenmesi. *Journal of Turkish Science Education*, 7(3), 86–99.
- Yıldız, S., & Turanlı, N. (2010). Investigation of university entrance exam students' attitudes to mathematics. *Selçuk Üniversitesi Ahmet Keleşoğlu Eğitim Fakültesi Dergisi*, 30, 361–377.