Evidence of Student Ideas in the Validation of Middle School Forces and Motion Assessment Items: Using Interviews to Confirm Student Idea Use

By Robert J. Ochsendorf

B.A. 1999, Colgate University
M.A. 2004, The George Washington University

A Dissertation Submitted to

The Faculty of
The Graduate School of Education and Human Development
of The George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Education

January 31, 2009

Dissertation directed by

Curtis Pyke

Professor of Secondary Education

The Graduate School of Education and Human Development of The George Washington University certifies that Robert J. Ochsendorf has passed the Final Examination for the degree of Doctor of Education as of October 16th, 2008. This is the final and approved form of the dissertation.

Evidence of Student Ideas in the Validation of Middle School Forces and Motion Assessment Items: Using Interviews to Confirm Student Idea Use

Robert J. Ochsendorf

Dissertation Research Committee:

Curtis Pyke, Professor of Secondary Education, Dissertation Director

Sharon Lynch, Professor of Secondary Education, Committee Member

George DeBoer, Deputy Director, AAAS—Project 2061, Committee Member

Acknowledgments

This dissertation would not have been possible without the research conducted by SCALE-uP, a collaboration between The George Washington University (GWU) and Montgomery County Public Schools (MCPS), Maryland. Drs. Sharon Lynch, Joel Kuipers, and Curtis Pyke of GWU, and Michael Szesze and Bonnie Hansen-Grafton of MCPS served as principal investigators of SCALE-uP. Funding for SCALE-uP was provided by the National Science Foundation, the U.S. Department of Education, and the National Institute of Health (REC-0228447). I would like to thank the committed efforts of my dissertation chair, Curtis Pyke, and committee members, Sharon Lynch and George DeBoer. Their patience, dedication, and wisdom greatly contributed to the development of the ideas in this dissertation. I would also like to thank the students and the interviewers, Carol and Bill, for the time and effort committed to making this study possible. Coding of the transcript data would not have been possible without the generous assistance of Jinsong Chen. Finally, I am grateful to have the loving and steadfast support of my wife, Vi. Thanks for always being there.

Abstract of Dissertation

Evidence of Student Ideas in the Validation of Middle School Forces and Motion Assessment Items: Using Interviews to Confirm Student Idea Use

This study explored the extent to which seven paper and pencil assessment items judged by experts to be aligned with forces and motion ideas actually measure those concepts for middle school students. To accomplish this, the study compared students' individual responses under paper and pencil and interview conditions while considering an elaborated model of a target science idea assessed by the items. The study sought to confirm the inferences provided by a written science assessment instrument by comparing them to alternative inferences based on student interview data.

Participants in the study were a diverse sample of 14 middle school students. Each student completed the written assessment instrument and then participated in an interview designed to confirm student idea use. All interview transcripts were coded for idea use and the results from coded data were then compared to students' performance on the written posttest. The study concludes that in most cases, the two sets of assessment evidence were in agreement and that the items provide valid inferences about students' actual understanding of the target ideas. The study also concludes that on a few occasions the written posttest overestimated students' performances when compared to data collected during the interviews. The study presents additional student ideas regarding forces and motion that have been previously undocumented in the literature. Conclusions and implications are discussed.

Table of Contents

Acknowledgments iii
Abstract of Dissertation iv
Table of Contents v
List of Figures vii
List of Tables viii
CHAPTER I: INTRODUCTION 1
The Problem and Its Setting 1
Theoretical Foundations 3
Context of the Problem 4
Problem of Research Gap 7
Accounting for Cognition in Assessment 8
Student Learning of Forces and Motion 13
Research Questions 16
Methods 17
Significance of the Study 19
Assumptions of the Study 21
Delimitations of the Study 22
Limitations 23
Definitions of Key Terms 24
Dissertation Overview 26
CHAPTER II: REVIEW OF THE LITERATURE 27
Overview 27
Organization of the Chapter 28
Literature Review Questions 29
What Assessment Instruments Have Been Used to Measure These Ideas? 30
What Ideas Do Middle School Students Have About Forces and Motion? 37
Target Idea 1: Changes in Speed or Direction are Caused by Forces 42
Target Idea 2: Objects in Motion Stay in Motion Unless Acted on by a Force 47
Target Idea 3: Objects at Rest Stay at Rest Unless Acted on by a Force 49
Summary 51
How Has Interviewing Been Used to Investigate Student Thinking? 53
Conceptual Framework 59
Comment on Validity 61
Conclusions 62
CHAPTER III: METHODOLOGY 64
Organization of the Chapter 64
Paradigm and Model of Cognition 64
Paradigm of Inquiry 64
Model of Cognition 65
Research Questions 71
Methodology 71
Background on SCALE-uP and Study Site 72
Background on SCALE-uP 72
Description of Study Site 74

Student Participants 74
Instrumentation 77
Descriptions of Assessment Items for this Study 85
Interview Protocol and Data Collection 92
Interview Protocol for MFA 92
Interviews and Data Preparation 95
Data Analysis 97
Phase I: Code Category Development 98
Phase II: Final Coding of Interview Data 103
The Coding of Transcripts for Research Question #1 104
Holistic Judgments to Address Research Question #2 105
Contributions to Literature Through Research Question 3 108
Human Subjects and Ethics Precautions 108
CHAPTER IV: RESULTS 110
Characteristics of the Video Data 110
Overview of Data Coding 111
Sample of Coded Interview 113
Research Question 1 116
Rolling Ball—Selected Response 118
Rolling Ball Constructed Response 122
Dropping and Throwing a Ball—Selected Response 128
Dropping and Throwing a Ball—Constructed Response 129
Carly's Skateboard—Selected Response 135
A Metal Ball—Selected Response 141
Pedaling a Bike—Selected Response 144
Research Question 2 149
Research Question 3 161
Summary 164
CHAPTER V: CONCLUSIONS AND RECOMMENDATIONS 165
Purpose and Significance of the Study 165
Summary and Interpretations of Findings 166
Discussion and Implications 183
Project 2061 Assessment Analysis 183
Assessment of a Motion and Forces Standard 185
Conclusions and Recommendations 191
Critique of Study 193
Appendix A: SCALE-uP Forces and Motion Student Interview Protocol 196
Appendix B: Evidence Statements and Student Idea Codes for Each Assessment Item 208
References 216

List of Figures

Figure 1. The assessment triangle (NRC, 2001). 9

Figure 2. A conceptual framework based on the assessment triangle for comparing inferences from student assessment data. 60

Figure 3. Assessment map of forces and motion target ideas used by SCALE-uP in development of MFA. 78

Figure 4. A Rolling Ball (two items). 86

Figure 5. Dropping and Throwing a Ball (two items). 88

Figure 6. Carly's Skateboard. 89

Figure 7. Pedaling a Bike. 90

Figure 8. A Metal Ball. 92

Figure 9. Sampling of interview transcripts for coding. 103

Figure 10. Frequency of student idea codes assigned across all students. 116

List of Tables

Table 1. *Questions Responded to by Students During Interviews.* 18

Table 2. *Journal Articles Resulting from ERIC Search on Forces and Motion.* 40

Table 3. *A List of Common Middle School Forces and Motion Misconceptions.* 52

Table 4. *MFA and Project 2061 Assessment Items with Intended Target Ideas.* 84

Table 5. *Student Idea Codes Used for Initial Coding of Interview Transcripts.* 100

Table 6. *Characteristics of Videotaped Interviews for All Students.* 111

Table 7. *Final Student Idea Codes Used to Code Interview Transcripts.* 112

Table 8. *Frequency of Student Idea Codes for Each Student Based on Interview Data.* 115

Table 9. *Judgments Made for All Students Interviewed per Item During SCALE-uP's Written MFA Posttest and Oral Interview.* 119

Table 10. *Interviewed Students' Level of Understanding as Judged by Two Methods.* .. 152

Table 11. *Judgments Made for Students During SCALE-uP's Written MFA Posttest and Oral Interview.* 169

CHAPTER I: INTRODUCTION

The Problem and Its Setting

This study explored the extent to which seven paper and pencil assessment items judged by experts to be aligned with forces and motion ideas actually measure those concepts for middle school students. To accomplish this, the study compared students' individual responses under paper and pencil and interview conditions while considering an elaborated model of a target science idea assessed by the items. The study determined if the paper and pencil assessment items elicited the expected student ideas, as determined by a review of the literature, by analyzing videotaped interview data of students responding to the assessment items and determining if responses provided during the interview agreed with written responses provided just prior to the interview. The study used an iterative code development procedure (Strauss & Corbin, 1990) to draw conclusions regarding which student ideas about forces and motion were being used during students' oral responses to assessment items and compared them to the ideas used during written responses to the same assessment items. The videotaped interviews were part of a transcribed video database assembled by the Scaling-up Curriculum for Achievement Learning, and Equity Project (SCALE-uP) during 2005-2006 as part of its ongoing research efforts on assessment development and validation.

Although assessment standards documents have proposed methodologies to explore construct validity from a psychometric, test structure perspective (American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing, 1999), few guidelines exist that provide researchers with the

methodological tools needed to explore construct validity in more qualitative ways that include a precise description of the content to be measured and how to document its use by students. Furthermore, little research has been conducted in the educational measurement or science education fields that offers a framework for student ideas about specific science content to assist in the construct validation process. Thus, validation efforts seeking evidence of concepts in use when students are responding to test items are left to work in absence of such frameworks or are left to propose their own (which few studies accomplish). This troubling gap in the measurement field is discussed in more detail by Leighton (2004) and will be revisited later in this dissertation.

This study created a framework for middle school student understanding in the domain of forces and motion by reviewing the relevant literature around research on student ideas. A student idea is defined in this study as an unobservable knowledge structure in a given domain possessed by an individual that is used by the individual to make sense of phenomena in nature. A student idea can be accepted as accurate or inaccurate by content experts (Driver, Squires, Rushworth, & Wood-Robinson, 1994; Pfundt & Duit, 1994). Determining the ideas that middle school students are likely to use, as defined by the research literature, and the ones that the students in this study actually do use to respond to standards-based assessments items is a central element in this study. The study explored whether the targeted forces and motion ideas are evoked by students as they responded to assessment items during interviews. A major goal was to confirm if the written assessment items, judged to be aligned to particular science ideas, satisfactorily elicited forces and motion student ideas (both the scientifically correct ideas that could be expected of middle school students and the misconceptions that they are

likely to have). Analysis of the ideas as described in the literature, the ideas as used by the students, and the inferences made about student understanding using a paper and pencil assessment instrument allowed an in-depth look at the validity of the assessment items in this study.

This chapter presents a basis for the study by outlining the context of the problem, the research problem statement, the research questions, the significance of the study, the theoretical framework of the study, the assumptions, the methodology, the delimitations, the limitations, and the definition of key terms. The chapter concludes with a summary of the contents presented.

Theoretical Foundations

This study is aligned with a cognitive perspective of learning (Bransford, Brown, & Cocking, 1999; Greeno, Collins, & Resnick, 1997; Mislevy, 2006). This means that conceptual understandings (i.e., ideas), while shared and negotiated in public spaces, exist within a personal mental framework. The evidence in the science education literature, as well as in the experimental and cognitive psychology literature (e.g., short term memory, long term memory), is relevant in understanding students' interactions with assessment tasks as individuals access their personal mental frameworks, select relevant ideas, and make sense of and explain phenomena represented by the assessment tasks. Such mental frameworks consist of ideas about the natural world that exist in individuals' minds. Well-designed qualitative inquiry should allow for inferences about the nature and use of those ideas by individuals. Although such ideas are clearly not directly observable, it is possible to collect evidence consistent with their existence and to observe when such evidence suggests ideas have changed in response to students'

experiences (Park & Han, 2002; Posner, Strike, Hewson, & Gertzog, 1982; Strike & Posner, 1992). By observing and explaining the changes in the ways in which students respond to and explain natural phenomena, one can arrive at reasoned conclusions about the nature of the student conceptions at different points in time.

The study is also based on a constructivist view of learning. Constructivist ideas refer to the process of knowledge being constructed by individuals as they are exposed to particular experiences (Phillips, 1995). As individuals personally construct their understanding of specific science ideas, assessment developers are challenged to arrive at instruments that accurately characterize the breadth of student thinking across large numbers of students. This challenge faced by assessment developers and researchers provides for a particularly rich setting in which to conduct research on the validity of assessment items.

Context of the Problem

Due to increased concerns about accountability, never before in the history of U.S. schools has educational assessment played as big a role as it does today. Assessments impact virtually all aspects of our educational system and have become ubiquitous in teaching and learning practices (National Research Council [NRC], 2001). State and national accountability systems require assessing K-12 students on large scales and addressing issues of failing schools, inequitable performance, and international comparisons that consistently place U.S. achievement well below where the U.S. would like to be. In short, educational policy such as the No Child Left Behind Act (NCLB) (U.S. Congress, 2001) has led to a heightened interest in measuring student progress at all educational levels and across all disciplines.

Accountability and improved performance are especially important in science education. In science, U.S. students consistently perform below expectations relative to their international peers, as presented in the most recent report of the Trends in International Mathematics and Science Study 2003 (TIMSS) (Gonzales et al., 2004). In addition, U.S. ethnic and socioeconomic achievement gaps in science education are remarkably persistent, leading to gaps in subsequent career choices (NRC, 2001). Improving science achievement for all students is viewed as a promising way forward to maintain a competitive advantage in a global economy (Augustine, 2006). As a way of monitoring the achievement of all students, large-scale assessment systems at the national level (e.g., National Assessment of Educational Progress[NAEP]) and high-stakes accountability systems at the state level (e.g., Virginia Standards of Learning, Maryland State Assessment) (mandated in part through NCLB) have become more significant for educators in schools and more visible to the public. In addition, research in the field of education has relied heavily on these assessment systems to determine the effectiveness of various educational interventions (Confrey & Stohl, 2004). Despite the centrality of good assessments to experimental and evaluation research in math and science, researchers often conduct studies with insufficient attention to sound measurement practice (Confrey & Stohl, 2004; NRC, 2001).

Given this inattention to sound measurement practice, it is important to begin to rely on assessment systems that involve more sensitive and accurate measurement techniques for specified content standards. Effective measurement techniques might be characterized as using high quality instruments that are well aligned to the target constructs and integral to understanding what students are learning in science. High

quality instruments may determine relative strengths and weaknesses of education programs. The fate of individuals is also often determined by assessments and so students deserve high quality testing of their understanding. This suggests that an assessment system should reflect more in-depth understandings of the target ideas and make a strong case that the assessment instrument does, in fact, measure those ideas. Procedures should be in place to ensure that data from instruments are valid and provide accurate information regarding the understanding of how target ideas may exist "in the minds of students" and how students use them in assessment situations. A validation study can provide guidance to the researcher and the public about the appropriate use of an assessment and the ability to make inferences about the ideas of individuals and groups.

In assessment validation, construct validation involves determining the degree of alignment between an assessment item, a student's response, and the idea or piece of knowledge that it intends to measure. The American Association for the Advancement of Science (AAAS) Project 2061 Item Analysis Procedure (DeBoer, 2005) uses expert panel review to evaluate the alignment of assessment items to learning goals, using specific review criteria. The review criteria are divided into three main areas: a *content analysis* which is interested in the extent to which the targeted science idea is both necessary and sufficient to respond to the assessment task; *confounding factors* such as comprehensibility, appropriateness of task context, and resistance to test wiseness; and, *cost-effectiveness* (DeBoer, 2005). Application of these review criteria during item development also is used to reduce the number of construct irrelevant features of a test item. For the set of assessment items investigated in this study, alignment is judged by using the Project 2061 assessment analysis criteria (Chapter III will include more details

on this procedure). Although these criteria are useful in providing some level of assurance that the items align with the intent of content standards and empirical evidence drawn from large numbers of students can provide further evidence regarding construct validity, in-depth analyses that delve into actual student responses can provide an additional, crucial piece of evidence regarding item construct validity. By closely examining the ways in which students respond to assessment items, focusing on student use of particular forces and motion ideas, this study provides more direct evidence for alignment of assessment items to target ideas.

Alignment at the item level (DeBoer, 2005) enables one to make more robust inferences about what students know or can do from a collection of items. Assessment instruments composed of a collection of such aligned items can have important implications for instruction, curriculum, or even local and federal policy. For example, inferences about student competency levels in a given domain, based on items having precise alignment to learning goals, could influence important curriculum and instruction decisions that are made at the state or local levels. Thus, it is vital that research in construct validation provides information about the nature of student knowledge with enough precision to make solid inferences about student knowledge.

Problem of Research Gap

Despite a long history of assessment validation research, few studies exist in assessment research that develop and use specific descriptions of student ideas for use in the construct validation process. One exception to this has been the extensive research done by physics educators on the development of students' conceptions of Newtonian mechanics (see for example, Champagne, Klopfer, & Anderson, 1980; Chi, Feltovich, &

Glaser, 1981; Clement, 1982; Driver et al., 1994; Hestenes, Wells, & Swackhamer, 1992; Minstrell, 1982; Pfundt & Duit, 1994). However, very few of those studies have explored the nature of those conceptions for younger middle school students in the context of research on assessment validation. This research gap will be briefly addressed here, but given more attention in Chapter II.

Accounting for Cognition in Assessment

Given that the usefulness of any assessment depends upon the quality of the inference that can be made from it, the trustworthiness of that inference is vital. Currently, validation of assessment items is often done by examining psychometric data and looking for patterns among student responses (i.e., factor analyses, inter-item correlations). Although certainly necessary, such psychometric approaches to validation if done in isolation of actual student reasoning, fail to capture ways in which students are actually thinking about a target idea. For example, inter-item correlations may help to determine which two items are statistically related to each other but offer little explanation as to why they are related or if students are thinking in similar ways when responding to both items. Another approach to validation, sometimes called content validation, involves asking content experts to examine test items and judge the extent to which items sample a specified performance domain (Crocker & Algina, 1986). The limitation here is that speculation by experts as to which content is necessary for responding to an item may often be quite different from what content students actually use when responding. Other common approaches for addressing issues of validity often involve criterion-related validation studies that focus on relationships between one measure of a concept (e.g., Mechanics Baseline Test) and another acceptable measure

(e.g., Force Concept Inventory). If the relationship between the two measures purporting to measure the same concept is strong, this is evidence of validity. These kinds of approaches are endorsed in national assessment standards documents for educational and psychological assessment (American Educational Research Association et al., 1999).

However, such approaches to determining or measuring validity do not address the student's ideas being tested (Gorin, 2006; Leighton, 2004). Validation research that fails to propose a model for the construct(s) of interest neglects a crucial aspect of validation work. Despite the promise of cognitive research for improving assessment, few large-scale assessment developers systematically account for the role of cognition. The corners of the assessment triangle (Figure 1) represent three key elements underlying any assessment: cognition, observation, and interpretation—as summarized by the NRC (2001).

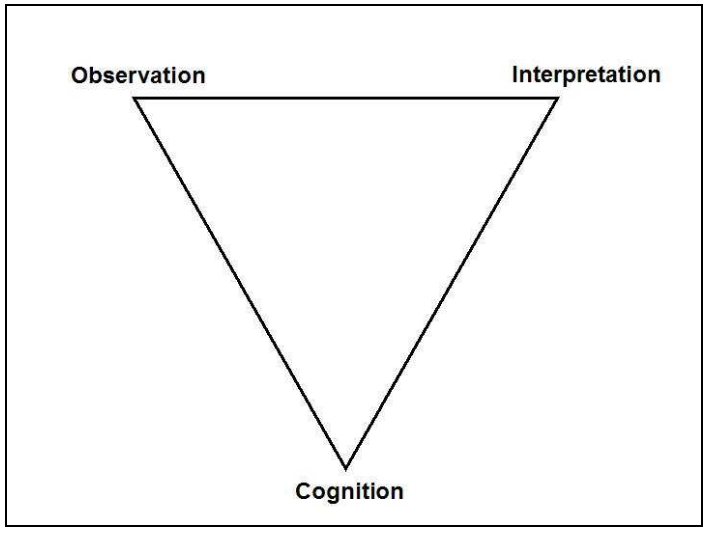


Figure 1. The assessment triangle (NRC, 2001).

The cognition corner of the triangle refers to a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain. In any assessment, a theory of learning in the domain is needed to help identify the set of knowledge and skills that is important to measure. The observation corner of the triangle represents a description or set of specifications for assessment tasks that will elicit illuminating responses from students. The observation aspect of the assessment is based on a set of beliefs about the kinds of tasks or situations that will prompt students to say, do, or create something that demonstrates their important knowledge and skills. The interpretation corner of the triangle encompasses all the methods and tools used to reason from the observations. The interpretation component should express how the observations derived from a set of assessment tasks constitute evidence about the knowledge and skills being assessed (NRC, 2001, pp. 44-48).

Knowing What Students Know (KWSK) (2001) suggests that for any assessment to be effective, the three components of the assessment triangle must be in synchrony. Most commonly, assessment research and development fail to consider the role of cognition in testing (Gorin, 2006; Leighton, 2004). KWSK recommends that assessment design be based on a representation or approximation of cognition that is both consistent with a fully developed psychological perspective of a target idea and at a level of detail sufficient to accomplish the job of assessment (2001, p. 46). KWSK suggests that any model of learning underlying an assessment will be a simplification of what is going on in the mind of the examinee and in the social situation within which the assessment takes place. The goal, therefore, should be to focus the assessment on those competencies that are most important to measure in light of the desired inferences about student learning

(2001, p. 46). For example, KWSK describes extensive research by Griffin and Case (1997) in which they put forth a detailed cognitive theory for how students develop whole-number sense, specify what knowledge is most crucial for early success in mathematics based on the theory, design assessment tasks that elicit relevant student knowledge and skills, and then provide feedback to educators regarding subsequent experiences and instruction necessary to develop competence in the domain (NRC, 2001, p. 179)

Leighton (2004) posits that many assessment developers do not develop such "cognitive models of task performance" because measurement researchers are not trained in cognitive psychology, the process is time consuming, or because verbal reports from students about their thinking are untrustworthy (a claim which has been extensively argued). As a way forward, Leighton summarizes some guidelines for incorporating components of student cognition into assessment development and validation processes and even questions the need for a new breed of psychometrician with training in basic cognitive psychological methods (2004). The use of verbal reports in assessment validation procedures, if done appropriately, would be one example of a cognitive psychological method.

Unfortunately, research that attempts to account for the ways in which students use their ideas in the context of educational measurement is sorely lacking. Existing research that privileges investigations with data of students' actual response methods in the validation procedures (Hamilton, Nussbaum, & Snow, 1997; Harlow & Jones, 2004) tends to be done on a small scale (not by major test developers) and often fails to account for how the evidence of a given student's ideas may shed light on his/her written

assessment responses (Leighton, 2004). These studies of student responses approach questions of construct validity by focusing on very general dimensions of cognition or on a very broad definition of the content areas and fail to carefully examine how specific science target ideas are used in responding to assessment items about related phenomena in science. For example, Hamilton et al. (1997) selected numerous science content items from the National Education Longitudinal Study of 1988 (NELS:88) (astronomy, chemistry, and biology domains) to determine if general cognitive dimensions (quantitative science, spatial-mechanical reasoning, or basic knowledge and reasoning) were used in responding to the questions across the domains. While such an approach may provide a general sense of a student's science abilities, it reveals little about which student ideas were actually elicited by the assessment items. Harlow and Jones (2004) interviewed 38 students to investigate the performance of items from the Third International Math and Science Study (TIMSS) from content areas of earth science, life science, physics, environmental science, and chemistry. However, due to the number of domains involved, the analysis was interested in students' "level of science understanding" and so the grain size allowed them to determine only if, for example, "physics knowledge" was used to respond to a physics assessment item (Harlow & Jones, 2004). Given these broad views of content alignment, these studies are unable to critically interrogate the nature of students' cognition in a single content area (target idea set) to shed light on how specific ideas are elicited by assessment tasks.

Other assessment validity researchers (Borsboom, Mellenbergh, & van Heerden, 2004) call for a narrower, causal model of validity that focuses validation efforts on whether variability in test scores can be explained by the construct of interest. They call

for a view of construct validity that is more narrowly interested in the extent to which the presence or absence of a student's ideas can explain variation in that student's item performance. Still, only a few studies in educational measurement have adopted such an approach to item construct validity (Leighton, 2004).

This study addresses this area of need. It explores the ways in which middle school students express their understanding of forces and motion target ideas when responding to aligned test questions both in writing and during interviews. The study determines if the variability in student responses to paper and pencil tests is related to their understanding of forces and motion as revealed in the videotaped interviews.

Student Learning of Forces and Motion

National content standards documents (AAAS, 1993; NRC, 1996) propose that students should begin to develop the basic conceptual understandings of Newtonian mechanics in the middle grades. For example, *Benchmarks for Science Literacy* states that, upon completing 5th grade, students should know that changes in speed or direction of motion are caused by forces. In addition, students should know that the greater the force is, the greater the change in motion will be, and, the more massive an object is, the less effect a given force will have (1993, p. 89). In addition, grade 6-8 students should be able to understand the basic premises of Newton's First Law of Motion: An object at rest stays that way unless acted on by a force, and …an object in motion will continue to move unabated unless acted on by a force (AAAS, 193, p. 90). The *National Science Education Standards* state that grade 5-8 students should know that an object that is not being subjected to a force will continue to move at a constant speed and in a straight line (NRC, 1996). While some students are successful at making appropriate use of these

concepts, these types of conceptual understandings called for in national standards documents often prove difficult for many students to achieve, as documented by numerous research studies.

Perhaps because of the many experiences that children have with moving objects and the forces that make them go and stop, a substantial body of research on student understanding in this area has developed (Caramazza, McCloskey, & Green, 1981; Champagne et al., 1980; Eryilmaz, 2002; Gunstone & Watts, 1985; Hammer, 1996; Hestenes et al., 1992; McCloskey, 1983; Minstrell, 1982). Among the conclusions drawn from all of these studies is the presence of surprisingly persistent misconceptions that students bring to their learning of forces and motion. In addition, evidence from this body of research shows that, in spite of explicit instruction for high school and college students, many retain their naïve understandings of forces and motion. Other studies have demonstrated the presence of naïve understanding in K-12 science teachers as well (McDermott, Heron, Shaffer, & Stetzer, 2006; McDermott & Redish, 1999). Thus, many individuals, including science teachers, approach and walk away from formal learning about forces and motion persistently holding a number of inaccurate ideas.

An example of a commonly held idea about forces and motion (Chapter II presents a more complete discussion of research on student ideas at the middle school level specifically) is that students tend to think of force as a property of an object ("an object has force," or "force is within an object"), rather than an interaction between objects. Piaget (1972) described this kind of thinking about forces and motion as being 'animistic.' In addition, many students who do think of forces as interactions tend to consider active interactions "forces" (i.e., a ball being kicked) but do not consider passive

interactions as "forces" (i.e., a book on a table) (Gunstone & Watts, 1985). Minstrell (1982) demonstrated students' difficulty in explaining the "at-rest" condition where only those passive forces were acting. As a way to address this issue, *Benchmarks for Science Literacy* supports teaching students to integrate the concept of passive support into the broader concept of force (AAAS, 1993). Students also may believe that constant speed of motion needs some constant cause to sustain it (Clement, 1982) so that if no force is acting, there can be motion. Similarly, they think that if a body is not moving there is no force acting on it. In other words, they do not understand that forces can be acting in opposite directions but be balanced (AAAS, 1993; Gunstone & Watts, 1985). Clement (1982) asserted that the persistence of this thinking is related to students' everyday perceptual-motor experiences with the pushing and pulling of objects. For example, in most everyday situations students will observe that the constant motion of a motor vehicle (or a bicycle) typically requires the constant addition of a forward force without being aware that there are also forces acting in the opposite direction (rolling friction and air resistance).

There is a considerable body of research in this area, but much of the work on student understanding of mechanical physics has been conducted with high school and college students (see McDermott & Redish, 1999). However, it is clear that the persistent, inaccurate conceptions that students bring to their learning at the high school and college levels are steadily being developed during the elementary and middle school years. The development of these conceptions is likely linked to the many experiences that students have had regarding the mechanics of everyday objects in motion (such as the motor vehicle example provided above), as Clement (1982) suggested. Thus, clarifying

understanding of how upper elementary and middle school students think about forces and motion could lead to more successful assessment and instruction at this age level and beyond. So, as students progress to high school and college level science, educators can be more aware of the ways in which student ideas may progress or evolve and therefore design instructional sequences to promote such growth—similar to learning trajectories research in mathematics education proposed by Simon (1995) and Clements and Sarama (2004) and learning progressions research in science education (Smith, Wiser, Anderson, & Krajcik, 2006). That is, by understanding the progression of student ideas at an early age, educators may be more able to design assessment and instructional sequences to help students arrive at a more scientifically accurate understanding of Newtonian mechanics.

Research Questions

Given the need to more carefully consider the call for a detailed understanding of how science ideas are evoked by students while responding to assessment items, the following research questions are addressed in this study:

1. To what extent do the inferences at posttest from each of seven paper and pencil science assessment items (aligned by expert judges using specific alignment criteria to particular forces and motion ideas) agree with the inferences about students' understanding of those ideas as determined during videotaped oral interviews at posttest?
2. During both a written posttest and a videotaped oral interview, to what extent are middle school students in this study able to consistently demonstrate their

understanding of the same forces and motion ideas targeted by a variety of assessment task contexts?

3. Based on the elicited ideas of sixth graders in this study, what can this research add to the literature base (as presented in the study) about common ideas students have regarding forces and motion?

Methods

This study was conducted in the context of a large, six-year, quasi-experimental research study (SCALE-uP) that was designed to explore the effectiveness and scale-up of promising middle school science curriculum materials in a large, diverse school district. The data for this dissertation study were collected by SCALE-uP researchers as part of ongoing assessment validation efforts.

The data analyzed in this dissertation included videotaped structured interviews (collected in the course of SCALE-uP research) with 14 sixth grade students who responded to seven assessment items judged to be aligned to specific ideas about forces and motion from the middle school content standards. (Table 1 shows which students were interviewed for which assessment items.) The seven assessment items of interest were part of an assessment instrument developed as part of the SCALE-uP research and the interviews took place at posttest, after students had a chance to learn about forces and motion ideas for several weeks. Each interview lasted between 20 and 30 minutes, and students were asked to respond to between four and seven assessment items, depending on time constraints. Prior to this dissertation study, students' posttest responses to each written assessment item were rated and a total score was computed for each student by SCALE-uP staff. One of the goals of this dissertation study was to compare students'

Table 1. *Questions Responded to by Students During Interviews.*

Student	Rolling Ball (1 selected and 1 constructed response)	Tennis Ball (1 selected and 1 constructed response)	Skateboard (1 selected response)	Bicycle (1 selected response)	Crane (1 selected response)
Alan	X	X	X		X
Barry	X	X	X	X	
Darlene	X	X	X		
Heidi	X	X	X		X
Ingrid	X	X	X	X	
Janet	X	X	X	X	X
Larry	X	X	X		X
Lisa	X	X			
Opal	X	X	X	X	
Robin	X	X	X		X
Scott	X	X	X	X	X
Teddy	X	X	X	X	
Tom	X	X	X	X	
Wendy	X	X	X		X

previously-scored written posttest responses to the same students' responses when they were interviewed.

Transcribed videotaped interviews were coded to describe students' ideas relative to forces and motion. Code development and procedures were modeled after frameworks suggested by Strauss and Corbin (1990) in which codes are developed in an iterative process involving cycles of refinement, elaboration, and extension of coding categories. Specific, coding of the videotaped interviews began with a set of student idea codes derived from the literature that was tested using a subsample of the data. These codes were improved upon and new codes added until there was consensus about their usefulness for categorizing students' ideas. During this process, the proposed set of codes was insufficient categorizing students' responses to each assessment item, so additional

codes were proposed during the first phases of data analysis. In the end, decisions were made regarding the seven judged-to-be aligned written assessment items' ability to accurately reveal 14 students' ideas about forces and motion. In addition, written and oral evidence was used to determine how consistently the 14 students in the study were able to apply their ideas about forces and motion across all of the assessment item contexts (i.e., falling objects, rolling objects, and objects at rest) during the interview and the written assessment. Of particular interest were instances in which students could use a particular idea to successfully explain one item context but were unable to use the same idea to explain a different item context, or instances in which an accurate response was elicited in one context but not the other. Often, the phenomena to which students have been exposed during instruction are different from the phenomena presented in assessment situations. Answering the second research question will help to gain a better understanding of the ways in which middle school students can or cannot demonstrate their understanding of ideas about forces and motion across a variety of phenomena and item contexts.

The third research question provides an opportunity for this dissertation to comment on what is known in the literature about the learning and assessment of forces and motion ideas at the middle school level.

Significance of the Study

This study makes an important contribution to the science education literature by filling an important gap in how aligned middle school science assessment tasks may be used to ascertain sixth-grade students' understanding of forces and motion ideas. Understanding how students at this age think about forces and motion should be

extremely beneficial in promoting effective instruction in middle school and at higher grade levels. This kind of research on effective assessment of middle school student ideas also has important implications for understanding how ideas are used during specific assessment situations; and how they may eventually lead to appropriate scientific understandings. The results of this research could lead to the development of better assessment instruments in this area of science education, which can be used to more precisely measure student understanding.

Also, the study provides a useful framework for developing an interview protocol for assessment validation research. The videotaped interview data that were analyzed in the study were collected paying careful attention to the ways in which evidence can be gathered to make claims about students' ideas in light of their responses to assessment tasks. In addition, the data analysis methods employed in the study may provide a model for other researchers interested in structured interviews for assessment validation purposes in other domains. If assessment researchers in areas of math, history, or English education are interested in investigations of assessment validation through analysis of student responses, this study contributed a reasonable framework for collecting data to determine how target ideas of interest come to light during a given assessment task.

This study is also significant because the sample of 14 students is highly diverse, which provides a wide range of student responses to the assessment tasks. Because judgments of assessment validity often rely on either adult experts in a given domain (Baxter & Glaser, 1998; Leighton, 2004) or psychometric analyses of patterns in large samples of dichotomously scored data, this study offers an important alternative by moving the level of analysis 'closer' to the actual students—in this case a highly diverse

sample of sixth graders—by focusing on the nature of student ideas that are used by students while responding to assessment items. This approach to validity can have important implications for the ways in which we design assessment situations, monitor student learning, and develop curriculum frameworks and instructional systems in science that can help all students achieve.

In addition, this study can provide evidence of the validity the Project 2061 assessment analysis criteria and the process of using expert panels and specific alignment criteria to produce a set of aligned assessment items for middle school forces and motion ideas.

Assumptions of the Study

This study assumes that ideas about key physical science topics exist in a mental framework for each student as he/she develops understanding in science. These ideas develop and change (Posner et al., 1982; Strike & Posner, 1992) as a result of relevant experiences that students have which may or may not be associated with explicit classroom instruction. Further, such ideas are called upon by students to guide responses to assessment items that are intended to detect the presence of such ideas. If the idea of interest does not exist for a given individual, then it is not evoked as that student engages with the relevant assessment task. Other, perhaps inaccurate, ideas may be evoked by the student instead. In addition, the study assumes that sixth graders' ideas about difficult science concepts like forces and motion may not be particularly stable and a given student's understanding may often be tied to a specific context.

Delimitations of the Study

Delimitations narrow the range of a research study and limit the generalizability of the results (Creswell, 1994). This study is limited in that it involves only a few students. Many more students would be needed to generalize to the larger population of U.S. sixth graders. In addition, the research involved only a few assessment items that target a narrow range of the possible forces and motion ideas. Due to the features of the assessment items under investigation in this study, the forces and motion ideas of interest in this study are the following:

1. Changes in speed or direction are caused by forces.
2. Objects in motion stay in motion unless acted on by a force.
3. Objects at rest stay at rest unless acted on by a force.

These ideas are derived from *Benchmarks for Science Literacy* (AAAS, 1993) and adapted for use by SCALE-uP during the assessment development procedure. Ideally, this research would involve considerably more assessment items targeting a wider range of middle school forces and motion ideas. This delimitation will be discussed as it relates to the potential findings and implications of the study in more detail during Chapter II.

It should also be noted that this research is being conducted with students who just completed an extensive unit on forces and motion ideas. Findings and conclusions from this research should take into consideration the possible effects of immediate prior instruction in the content area. Such instruction may make it more likely that students will have multiple, contradictory ideas and explanations for the phenomena involved.

Limitations

Limitations are described as potential weaknesses in a research study (Creswell, 1994). This study had several weaknesses. Although, the study assumed that a well designed interview protocol, which elicits concurrent and retrospective verbal reports from students, can result in an accurate data source for making judgments about students' understanding of key science concepts, the study recognized that interviewers may not have always asked the questions that most effectively elicited student thinking about key aspects of the construct, or sometimes failed to probe student thinking deeply enough. It was also possible that an interviewer introduced bias into the data by influencing the nature of the student responses. To counter this limitation, efforts were made to carefully observe and reflect on how effectively the interviewer probed student understanding of the target construct. Another limitation was that interviewers did not have much opportunity to develop a high level of rapport with students, which is generally considered to be needed for reliable data collection with interviews. Therefore, it is possible that the low level of familiarity with the interviewer could have negatively affected the completeness of students' responses during the interview. Another limitation of the study was that the assessment items involve a relatively small number of contexts. It is possible that particular assessment contexts offer disadvantages or advantages to some students as they attempt to express their understanding of a given target idea and that the small number of contexts did not provide equal advantage and disadvantage to the sample of students. However, this is a limitation of assessments in general because they can only include a relatively small number of contexts to test each idea.

Definitions of Key Terms

This section contains definitions used in this study. The key terms include the following:

Alignment—a process of ensuring that assessment items measure the student ideas that are of interest to the assessment developer.

Assessment—a systematic, multistep process involving the collection and interpretation of educational data.

Construct—a postulated attribute assumed to be reflected in test performance.

Construct Irrelevant Variance—the inclusion of excess variance that influences measurement of the construct.

Construct Validity –the extent to which an assessment item can be used to draw valid inferences about student understanding of the concept that it claims to measure.

Content Validity—in measurement, how representative the sample of assessment situations is relative to all the possible situations of interest (Cronbach & Meehl, 1955).

Inference—an interpretation or judgment made about student understanding.

Interview Protocol—an instrument used to guide data collection process for clinical interview.

Motion and Forces Assessment (MFA)—instrument developed by SCALE-uP researchers to assess middle school students' understanding of key middle school physical science ideas.

Oral Interview—procedure used to collect data from students regarding their thinking during assessment item responses.

Student Idea— an unobservable knowledge structure in a given domain possessed by an individual that is used to make sense of phenomena in nature.

Target Idea — a student idea that is meant to be assessed by a given test question.

Task Context – an aspect of an assessment task that sets up a particular situation, story, event, or object as a way to cue students (DeBoer, 2005).

Validation—the process of collecting evidence about the adequacy and accuracy of assessment inferences.

Verbal reports – data collected during clinical (oral) interviews to investigate individuals' cognition (Ericsson & Simon, 1993)

Dissertation Overview

This study consists of five chapters. Chapter I has provided a rationale for the study by discussing the topic, context, and background of the problem. This chapter presents the research problem statement, the research questions, the significance of the study, assumptions, delimitations, limitations, and the definition of key terms.

Chapter II provides a theoretical foundation for the study. First, the literature review provides a summary of research on assessment efforts at the middle school level for forces and motion ideas. The chapter also reviews research on students' ideas of forces and motion at the middle school level and offers a framework of ideas to be used in the study. Finally, the literature review summarizes clinical interview methods employed by other researchers who collected verbal reports as data—some for the sake of assessment validation work.

Chapter III summarizes the methodology and the procedures used to answer the research questions. This chapter presents the research design, the theoretical framework of the study, a reiteration of the research questions, the study participants, the instruments used, data analysis procedures, and the consideration of human participants.

Chapter IV presents the data analysis that answers the research questions by focusing on the coding of student interviews. Finally, Chapter V includes interpretations and conclusions about the results presented in Chapter IV. The conclusions are connected to the literature and recommendations are made for future research.

CHAPTER II: REVIEW OF THE LITERATURE

Overview

The purpose of this study was to explore the extent to which a set of seven paper and pencil assessment items, judged to be aligned with specific middle school level forces and motion ideas, accurately measure those ideas. Analysis of videotaped clinical interviews with sixth grade students showed how the ideas middle school students used when responding orally to items compared to the ideas they used when responding in writing to the same items during a posttest.

The following research questions were addressed:

1. To what extent do the inferences at posttest from each of seven paper and pencil science assessment items (aligned by expert judges using specific alignment criteria to particular forces and motion ideas) agree with the inferences about students' understanding of those ideas as determined during videotaped oral interviews at posttest?
2. During both a written posttest and a videotaped oral interview, to what extent are middle school students in this study able to consistently demonstrate their understanding of the same forces and motion ideas targeted by a variety of assessment task contexts?
3. Based on the elicited ideas of sixth graders in this study, what can this research add to the literature base (as presented in the study) about common ideas students have regarding forces and motion?

Organization of the Chapter

This chapter synthesizes the literature relevant to this study and is divided into three sections. The first section of the literature review provides an overview of the methods that researchers have used to assess ideas middle school students have about forces and motion. One of the goals of this study was to explore the validity of assessment items being used to detect student ideas in this domain. Therefore, it is beneficial to know what kinds of assessment techniques have been used in the past to determine student understanding of forces and motion. Research that has proposed methods and/or instruments for assessment of these ideas with middle school students is presented and discussed.

Because the study was interested in students' use of their ideas about forces and motion (both correct and misconceptions) in assessment situations, the second section of the literature review provides an overview of the research into middle school students' ideas about forces and motion. The goal of this section of the review is to synthesize the research on student thinking and provide a framework to guide the coding and analysis of videotaped, transcribed interviews. Thus, this section establishes the relevant ideas that middle school students may have about forces and motion that may be detected by the interviews. It summarizes the major findings from the research on student thinking in this domain. These ideas identified in the literature were ultimately used in coding procedures for determining the ideas that students in this study express in response to test questions. These ideas were then compared to students' ideas as uncovered by the previously scored (by SCALE-uP) written assessment instrument.

The third section of this chapter provides an overview of clinical interview methodologies that have been used to detect student ideas and conceptions. Also, research methodologies were included in which clinical interviews were used to help validate assessment items. This chapter concludes with a presentation of the conceptual framework used in this assessment validation study.

Literature Review Questions

This literature review focuses on the following questions:

1. What assessment instruments/strategies have been used by researchers to measure student understanding of forces and motion at the middle school level?
2. What ideas do middle school students have about forces and motion?
3. How has interviewing been used to investigate student thinking?

As presented in this literature review chapter, a large body of research has emerged that builds a case for the existence of numerous, robust misconceptions, naïve conceptions, alternative frameworks, or intuitive ideas in the area of forces and motion. These student ideas have been thought of as impeding or complicating the acquisition of the canonical or accepted views of scientific knowledge (e.g., Strike & Posner, 1992). Some have speculated that the existence of such ideas prior to, during, and even after instruction represents natural progressions or "learning trajectories" in which students are engaged that could be used to an instructional advantage (Simon, 1995; Smith et al., 2006). While proposing a specific learning trajectory for these ideas goes beyond the scope of this study, the study could have implications for learning trajectory research.

In the context of this research on forces and motion, it is beneficial to reflect on the research at the middle school level as being distinct and often quite different from research that addresses student thinking at the high school or college level. In fact, it could be disadvantageous from the perspective of instructional design to assume that middle school students approach the learning of forces and motion with the same, well-documented ideas that high school students bring to their physics classes. Thus, a review of middle school forces and motion research can help clarify the appropriate distinctions among student thinking across the age levels and lead to a better understanding of the ways in which assessment and instruction could proceed so that students gain a more robust understanding of mechanics from both middle school and high school. It is conceivable that a synthesis of the research at the middle school level could be informative from the perspective of developing an appropriate learning trajectory for the assessment and instruction of forces and motion ideas.

What Assessment Instruments Have Been Used to Measure These Ideas?

Perhaps more than other topics in science education, considerable effort has gone into the development of stand-alone assessment instruments to measure students' developing understanding of force and motion ideas. The goal of this aspect of the literature review will be to answer the question: What assessment instruments/strategies have been used by researchers to measure student understanding of forces and motion at the middle school level?

As discussed by McDermott and Redish (1999), comprehensive instruments to assess student understanding of motion and force have been published at the upper secondary and college levels. These include the Test of Understanding Graphs in

Kinematics (TUG-K), the Force and Motion Conceptual Evaluation (FMCE), and the Mechanics Baseline Test (MBT), which later became the Force Concept Inventory (FCI). Items from the MBT and FCI have been used to develop other instruments, including the Force Achievement Test (FAT) and the Force Misconception Test (FMT) (Eryilmaz, 2002). Each of these instruments is a multiple-choice assessment instrument that is designed primarily for use in determining high school or college students' understanding of Newton's Laws. The instruments have often been used to determine the effectiveness of introductory physics instruction at helping students to develop deeper understanding of the mechanics concepts beyond the abstract problem solving that is common in many physics courses (see, for example, Hestenes et al., 1992). A general conclusion of the research at this level is that high school and college physics students retain strong misconceptions around Newtonian mechanics, despite being able to perform quite well in their physics courses.

Other assessment techniques have been used with younger students, but they have been less widely available and perhaps less reliable. A questionnaire developed by Watts and Zylbersztajn (1981) for 14-year-olds involved a multiple-choice with explanation format. Their questionnaire involved 12 multiple-choice questions always having one choice as the scientifically correct response and at least one of the other choices inspired by an alternative framework described in the literature (Watts & Zylbersztajn, 1981). The contexts for the items included a man dropping a wrench on the moon, two equal mass objects suspended from a pulley, a cannon ball in flight, a game of tug of war, and a car on an incline. The authors are not specific as to what knowledge is being targeted by their questionnaire except to say that they are interested in the prevalence of non-Newtonian

views of forces and motion. While their questionnaire has been used to assess student understanding in this area, response data has not been compiled in a way to lend more authoritative credibility to the validity or reliability of the instrument (i.e., the psychometric analyses that are more common with the FMCE or FCI). In other words, they seem to rely on expert judgment of content alignment to conclude that the questionnaire items are indeed measuring the prevalence of non-Newtonian reasoning in students. This particular questionnaire is not widely used and seems more suited for small-scale, qualitative investigations of student ideas. For example, Park and Han (2002) conducted one study in which it was used in research with 27 seventh grade students in Korea.

Research by Osborne and Gilbert (1980) involved the development and use of "interviews-about-instances" cards. They developed cards that depicted four everyday instances in which various forces (i.e., pushes, gravity, friction) are acting on both stationary and moving objects (cars, bicycles, golf balls, people). The items can be used as prompts or guides for qualitative and exploratory interviews with students about forces and motion ideas.

In other research by Watts (1982b), elementary, middle, and undergraduate students were asked to look at cards with "some simple line drawings depicting various situations where the concept of force may or may not be applicable" (p. 116). These drawings include situations involving a golf ball in flight, a held helium balloon, a diver on a diving board, and a book on a table. Students are asked the same question each time, "Are there any forces here?" The cards can be used to uncover student thinking about

various kinds of forces and to prompt discussions about students' explanations for natural phenomena.

In addition to these efforts by researchers, other research at the middle school level has involved the use of computer simulations of scientific models (Weller, 1995). Through a series of pilot studies and interviews with students, Weller developed assessment items designed to measure "three particular naïve models of the natural world that are resistant to traditional methods of teaching" (p. 271). Weller developed and used the same eight assessment items at pretest, posttest, and delayed posttest as part of an investigation to determine the effectiveness of computer simulations for remediation of misconceptions. The items asked students to consider scenarios of steam rising, a spoon falling, an anvil and can falling, a cup sliding, a bicycle slowing, and a softball and brick falling. Seven of the eight questions asked students to evaluate only two answer choices (one correct and one naïve model). This could help to explain the poor inter-item correlations that Weller found. It is likely that error due to guessing could have been a significant source of variability that contributed to the low correlations.

Montanero, Suero, Perez, and Pardo (2002) investigated elementary, middle, and high school students' use of the concept of force when applied to static bodies in contact as compared to two bodies in motion (collisions). They developed and used an instrument that depicted four sets of animate and inanimate objects in positions of static contact and asked students to select the answer that best described how the forces were being exerted on each object. Their research involved analyzing data from 12 items. Inter-item correlations with the items suggested a stable factor structure that was supported by evidence from student reasoning.

Ioannides and Vosniadou (2002), as part of research around the development of force ideas, conducted individual interviews with 105 Greek students ranging in age from 4 to 15 years. Students were shown 27 drawings of physical objects in combinations of different sizes and kinetic states and were asked to determine which forces were being exerted on the objects, if any. Each of the drawings was accompanied by the question, "Is there a force on x? Why?" which initiated a verbal explanation from the students for the purposes of the interviews. The 27 items were grouped into five item-types: stationary objects, stationary objects pushed by a human agent, stationary objects being on top of a hill, objects in a free fall, and objects that have been thrown. The research involved interviews with students to learn more about the development of the force concept in children. Thus, the 27 questions (drawings) were not intended to stand alone as an assessment instrument but, rather, to serve as prompts for the interviews. Vosniadou, Ioannides, Dimitrakopoulou, and Papademetriou (2001) used similar drawings in previous research with fifth and sixth grade students to determine the effectiveness of research-based learning environments in teaching students about mechanics within an experimental research design.

Clearly, assessment development efforts have resulted in some items targeting forces and motion ideas for middle-school-aged students. However, many of these efforts have been used as part of interview research methodologies. The focus has not been necessarily on the development of a comprehensive assessment instrument but, rather, more explicitly focused on providing data for qualitative descriptions of the nature of students' conceptions. This is a limitation for the field because, despite our growing understanding of how students think about forces and motion, there is a limited

understanding about how to effectively determine their ideas about forces and motion. Aside from the research presented above, there are a few other assessment development efforts worth mentioning.

One source of science assessment items has been the National Assessment Educational Progress (NAEP) released items. Education researchers have commonly used such items in mathematics effectiveness studies (Confrey & Stohl, 2004). According to the NAEP Web site (http://nces.ed.gov/nationsreportcard/ITMRLS/), NAEP released items in the forces and motion content area are aligned to content areas that address "an understanding of frames of reference, force and changes in position and motion, action and reaction." For example, NAEP currently makes three items available (2 multiple choice and 1 short answer) that target each of the above content areas. It is unclear and difficult to determine the extent to which researchers or educators make use of such released items from NAEP (or TIMSS) assessment databases to measure student understanding of force and motion ideas. However, given the scant item specifications available to potential users (i.e., vague content standards), it would be difficult to determine what students did or did not know if they responded correctly or incorrectly to these items.

A group at Harvard University led by Philip Sadler has undertaken another assessment development effort. They developed a series of assessment instruments called Misconceptions-Oriented Standards-Based Assessment Resources for Teachers (MOSART) with over 1,000 science items spread across grades K through 12 (http://www.cfa.harvard.edu/smgphp/mosart/index.html). One of the 10 assessment instruments is a 20-item instrument for Grades 5-8 physical science that covers topics

such as light, heat, magnetism, electricity, energy, and motion. Relative to assessment of force and motion ideas, this particular MOSART instrument includes three items that target the following two content standards:

1. The motion of an object can be described by its position, direction of motion, and speed. That motion can be measured and represented on a graph (two multiple-choice items).
2. If more than one force acts on an object along a straight line, then the forces will reinforce or cancel one another, depending on their direction and magnitude. Unbalanced forces will cause changes in the speed or direction of an object's motion (one multiple-choice item).

In this particular case, it is unclear if three assessment items could provide an educator enough information about what students did or did not know about the two dense content standards listed above. However, given the focus on misconceptions in the development of the items, perhaps there is useful information to be gained from these items. There are other physical science MOSART assessment instruments at the K-4 and 9-12 grade levels with similar numbers of items targeting forces and motion.

Recent assessment development efforts by AAAS Project 2061 (see, for example, Abell & DeBoer, 2007; DeBoer et al., 2007; Gogos & DeBoer, 2007) have resulted in the assembly of an online collection of student assessment items in middle school science that are aligned to national content standards. Though specific forces and motion items are being developed by Project 2061, none have been released to date. Once complete, this resource could potentially be helpful for educators and researchers interested in detection of students' ideas in this area of physical science.

In addition to these efforts, SCALE-uP at GWU has developed a Motion and Forces Assessment (MFA) that has been used at the sixth grade level with thousands of students over the past several years (Pyke & Ochsendorf, 2006). Details about this assessment will be discussed in more detail in Chapter III, as it is an integral part of this study.

In summary, this section established the existence of several comprehensive assessment instruments at the high school level, but concludes that there has been a limited effort to develop such comprehensive instruments that assess forces and motion understanding at lower grade levels (i.e., 6-8). Despite a wealth of research on forces and motion at the middle grades and earlier, it is apparent that few systematic efforts have been made to develop forces and motion assessment instruments at those lower grade levels. Working toward the development of effective assessment instruments in this area might be extremely productive in monitoring and promoting effective teaching and learning of this difficult physical science topic.

What Ideas Do Middle School Students Have About Forces and Motion?

In this portion of the literature review, the existing literature on research on middle school students' ideas about forces and motion is presented. A systematic search of the Education Resources Information Center (ERIC) database for research on middle school students' understanding of motion, force, mechanics, and other relevant subject headings was completed to find out what ideas middle school students have about forces and motion?

The goal of this literature review will be to focus on systematic investigations into the learning of forces and motion ideas by middle school students. A staggering amount

of research on physics education has been conducted at the upper secondary level and introductory college level (see McDermott & Redish, 1999, for a thorough review of the literature in this area). There has been considerably less research surrounding younger students' understanding of force and motion ideas in physics education research. This summary was based on a number of syntheses of research at this level (AAAS, 1993; Driver, Guesne, & Tiberghien, 1985) as well as a review of the literature cited in the ERIC system that was conducted between 1965 and April 2007.

The following paragraph details how the ERIC literature search was conducted to locate existing research on middle school student ideas in forces and motion. First, the subject headings "Mechanics (Physics)," "Physics," and "Motion" were entered as search terms, resulting in 14,598 references. These initial search terms were "exploded" such that references including other headings that "used for," "related" or "broader" than the above terms were also included in the results. The use of the term "force" complicated the search process by including references pertaining to "labor forces" and other kinds of forces that are not relevant to physical science education, so it was excluded. Next, these 14,598 references were further narrowed by limiting the search to only those references involving "Secondary Education," resulting in a total of 2,945 references. Next, these 2,945 references were limited to only publications that were journal articles, which left 1,841 references. Finally, these references were limited by exploding the subject heading "Concept Formation." This resulted in the inclusion of broader, narrower, and related headings that included the following subject headings: learning processes, concept development, cognitive development, cognitive structures, comprehension, concept teaching, concept mapping, constructivism, developmental stages, fundamental concepts,

knowledge representation, and misconceptions. Thus, a reference could contain any one of the above headings and, therefore, be included in the search results. A total of 51,777 references included at least one of the above subject headings. By looking at common references between these 51,777 and the 1,841 references described above, a total of 248 journal article references remained. The list of references was further narrowed by removing articles pertaining to "light," "optics," or "heat." Also, articles that included the subject heading "High Schools" were eliminated. This left 188 journal articles. Abstracts were then read to determine relevance to the current literature review. Articles were determined to be relevant if they focused, at least in part, on students' conceptions of motion, force, or mechanics and included subjects at the middle school level (ages 12 to 15). Of the 188 abstracts reviewed, 46 journal articles were selected because of their explicit focus on force, motion, or mechanics (as opposed to other aspects of physical science such as chemistry, thermodynamics, or electricity) and their potential for inclusion of students at the middle school level. Many articles were not selected because the subjects in the studies were older than the interest of this particular research (middle-school-aged-students). Upon closer inspection of these 46 journal articles, it was apparent that many still did not involve research with students at the middle school level. Upon reviewing all 46 journal articles, 17 journal articles remained that explicitly focused on forces and motion research at the middle school level. These studies are listed in Table 2. From this search of the literature, it became apparent that the vast majority of research in physics education has taken place at the upper secondary level (i.e., high school juniors and seniors) and at the college level (predominantly with introductory physics courses).

Table 2. *Journal Articles Resulting from ERIC Search on Forces and Motion.*

Author(s) (year).
1. Gamble, R. (1989).
2. Gorsky, P., & Finegold, M. (1992).
3. Ioannides, C., & Vosniadou, S. (2002).
4. Kibble, B. (2006).
5. Larrabee, T. G., Stein, M., & Barman, C. (2006).
6. Levin, I., Siegler, R. S., & Druyan, S. (1990).
7. Marioni, C. (1989).
8. Montanero, M., Suero, M. I., Perez, A. L., & Pardo, P. J. (2002).
9. Osborne, R. (1984).
10. Osborne, R. J., & Gilbert, J. K. (1980).
11. Palmer D. H. & Flanagan, R. B. (1997).
12. Park, J., & Han, S. (2002).
13. Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000).
14. Watts, D. M. (1982).
15. Watts, D. M. (1983).
16. Watts, D. M. & Zylbersztajn, A. (1981).
17. Weller, H. G. (1995).

In presenting the research articles found at the middle school level, this review focuses on findings of student ideas related to the target ideas of interest in this particular study. During development of the MFA instrument, SCALE-uP selected and targeted for assessment a number of middle school forces and motion ideas from *Benchmarks for Science Literacy* (AAAS, 1993). These ideas were broken down into sub-ideas for assessment purposes. The items selected for this dissertation study target the following subset of ideas assessed by the MFA:

1. Changes in speed or direction are caused by forces.
2. Objects in motion stay in motion unless acted on by a force.
3. Objects at rest stay at rest unless acted on by a force.

This portion of the literature review of student ideas first reports and summarizes existing research relative to middle school forces and motion ideas in general. The review includes the various kinds of motion that can be observed (i.e., sliding, rolling, flying, or falling) and the various kinds of forces that are often appropriate for this age level (i.e., pushes, pulls, friction, or gravity). The literature review includes research on student ideas about gravity, given the popularity of research on this topic. Research on common student ideas related to each of the above target ideas is reported, discussed, and summarized in a table for use in the data analysis portion of this dissertation. Finally, the scope of the relevant student ideas will be narrowed to include only those ideas that are likely to be elicited by the assessment items involved in this study, which are discussed in more detail in Chapter III.

As the research is presented from these studies, supporting evidence is drawn from a frequently cited chapter in Driver's *Children's Ideas in Science* (Gunstone & Watts, 1985) and from supporting text statements on research from *Benchmarks for Science Literacy* (AAAS, 1993). In addition, Project 2061 has made available comprehensive internal reports that summarize findings from assessment item piloting on forces and motion target ideas of interest in this study. Findings from these Project 2061 reports are discussed as well.

To begin, research conducted by Jean Piaget (1972) is discussed briefly to provide some historical context for research on students' ideas about forces and motion. Piaget

proposed that understanding of the concept of force develops through stages. Piaget proposed that children equate force with movement or speed during concrete stages of development. Piaget also proposed that older children (9 to 12 years) are able to differentiate movement and force and understand that force represents the cause of the movement or of its changes. However, as pointed out by Ioannides and Vosniadou (2002), a considerable amount of research has been conducted since then that suggests that older students and even many adults are not capable of drawing the scientifically accurate distinctions between movement and force or identifying force as a causal factor in motion.

In the following sections, the research on student ideas is presented relative to each of the target ideas of interest.

Target Idea 1: Changes in Speed or Direction are Caused by Forces

It is often not difficult to get younger students to "understand the effect of constant forces to speed up (acceleration), slow down (deceleration), or change the direction of an object" (AAAS, 1993, p. 339). However, a significant challenge for many younger students seems to be learning how to correctly use the word "force" to describe how forces influence speed and direction. Osborne and Gilbert (1980) interviewed 40 students and young adults ages 7 to 19 about the concept of force using their interview-about-instances technique. They found that several 11 and 13-year-old students possessed anthropocentric and egocentric views of force (i.e., rolling balls *have* a force). They surmised that our everyday use of the word force as it relates to human action tends to reinforce these kinds of views. Kibble (2006) points out that the word force has a currency in everyday dialogue when used in ways like "you can't force me to do that" or

"this is a forceful argument" (p. 228). Students face the daunting task of attempting to use the word force in appropriate ways, and national standards documents have recognized this difficulty (AAAS, 1993).

Related to the language difficulties, Osborne and Gilbert (1980) noted that younger students sometimes endow objects with feeling or purpose (i.e., the ball could be trying to stop). In addition, they found that many 11 and 13-year-old students tend to endow an object with a certain amount of a physical quantity (i.e., force, momentum, velocity); and, when the quantity runs out, the object stops.

In an assessment item targeting the related idea that when unbalanced forces act on objects in the direction of motion the object's speed increases, Project 2061 (2006) found that 63% of sixth grade students (n=24) understood that a falling ball's speed increases because the force of gravity is stronger than the force of air resistance, and 50% of students (n=14) understood that if an object's speed is increasing, there must be an unbalanced force acting on the object in the direction of motion. Only 19% of sixth graders (n=26), however, understood that if two forces (wind and water) act on a sailboat in opposite directions and the larger force is in the same direction as the sailboat's motion, the sailboat's speed will increase.

Related to this target idea, they summarized the misconceptions in the following way:

> Some students thought that when there is an unbalanced force acting on an object in the same direction that the object is moving, the object will move at constant speed. Others thought that the object will speed up and then slow down, or speed up and then move at a new, higher constant speed. (Project 2061, 2006)

In connection with the idea that when an unbalanced force acts on an object in the opposite direction of motion the object's speed decreases, Project 2061 found that 73% of sixth graders understood that the speed of a ball kicked upward decreases because the force of gravity is in the direction opposite to the ball's motion. On the other hand, only 27% of students recognized that as a box slides across the floor, its speed decreases because the force of friction is in the direction opposite to its direction of motion. In regard to this target idea, Project 2061 summarized the common misconceptions:

> It is difficult for some students to recognize when friction is acting, and other students who do recognize when friction is acting do not identify friction as a force. Thus many students do not explain slowing down in terms of a force of friction acting on an object, thinking instead (incorrectly) that there is a "force of motion" on (or in) a moving object that "runs out" as the object slows down. (p. 6)

This finding of students' ideas that slowing objects have less force or are running out of force is related to the common misconception that "the amount of motion is proportional to the amount force," reported by Gunstone and Watts (1985).

A challenge for many students in developing understanding of forces and motion seems to be recognizing the force of gravity and its effect on objects both at rest and in motion. Many students have trouble understanding when gravity acts and on which objects under what conditions. Watts and Zylbersztajn (1981) found that for objects thrown in the air, some students believed that gravity "turned on" at the peak height of the throw; others thought that gravity acted only as objects move downward, and some thought correctly that the force of gravity acted all the time. Park and Han (2002) found that 63% of Korean eighth graders believed that for objects thrown in the air there was an upward and downward force, while 27% of students believed that there was only an upward force. (Gorsky and Finegold [1992] also observed this idea about upward forces.)

Only 10% correctly believed there was only the downward force of gravity acting. Watts and Zylbersztajn also found that some students believe gravity increases with height. In general, they concluded, "14-year–olds' language lacked precision and their words held global, generous meaning. . . Terms like potential energy, power, energy, gravitational force, weight and force were all used in different guises to support their answers" (p. 365).

Subsequent research on student ideas about gravity (Watts, 1982a) reports several common misconceptions related to gravity. Watts describes the following eight misconceptions (termed "frameworks") regarding gravity:

1. Gravity is a force that requires a medium (i.e., air) to act.
2. Where there is no air (i.e., the moon), there is no gravity.
3. Gravity increases with height.
4. Gravity is constant—moving objects try, and fail, to 'counteract' gravity.
5. Gravity begins to operate when objects start to fall down and continues to until they are at rest on the ground.
6. Gravity is selective—it does not act on all things in the same way, nor on the same things in the same way at all times.
7. Gravity is a large force.
8. Gravity is not weight.

Watts emphasizes that the importance is not that younger students misunderstand gravity, but that the focus should be on exploring the ways students try to understand gravity. Watts' work attempted to bring to light all the possible ways in which students were "trying" to make sense of gravity as a force.

Marioni (1989) reports several misconceptions when 12 and 13-year-olds are asked when a runner should release a marble for it to land in a cup on the ground. The study found that students believed that objects must always fall vertically, heavier objects have a greater tendency to fall vertically, and that students have difficulty adopting an appropriate frame of reference for describing motion.

Ioannides and Vosniadou (2002) used the 27-item questionnaire discussed previously to ask students about their understanding of force. For free-falling objects (large and small stones, large and small balloons), they reported that 60% of sixth graders believed there was a force acting on all the objects due to the object's motion; 17% believed that a force acted only on the large and small stones because they were heavy; 7% believed a force acted on all objects because they had weight; and only 3% believed that the force of gravity acted on all the objects. For the same objects thrown by a man, 83% believed that the force from the man was still on all the thrown objects; 10% believed that there was only a force on the large and small stones; and 3% believed that, once thrown, no force was acting on any of the objects. When comparing the forces acting on a falling stone and a stationary stone, 50% believed that there was only a force acting on the falling stone; 37% believed that there was a greater force acting on the falling stone; and 7% believed there was only a force (or a greater force) acting on the stationary stone.

When considering rolling objects, middle school students often have difficulty understanding how friction forces act to change the speed of rolling objects. Students tend to believe that objects slow because the forces in them "run out" or the object "loses speed." They also will attribute the slowing to there not being enough force or energy

from the initial push. In assessment item piloting, Project 2061 (2006) reported this "decreased motion implies decreased force" misconception as follows:

> A number of students associated slowing down with decreasing force. Some students did not explain slowing down in terms of a force of friction acting on an object, thinking instead that there is a "force of motion" on (or in) a moving object that "runs out" as the object slows down. Alternatively, they thought that an object moves because of the lingering influence of the force that set it in motion, such as the "force of the shove," and that it slows down because the influence or strength of this force decreases. (p. 5)

There are some commonalities that can be found in the research on forces and motion ideas as they relate to Target Idea 1. Most commonly, students tend to associate force with motion as opposed to considering the more accurate idea that force is associated with a change in motion. Perhaps the idea of a "changing motion" (i.e., acceleration and deceleration) resulting from a force acting is not intuitive for younger students.

Target Idea 2: Objects in Motion Stay in Motion Unless Acted on by a Force

This premise of Newton's First Law is often complicated for adults to understand, let alone middle school students. One of the major difficulties for students seems to be their belief that for objects to stay in motion, they require the acting of a continuous force. They are not likely to believe that objects in motion will stay in motion because of the difficulty in imagining a frictionless environment. The difficulty in understanding this target idea lies in the persistent misconception that a force must continue to act on an object if it is to continue in motion even under (simulated) friction-free conditions (Gunstone & Watts, 1985).

Weller's study (1995) with 31 eighth graders involved structured interviews with the goal of diagnosing the prevalence of Aristotelian alternative conceptions. Weller

reported that many students believed that if an object is moving and no unbalanced force is acting on it, the object would come to rest.

Developing a full and robust understanding of this target idea seems to be complicated by students' persistent misconceptions about the nature of forces. Namely, many students believe that moving objects *have* forces within them or that forces act in the direction of motion. Osborne and Freyberg (1985) found that more than half of the students at age 13 believed that a force still existed in a golf ball in flight acting in the direction of its motion. The cause of this force was usually attributed to the original mover (i.e., the golfer with the club).

In research with 14-year-old students in the United Kingdom, Watts and Zylbersztajn (1981) found that about 85% of students associated force with motion. They summarized the misconception as "If a body is moving, there is a net force acting upon it in the direction of movement. . . . . If a body is not moving, there is no force acting on it" (1981, p. 361). Palmer and Flanagan (1997) found that about 66% of 12-year-olds in Australia held the motion-implies-force misconception. Watts and Zylbersztajn reported that a "considerable number of children thought that as the movement decreases, so the force decreases" (p. 363).

Project 2061 (2006) reported a number of "motion implies force" misconceptions in assessment item piloting with sixth graders.

> Students often do not understand that all forces are interactions between two objects. Therefore, when students consider a moving hockey puck (that is no longer touching the hockey stick) they (1) reject the correct view that there are no horizontal forces acting on it, and (2) suggest incorrect forces, such as a 'force of motion,' acting on the puck. (p. 2)

It seems clear that this "motion implies force" misconception is a common idea for many middle school students that may influence their readiness to adopt the idea that objects in motion will stay in motion unless acted on by a force.

Target Idea 3: Objects at Rest Stay at Rest Unless Acted on by a Force

Many students have little trouble believing that a stationary object will remain stationary unless pushed or pulled in some way. The difficulty lies in getting students to appreciate the balanced nature of forces in the "at rest" condition. That is, in order to develop a full understanding of this target idea, students should appreciate that even for objects at rest, there are always at least two forces involved (the force of gravity and the passive supporting force of a table, floor, or ground).

In research with 30 sixth grade Greek students, Ioannides and Vosniadou (2002) asked students about the forces acting on stationary objects (large and small stones, large and small balloons). They found that 47% of sixth graders believed that there were no forces acting on stationary objects because nobody was pushing on them. Gorsky and Finegold (1992) reported similar student ideas for objects in the "at-rest" condition.) Ioannides and Vosniadou found that 17% of sixth graders believed that a force was acting on all stationary objects because they have weight, while 7% believed that the force of gravity acted on all stationary objects. However, 13% of students believed that forces only acted on heavy stationary objects (stones) and not on light stationary objects (balloons).

The relationship between stationary objects and forces is complicated by the equating of force and motion discussed previously. Because of this, students will also believe that if an object is not moving, there are no forces acting. For example, in

research by Osborne and Freyberg (1985), when New Zealand students ages 7 to 19 were asked if there was a force on a stationary car being pushed unsuccessfully by a man, many students believed that there was no force acting (i.e., no movement = no force acting). When Ioannides and Vosniadou (2002) asked students to consider stationary objects (large and small stones, large and small balloons) being pushed by a human agent (but not moving), 60% of sixth graders believed that there was a force on all the objects because the man was pushing them. Some students (13%) believed that a force was acting on all the objects because they had weight, and another 13% believed that a force was only acting on the large and small stones being pushed (and not acting on the large and small balloons).

For stationary objects on top of a hill, Ioannides and Vosniadou (2002) reported that 17% of students believed that no force was acting on any object because no one was pushing them; 27% believed that no force was acting on any object because they were not moving; 20% believed there was a force acting on all objects because they had weight; and only 3% of students believed that the force of gravity was acting on all objects. In situations involving stationary objects (large and small stones) being pushed by people (man and child), 33% of students believed that the magnitude of the force exerted was related to the size of the object; 37% believed that no force was acting on any stone because they were not moving; and 23% believed that the force from the human agent was related to his effort and supply of force.

Kibble (2006) summarized that for many students, “Bodies in equilibrium don’t seem to have any obvious connection with forces” (p. 229). A bottle in the at-rest position on a table is not moving, so, therefore, must not have any forces acting on it. To

apprentice students to understand that in such situations, there are at least two forces acting on an object is a very difficult proposition. The idea of forces in equilibrium resulting in no motion or constant motion is important to helping students arrive at an understanding of Newton's First Law.

Summary

In summary, it is important for the purposes of this study to condense the findings from the research just presented around these three target ideas. This collection of student ideas from the literature served as a starting point for initial code development for analysis of the videotaped interviews.

In conclusion, Table 3 presents the student ideas that were uncovered from the existing literature on forces and motion at the middle school level. While the student ideas presented in the table are representative of findings in the literature on forces and motion ideas, not all are relevant to this dissertation study. Based on the features and intended alignment of the items under investigation (discussed in detail in Chapter III), several of the common student ideas shown in Table 3 will not be helpful in categorizing and making sense of student thinking during the interviews. Therefore, the ideas shown in the table will be narrowed to include only those that are most relevant, given the focus of the assessment items under investigation.

MacKenzie Stetzer, a Research Assistant Professor and subject-matter expert (SME) at the Physics Education Group in the Department of Physics at the University of Washington, reviewed, for accuracy and completeness, these 13 common student ideas from the research literature on forces and motion. She evaluated the literature review and provided feedback on the summary of research based on her knowledge of physics

Table 3. *A List of Common Middle School Forces and Motion Misconceptions.*

Common Forces and Motion Misconceptions based on Literature Review
1. Inanimate objects try to stop (or keep going) on their own accord (Osborne & Gilbert, 1980)
2. Students associate amount of motion with amount of force (Gunstone & Watts, 1985; Ioannides & Vosniadou, 2002)
3. Constant force results in constant motion (Gunstone & Watts, 1985; Project 2061).
4. Decreasing motion (slowing down) is the result of decreased force or force running out (Gunstone & Watts, 1985; Project 2061) as opposed to a force (i.e., friction) in the opposite direction (Project 2061).
5. Forces must act if objects are to stay in motion (Gunstone & Watts, 1985; Weller, 1995)
6. Objects in motion *have* a force in the direction of their motion (Osborne & Freyberg, 1985, Reiner, Slotta, Chi, & Resnick, 2000).
7. If a body is not moving, there is no force acting on it (Gorsky & Finegold, 1992; Ioannides & Vosniadou, 2002; Osborne & Freyberg, 1985; Watts & Zylbersztajn, 1981).
8. Gravity turns-on or only acts when objects are falling (Watts & Zylbersztajn)
9. Rising objects experience an upward force only (Park & Han).
10. Gravity changes (increases/decreases) with height or is selective (Watts & Zylbersztajn)
11. Objects must always fall vertically (Marioni, 1989).
12. Forces (like gravity) act only on (heavy) objects because they have weight (Ioannides & Vosniadou, 2002)
13. Force is the acquired property of objects that move (Ioannides & Vosniadou, 2002; Project 2061, 2006; Reiner, Slotta, Chi, & Resnick, 2000).

education research at the middle school level. From this feedback, the categories of student ideas were revised to incorporate suggestions (e.g., two codes were consolidated).

Many of the studies reported on in the above section of the literature review involved investigative, qualitative research with students in which the authors were interested in exploring and reporting on the range of student ideas around forces and motion. That is, no prior claims were made about the quality of the assessment

instruments used to assess students' ideas. One exception to this may be the summary report authored by Project 2061. Their observations were based on students' answer choice selections and written comments on items that were under development. The items had gone through several quality control stages, but were not yet finalized.

It is evident that in the past few decades, researchers have developed a sophisticated understanding of the ways in which students express their understanding of forces and motion. Despite a wealth of research on student learning in this domain, no studies, with the possible exception of the Project 2061 work, were found that investigated how forces and motion ideas do or do not get used by students responding to judged-to-be-aligned assessment tasks.

How Has Interviewing Been Used to Investigate Student Thinking?

This section of the literature review addresses clinical interviewing methods from a broad perspective, but also focuses on studies that have been conducted that involved interviewing for the purpose of assessment instrument validation. There is a considerable amount of research that investigates the nature of students' ideas and cognitive processes. Naturally, the characteristics and design of a given research study can affect the findings and conclusions drawn by the researcher. Some of the factors affecting the outcomes include the nature of the instrument used to assess understanding, the degree of interaction between student and investigator, the depth of probing, the form of the data, the physical setting, the time frame, and the goals of the investigator (McDermott, 1984). These characteristics are addressed in this section of the literature review to provide a complete picture regarding what is known from the literature about conducting interviews to reveal important aspects of an individual's cognition.

Interviewing as a research methodology for data collection has a long history across many disciplines. Alfred Binet was an early pioneer in interviewing to determine the ways in which individuals responded to intelligence tests. His research led to a greater understanding of individual differences in mental performance, and he later developed tests to identify areas of cognitive weakness in children. Ultimately, his work led to the Stanford-Binet intelligence test, which is widely used today. Clinical interviewing was further utilized by a Binet protégé, Jean Piaget, in the 1920s; and his work in this area continued throughout his career (see, for example, Piaget, 1928, 1952). Piaget (1983) was very much interested in children's progressions in their explanations of natural phenomena (i.e., conservation, quantity, motion). Data collected through his use of structured, clinical interviews lent support to theories surrounding his four stages of cognitive development in children (sensory motor, preoperational, concrete operational, and formal reasoning).

Since the days of Piaget's clinical interviews, interview protocols have been adapted to suit the needs of many in the fields of education (see, for example, Novak & Gowin, 1984; Osborne & Freyberg, 1985) and psychology (Ericsson & Simon, 1980; 1993). In referring to interviews as data sources in research, Clement (2000) writes,

> Their strengths, in comparison to nonclinical, data-gathering techniques, include the ability to collect and analyze data on mental processes at the level of a subject's authentic ideas and meanings, and to expose hidden structures and processes in the subject's thinking that could not be detected by less open-ended techniques. (p. 547)

Ericsson and Simon (1993) refer to the data collected in such interviews as verbal reports and defend such reports as being valid for studies on human cognition. While some have questioned the trustworthiness of clinical interview data (see, for example, Nisbett &

Wilson, 1977; Wegner, 2002), Ericsson and Simon have addressed concerns about the data collection methods and offer important guidelines for those interested in collection of this kind of data. These guidelines proved important in designing the data collection methods for this study. They will be discussed in more detail below.

Currently, verbal reports are classified into two general areas—retrospective and concurrent (Ericsson & Simon, 1993). Retrospective verbal reports involve students reflecting on particular events or instances of their cognition and explaining their reasoning around their remembered thoughts or actions. According to Ericsson and Simon, retrospective verbal reports as methods of data collection can be reliable, provided the temporal separation between thought occurrence and thought reporting is minimized—no more than 10-15 seconds. If the elapsed time is longer than that, the concern is that subjects are not always capable of accurate retrieval of reports on cognitive processes from long-term memory (LTM), as their verbalization is likely to be influenced by other similar memory structures, thus compromising the validity of the subjects' verbal reports about the instance in question (as discussed in Hamilton et al., 1997). So, it is considered more reliable to collect retrospective verbal reports immediately following the cognitive processes of interest as opposed to asking individuals to recall their thought processes from some previous time period.

Nisbett and Wilson (1977) found that at the time of retrospective verbal reports made a long time after the event (sometimes years), individuals were not basing their answers on memory of the specific events but, rather, theorizing about their processes. Therefore, the data collected would be less a reflection of cognition at the time of the task and more of an explanation or justification of their actions that were generated at that

time. Also, whether or not an individual calls upon LTM to explain cognitive processes can depend greatly on the instructions given during the interview. Often, it is possible to answer the interviewer's question without accessing any information from LTM. Considerable attention must be given to the ways in which questions are asked to promote the use of LTM, leading to accurate disclosure of the thought processes during retrospective kinds of studies (Nisbett & Wilson, 1977).

Concurrent verbal reports are typically considered more valid and reliable, and one accepted method for acquiring such data is through 'think-aloud' strategies with individuals. Students are asked to think aloud (TA) about the task they are performing as direct evidence of their cognitive processes related to that particular task. It is generally considered reliable despite some criticism that TA can alter and impede the sequence and content of the thought processes (Flaherty, 1974). Ericsson and Simon (1993) report several strategies to minimize any impacts of interviewer behavior on thought processes due to TA, including decreased attempts at social interaction, limited or no explanations/ descriptions of the phenomena of interest on the part of interviewer, and ample opportunities for training of interviewers to accurately elicit concurrent verbal reports from subjects using well-designed practice situations.

Generally, it is considered sound methodology when a well-designed mix of both retrospective (obtained immediately) and concurrent verbalizations is used to collect data from subjects (Ericsson & Simon, 1993). By accomplishing this, there is minimal influence from the confounding factors of time and accurate thought retrieval from long-term memory. It is this mixed approach that provided a foundation for the development of the SCALE-uP interview protocol that was used for this study. In earlier versions of

protocol development, SCALE-uP attempted to elicit concurrent reports from sixth and seventh grade students, but found that many students were unable to fluidly describe their current thought processes. Despite repeated efforts to stimulate TA with middle schoolers, interviewed students tended to be most comfortable responding to assessment questions silently and then describing their thought processes immediately following their responses (i.e., 5-10 seconds after). Similar findings with middle school students in the domain of science have been reported elsewhere (DeBarger, 2006). During the clinical interviews used to collect data in this study, students were asked to respond to a given assessment item in writing, which was immediately (5-10 seconds) followed by the elicitation of students' thought processes and rationale for responses.

It is worthwhile to briefly discuss verbal reports as they relate to assessment. Ericsson and Simon (1993) discuss two types of methods for verbal reports: indirect methods (ability tests and self-reports) and direct methods (concurrent and retrospective reports). In ability tests, for example, where an individual's shorter written response is marked as correct or incorrect, the cognitive processes can only be inferred from the limited data that is provided, hence the label "indirect." Direct methods, such as verbalization, are assumed to provide a more accurate picture of the cognitive processes involved in the task. Responses are more likely to be an accurate and more detailed reflection of the cognitive processes involved. Direct assessment techniques offer more insight into students' thoughts and, therefore, have greater implications for understanding how learning occurs and how effective and reliable our indirect assessments can be. It is this reasoning that supports the primary line of inquiry in this study: a better

understanding of the ways in which direct assessment methods can reveal more information than is possible from indirect assessment methods.

There are some other limitations to conducting interviews with students that are worth discussing. Despite the usefulness of interview data analysis, the process is often time-consuming and requires some level of subjective interpretation. Also, the procedures place considerable demands on an interviewer who has to listen carefully to what the individual is saying and follow up with additional probing questions when responses are not clear. In addition, there is typically another level of interpretation undertaken by the coder of the verbal data. Some of these limitations have been cited as reasons for interviews not being as widely used in some areas where they may be needed, such as assessment development (Leighton, 2004).

Leighton (2004) described the conditions that support successful use of verbal reports. Assessment tasks of moderate difficulty levels are best used to collect verbal reports. These kinds of tasks are most appropriate because they require students to engage in cognitive processing in which knowledge and strategies are selected consciously, thus allowing the student to be aware of his or her thinking while solving the problem (p. 11). Being aware or conscious of their thinking provides students with something to talk about during the verbal report. On the other hand, if the problems are too easy, students will engage in cognitive processing that is more automatic. Thus, problem solving occurs too quickly and fails to leave a conscious trace in working memory, so the student is unaware of how the problem was solved (Anderson, 1990; Leighton, 2004, p. 11). Conversely, problems that are more difficult require a lot of mental resources; and, therefore, students are less likely to provide accurate verbal reports.

Osborne and Gilbert (1980) reported some limitations in their interviewing-about-instances methodology. First, they noted that there was the problem of choosing a limited but adequate set of instances. That is, the phenomena used to guide the discussion and probe for student ideas invariably affect the kinds of data that are collected and conclusions that are made. Second, they noted that the order of instances may influence the student responses. These limitations should be considered carefully in conducting interviews with students.

Conceptual Framework

This section of the chapter presents a conceptual framework for research on assessment item validation using clinical interview methodologies. In light of the bodies of literature presented in this chapter, it is reasonable to conclude that more carefully designed assessment validation research incorporating data on actual students' ideas collected during clinical interviews is needed at the middle school level in this physical science domain. The assessment triangle (Figure 1), which is discussed commonly in measurement literature (NRC, 2001), provides a useful framework for considering alignment issues for assessments and their development. Figure 2 proposes a conceptual framework adapted from the assessment triangle to guide and focus the work presented in this dissertation study. This framework is useful for understanding how all of the aspects of the assessment process are aligned in the analysis of data collected using the MFA instrument. The observation portion of the assessment triangle in Figure 1 is represented in two places in Figure 2; both the written and oral student responses to the assessment tasks and contexts of interest (boxes B, D and E). The interpretation portion of the assessment triangle is also represented in two places in Figure 2; the two

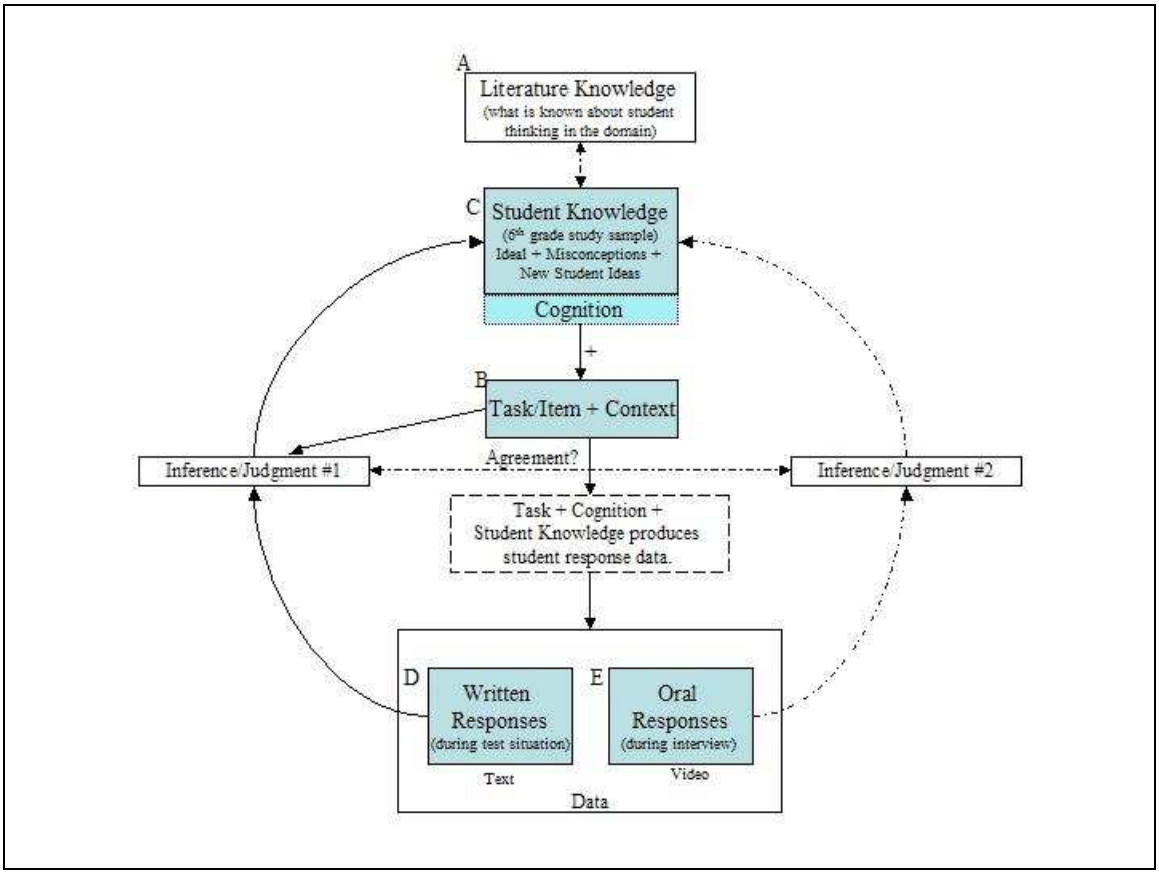


Figure 2. A conceptual framework based on the assessment triangle for comparing inferences from student assessment data.

Inference/Judgment boxes on the right and left sides of the figure. The cognition portion of the assessment triangle is represented in the framework as a portion of Box C. A theory about student cognition in the domain of forces and motion is relevant for making sense of responses to assessment student knowledge in the domain. The cognitive assumptions that underlie this study pertaining to the forces and motion domain are presented and discussed in more detail at the beginning of Chapter III.

Figure 2 presents a conceptual framework for research on assessment validation that incorporates aspects of what is known about student ideas from the literature to judge the quality of inferences being made about study sample students' ideas using assessment

tasks/items for both the written assessment and the verbal interview. This conceptual framework helped to guide the research conducted in this study by clarifying the data sources and the interpretive work in the dissertation. Referring to Figure 2, the dashed lines represent areas where this dissertation research made connections and contributions. First, the task of making Inference/Judgment #2 from the videotaped oral interview responses back to student knowledge was critical in determining what student ideas (both ideal and misconceptions) are evoked by the assessment tasks during the interviews. Second, this judgment was compared against the judgment made by the written assessment item (Inference/ Judgment #2), looking explicitly for areas of agreement and disagreement between the two judgments and across all the assessment items. In addition, the analysis focused on the variety of task contexts to which individual students are exposed by looking for consistent response processes for a given student across an array of item contexts. Finally, the framework shows how this study contributed to the literature regarding common forces and motion student ideas, especially as it relates to aligned assessment items.

Comment on Validity

Finally, it is worth commenting on the concept of validity and how it was interpreted in this dissertation study in light of recent scholarly debates on validity research (see *Educational Researcher*, 2007, Vol. 36, No. 8). Given the nature of the content domain targeted (specific middle school forces and motion ideas), this dissertation study adopts a perspective on assessment validity that is best interpreted at the individual item level. That is, the dissertation is primarily interested in the extent to which individual items provide valid information about student understanding of the

targeted ideas. However, the narrow focus of the assessment instrument under investigation in this study seeks only to make an inference about students' science understanding that is closely linked to the focus of individual items. While many assessment instruments often attempt to make more general inferences about students' skills or ability in broader domains (i.e., science or math ability more broadly construed), this instrument's focus is narrow and therefore, the concepts of item, test, and inferential validity could become interrelated and confounded. However, throughout this dissertation, the author uses the term validity to more narrowly describe item level inferences about student understanding as opposed to broader notions of validity regarding inferences drawn from test performance about students' skills or abilities.

Conclusions

This chapter synthesized the literature about three areas of the research relevant to this study. First, it provided a review of the assessment efforts that have been conducted by other researchers in the area of middle school forces and motion. Despite a wealth of instrument development at the higher grade levels, very few systematic assessment development efforts around forces and motion have been undertaken in the middle grades. It appears, though, that this is an auspicious time to be conducting research in this area, as more assessment development efforts in this domain have recently begun, and there are likely to be more sophisticated measurement options in this particular domain of science.

The second section of the literature review presented research around student thinking about forces and motion at the middle school level. While most of the misconceptions research in physics education appears to be at the high school and college

levels, a considerable number of studies have been conducted chronicling the way younger students think about forces and motion. Depending on the focus of the research, many of these studies had similar findings (when compared to research on high school students) in terms of how students think about motion and the forces acting. These ideas have been summarized in the chapter and are important for the methodology of this study.

The third section of the literature review addressed the use of interview methodologies in studies of human cognition and psychology. Interviews have been and continue to be important to researchers in a wide range of disciplines to understand more about how subjects understand various issues. While some have criticized the ability of interviews to uncover some aspects of cognition, many still believe that clinical interviewing is an important and reliable way to collect data about thought processes.

Finally, the last section presented the ways in which the separate parts of the literature review come together to help inform the development of a conceptual framework for the assessment validation work using an interview methodology. This conceptual framework was useful for guiding the data analysis, interpretations, and conclusions.

CHAPTER III: METHODOLOGY

Organization of the Chapter

This chapter presents the paradigm of inquiry to be used in the study, the research design, a reiteration of the research questions, a description of the data sources and their origin, the participants, the instrument under investigation, the data analysis methods and procedures, the data presentation, and the consideration of human subjects issues.

The purpose of this study was to explore the extent to which paper and pencil assessment items, judged to be aligned with specific middle school forces and motion ideas, actually measure those ideas. The problem to be addressed was to develop an approach to validate assessment items for forces and motion ideas based upon the responses of middle school students. The study determined the extent to which judged-to-be aligned test questions accurately elicited student thinking about the forces and motion ideas of interest.

Paradigm and Model of Cognition

The purpose of this section is to provide a presentation of the paradigm of inquiry and a discussion of the cognitive model adopted in the study. The cognitive model presented in this section was used to guide the interpretations of student responses to assessment items aligned with forces and motion target ideas.

Paradigm of Inquiry

A research paradigm serves as a guide to a specific context of social inquiry and includes the questions to ask and the type of research method to use (Donmoyer, 2001).

A paradigm helps to provide a basis for understanding phenomena and allows an individual to make assumptions during the research process (Creswell, 1994; Denzin & Lincoln, 1994). The paradigm adopted for this study is constructivism. Constructivism, in general, refers to knowledge and the learning process as being constructed by individuals as they are exposed to particular experiences (Phillips, 1995). Fundamentally, constructivism is about individuals creating individual meanings, often in social settings (Carlsen, 2007). In the context of this research on assessment validity in science education, constructivist views of learning are particularly salient as not only do individual students construct their own understandings of science concepts, but this individual construction of knowledge challenges assessment developers to accurately characterize and measure science understanding across large numbers of students.

Model of Cognition

Sfard (1998) discusses two metaphors for learning, which are useful for considering learning theory in science: the *acquisition* metaphor and the *participation* metaphor. In the acquisition metaphor, there is a belief that learning is the process of gaining possession over some commodity as "concepts are to be understood as basic units of knowledge that can be accumulated, gradually refined, and combined to form ever richer cognitive structures" (p. 5). On the other hand, learning is increasingly viewed as participation, and researchers talk about learning as a legitimate peripheral participation or as an apprenticeship in thinking (Sfard, 1998). Sfard's distinctions about knowledge and learning are useful for educational, psychological, or anthropological researchers to make sense of human behaviors in a wide array of situations in schools, but also in everyday settings (see, for example, Bransford et al., 1999). However, as is sometimes

the case with assessment development efforts, groups often adopt a particular view of learning that will hopefully justify, facilitate, and clarify the development procedures and assumptions about the target of inference (NRC, 2001). Mislevy (2006) characterizes this process in the following way:

> Any particular assessment is shaped by its purposes, its constraints, and the context of its use, but also fundamentally by the psychological perspective that gives the argument meaning. This grounding entails the kinds of claims the assessment is meant to support, the kinds of data that are sought and how they are interpreted, and the rationale for each link in a chain of reasoning from observation to inference.

When considering assessment development efforts intended for large numbers of students in an experimental research study, it is often the case that a model of learning that is more consistent with a cognitive acquisition metaphor of learning will be adopted. This is perhaps partly because developing an assessment strategy for large numbers of students that is consistent with a participation metaphor for science learning would be both logistically complicated and expensive. Thus, it is a view of learning that is more consistent with the acquisition approach that primarily guided SCALE-uP in the development of the MFA instrument (Pyke & Ochsendorf, 2006). Because the research presented here investigates the MFA instrument, it adopted similar assumptions about the ways in which student knowledge is situated and used by individuals.

Therefore, within this acquisition view of learning, this dissertation research adopted a cognitive perspective in the forces and motion domain of interest. This perspective is consistent with what many have described as an information processing perspective (see Mislevy, 2006). To help clarify the assumptions around this cognitive view of learning, it is helpful to consider the essential, widely accepted constructs that are relevant in this information processing model of learning (e.g., knowledge types, working

memory, and long-term memory). Working memory is what individuals use to act on information immediately before them (Baddeley, 1986) (e.g., assessment tasks or novel problems). Working memory receives information both from sensory inputs and from long-term memory (e.g., ideas stored about forces and motion ideas). Working memory often refers to the active portion of long-term memory, but the capacity of working memory has been shown to limit an individual's performance on a variety of tasks (Huitt, 2006). Long-term memory, however, is typically considered to be a limitless information store and contains two kinds of information—semantic ("the way the world is") and procedural ("how things are done") (NRC, 2001). These two kinds of information have commonly been equated with what science education has termed the "content" and "inquiry" pieces of science learning standards. When it comes to effective assessment of such learning standards, what matters most in the assessment of learning is "not the capacity of working memory but how well one can evoke the knowledge stored in long-term memory and use it to reason efficiently about information and problems in the present" (p. 68). Thus, efforts should be made in assessment development to restrict the working memory resources needed to respond to a test item to those absolutely necessary to demonstrate understanding of the science ideas of interest.

In addition, the information processing model is distinguished by other essential characteristics. First, the entire model is assumed to be constrained by some limited capacity (i.e., bottlenecks) most likely distinguished by limits on working memory, classically surmised by Miller (1956) with his "seven plus or minus two" principle. Second, some control mechanism is required to oversee encoding, processing, storage, retrieval, and utilization of information in such a way that an executive function "uses

up" some of the capacity of the entire system. Therefore, a new task (i.e., novel assessment task) requires more executive processing than a routine task. Third, there is a two-way flow of information as individuals simultaneously use information gathered from the environment and information stored in long-term memory (Huitt, 2006). These components are helpful for considering the ways in which individuals access and utilize ideas stored in long-term memory to explain and make sense of phenomena presented to them in a given assessment task.

Assessment that is interested in what knowledge or ideas are possessed by individuals in long-term memory should limit the demands placed on working memory so that performance more accurately reflects stored knowledge structures as opposed to any limits on an individual's working memory. This point is important to consider for individuals who are learning English and attempting to demonstrate their understanding of specific science content on written assessments, such as the MFA. Potential for successful second language learning has been shown to depend greatly on the capacity of working memory (Gonzalez, 1999; Skehan, 1998). This suggests that individuals currently learning English who are engaged in the processes of interpreting and responding to assessment items (i.e., decoding, translating, and generating text) may be restricted in their ability to demonstrate their understanding of a given science idea. That is, the demands placed on working memory by the decoding and translating of text in the second language may use up valuable cognitive resources that are essential in arriving at responses that accurately reflect stored understandings of key science content.

Cognitive scientists draw distinctions between domain-general knowledge structures, which are applicable to a wide range of situations (similar to scientific inquiry

skills), and domain-specific knowledge structures, which are relevant to specific problem areas (NRC, 2001). Assessment of domain-general knowledge structures typically is interested in general problem-solving abilities in a range of novel situations. (The Law School Admission Test is a good example of this kind of assessment.) Assessment of domain-specific knowledge is more common in K-12 education, as it is more interested in individuals' understanding of a domain of interest (NRC, 2001). In these kinds of assessments, individuals typically try to interpret situations so they can apply *schemas*—previously learned and somewhat specialized techniques for organizing knowledge in memory in ways that are useful for solving problems (NRC, 2001). Schemas enable competent performers to recognize situations as instances of problems they already know how to solve, to represent such problems accurately, and to know which strategies to use to solve them (Baddeley, 1986, 1990). Therefore, the ability to problem solve in science, for example, is dependent upon the organization of an individual's knowledge structure in that domain (long-term memory). These assumptions about cognition and assessment are relevant for understanding the response processes that middle school students undergo when responding to assessment tasks regarding their domain-specific knowledge about forces and motion.

The cognitive perspective adopted here shares the views of others in the science education research community (see Scott, Asoko, & Leach, 2007) and shares common principles about learning and knowledge. These principles can be applied to the learning and assessment of middle school student learning in the domain of physical science. This dissertation research was aligned with the following four cognitive assumptions about the nature of knowledge and the assessment of student ideas in this domain:

1. Students' ideas about the natural world are constructed, rather than received. Students develop accurate and inaccurate understandings about objects in motion and the forces acting on them based on both relevant life experiences and instruction in school and out of school.
2. There are strong commonalities in how individuals appear to think about the natural world and about forces and motion specifically. These commonalities have been reported in the literature (summarized in the literature review of this dissertation) and are detectable on the individual level with well-designed assessment tasks or interviews.
3. A person's existing ideas about a given subject greatly influence his/her subsequent learning about that subject. In the forces and motion domain, despite well-designed instruction, many students have difficulty constructing accurate ideas about forces and motion due to the competing presence of often strong, personally helpful, but scientifically inaccurate ideas about the natural world.
4. Individuals' ideas are cognitively available in relevant contexts to offer and evaluate explanations, make predictions, analyze situations, and interpret representations. In assessment situations such as the MFA, students call upon their understandings of moving objects and the forces acting on them to make sense of and explain phenomena in a variety of contexts. However, students may not consistently be able to recognize when particular ideas are useful to explain all relevant contexts.

These assumptions of learning and knowledge in this forces and motion domain guided the development of this research design, the research questions, data analysis, and conclusions.

Research Questions

Given the interest in understanding how students' ideas are used while responding to science assessment items, the following research questions were addressed:

1. To what extent do the inferences at posttest from each of seven paper and pencil science assessment items (aligned by expert judges using specific alignment criteria to particular forces and motion ideas) agree with the inferences about students' understanding of those ideas as determined during videotaped oral interviews at posttest?
2. During both a written posttest and a videotaped oral interview, to what extent are middle school students in this study able to consistently demonstrate their understanding of the same forces and motion ideas targeted by a variety of assessment task contexts?
3. Based on the elicited ideas of sixth graders in this study, what can this research add to the literature base (as presented in the study) about common ideas students have regarding forces and motion?

Methodology

This study was conducted in the context of a large, six-year, quasi-experimental research study (SCALE-uP) that was designed to explore the effectiveness and scale-up of promising middle school science curriculum materials in a large, diverse school

district. SCALE-uP researchers, as part of ongoing assessment validation efforts, collected the data being analyzed in this dissertation study.

Background on SCALE-uP and Study Site

This section will present background on the goals of the Scaling-up Curriculum for Achievement Learning and Equity Project (SCALE-uP), the student participants involved in this dissertation study, and the selection methods used in the dissertation. Because this study involved only 14 participants, data and interpretations derived from student data are not intended to be generalized to the population from which they are drawn.

Background on SCALE-uP

In 2002, researchers at GWU received funding from the Interagency Education Research Initiative (IERI) to study the effectiveness and scale-up of promising middle school science curriculum units (Lynch, Kuipers, Pyke, & Szesze, 2005; Lynch et al., 2007) in Montgomery County Public Schools (MCPS), a large, suburban school district in Maryland. One of the aims of SCALE-uP was to explore possibilities for closing science achievement gaps in a school district that has an increasingly diverse student population. SCALE-uP's interdisciplinary research employs a quasi-experimental design along with a video ethnography to better understand how highly rated curriculum materials function in diverse classrooms. In the initial phases of the study, SCALE-uP researchers sought to determine if the promising middle school curriculum materials were more effective than existing materials in use in MCPS. (See Lynch et al., 2006 for a discussion of materials used by a comparison group in a SCALE-uP study.) Subsequent

phases of SCALE-uP focused on understanding the conditions under which effective curriculum interventions can be scaled-up across MCPS (Lynch et al., 2005; Lynch et al., 2007).

As part of the quasi-experimental design, SCALE-uP researchers designed assessment instruments that would be used to measure student learning of the content standards targeted by the highly-rated curriculum materials. Three such instruments were developed as part of SCALE-uP research: (a) a Conservation of Matter Assessment (eighth grade), (b) a Causes of Seasons Assessment (CSA) (seventh grade), and (c) a Motion and Forces Assessment (MFA) (sixth grade) (Pyke & Ochsendorf, 2006). The dissertation researcher (Ochsendorf) was part of the assessment development team for both the CSA and the MFA instruments. As part of the ongoing assessment development and validation process, SCALE-uP researchers have been interested in better understanding the ways in which students respond to the assessment items in each of the instruments. Thus, classrooms participating in SCALE-uP's ethnographic research have proven useful for collecting videotaped, structured interviews of student responses to assessment questions. Interview protocols have been developed for each of the assessment instruments by SCALE-uP staff to generate validation data. The dissertation researcher, in conjunction with Dr. Curtis Pyke and other SCALE-uP staff, played an important role in the development of the interview protocols that have been used by SCALE-uP for this purpose. This dissertation research analyzed the interview data collected in one year of SCALE-uP research as it relates to one of SCALE-uP's assessment instruments.

Description of Study Site

This study analyzed videotaped interviews conducted during April and May of 2006 as part of the SCALE-uP study with 14 sixth grade students from a Montgomery County (MD) middle school. The participants in the study were selected from Montgomery County Public Schools (MCPS), a large Maryland school district (approximately 139,000 students total; 31,000 in grades 6-8 during the 2005-2006 academic year) located in the Washington, DC, metropolitan area. There is no ethnic majority in MCPS. Among the 31,000 students in grades 6 through 8, 27% are eligible for Free and Reduced Meals Status (FARMS), and the percentages of various ethnic groups are as follows: 23% African American, 15% Asian American, 19% Hispanic, and 43% White. MCPS has been among the top-performing school districts in the State of Maryland.

Student Participants

In the SCALE-uP study, one middle school science classroom was selected to be videotaped for the implementation of the *Motion & Forces* (*M&F*) curriculum materials (Harvard-Smithsonian Center for Astrophysics, 2001) during the 2005-2006 school year. The *M&F* materials are hands-on curriculum materials designed for grades 5 through 8 that emphasize students' direct experience with phenomena. During 2003, a curriculum analysis of the *M&F* written materials using the Project 2061 Curriculum Analysis procedures was conducted by SCALE-uP (see Kesidou & Roseman, 2002, for details and background on the procedure; see Ochsendorf et al., 2003, for the *M&F* curriculum analysis report). The *M&F* curriculum materials were determined to be aligned to the forces and motion ideas of interest. *M&F* met, at a satisfactory level, 10 of the 22 Project

2061 instructional criteria and, when implemented as part of SCALE-uP research, proved to perform better than existing curriculum materials in use in MCPS, especially for underserved subgroups of students (for more discussion, see Lynch, 2007). *M&F* showed positive effects for subgroups of students that often do not perform well in science, including African American and Hispanic subgroups, prior and now English Language Learner (ELL) subgroups, and students eligible for special education (Watson, Pyke, Lynch, & Ochsendorf, submitted for publication). According to the curriculum analysis (Ochsendorf et al., 2003), the *M&F* materials were rated well in terms of their potential to provide a sense of lesson purpose for students, help teachers identify their own students' ideas, engage students with relevant phenomena, and encourage students to explain their ideas. The *M&F* materials appeared to be very promising in terms of their effectiveness (as indicated by findings from SCALE-uP's third quasi-experimental replication study with the materials in MCPS). (See Watson, Pyke, Lynch, & Ochsendorf, 2007, for more discussion of three consecutive quasi-experimental studies of *M&F*.) Prior to participating in this study, these sixth grade students had ample opportunity to experience phenomena directly related to the target ideas by engaging with the *M&F* curriculum materials.

The classroom that was videotaped in the SCALE-uP study was selected based on the willingness of the teacher to participate in that study. SCALE-uP research staff worked with the classroom teacher to select an academically, ethnically, linguistically, and socioeconomically diverse classroom that would meet the needs of an ethnographic research team, who were studying how the *M&F* curriculum materials function in a diverse classroom setting. (For more discussion on the video-ethnography of the *M&F*

curriculum materials, see Kuipers, Viechnicki, Fugate-Brunino, Jones, & Wright, 2006.) The students in the selected classroom and their parents signed an Internal Review Board (IRB) agreement and video release form prior to filming and were aware of the conditions of the study. Students and parents were aware that SCALE-uP researchers during, or immediately following, the implementation of the *M&F* could interview students in the class. A total of 25 students in this single science classroom were filmed for 35 class periods during the spring of 2006 in an alternating block schedule (a combination of 90- and 45-minute class periods). Two SCALE-uP research members, Lindsey Jones and Laura Wright, were in the science classroom with the students every day setting up cameras and managing the videotaping procedures, which typically involved two stationary cameras. In this particular classroom, *M&F* lasted 31 instructional days. One camera filmed the same four students every day, while another camera circulated around the room filming one of the other tables of three or four students each day. If necessary, cameras and SCALE-uP staff followed students into hallways or other rooms when *M&F* implementation required students to leave the classroom to conduct its various Explorations. By the end of M&F implementation, all students in this science classroom had been exposed to video cameras for a considerable number of hours. It is, therefore, reasonable to assume that students had some level of comfort with cameras by the end of the study, although students were not directly queried about their comfort level with cameras or the presence of researchers in their science classroom.

The videotaped classroom was located in a diverse middle school with the following ethnic percentages in 2005-2006: 39% African American, 13% Asian

American, 24% Hispanic, and 23% White. The school is somewhat more diverse than the overall middle school population in MCPS. The ethnic makeup of the 25 students in the classroom selected for videotaping was as follows: 32% African American, 14% Asian American, 18% Hispanic, and 32% White. Only 3 of the 25 students in the class were classified as having either prior or now ELL status.

At the completion of the *M&F* unit, the teacher selected 14 students to participate in the interviews and data collection. The students were selected by the teacher to represent a diverse sample across a number of criteria, including perceived knowledge of forces and motion ideas, gender, and ethnicity. The videotaped interview data collected on these 14 students were used as the basis of this dissertation study.

Instrumentation

The Motion and Forces Assessment (MFA) instrument, under investigation in this study, is a curriculum-independent assessment designed to measure understanding of specific middle school physical science benchmarks (mapped in Figure 3) for the SCALE-uP research. This dissertation study was greatly interested in the quality of the inferences made using the written MFA instrument. Sixth grade students in this research had previously taken the MFA in written paper and pencil format; their responses had been rated by SCALE-uP researchers; and each student had been assigned a scale score. Assessment and scoring procedures for the MFA are described below.

An analysis and development procedure developed by AAAS Project 2061 (DeBoer, 2005; Stern & Ahlgren, 2002) was used as part of the process to develop the MFA for the SCALE-uP study (Pyke & Ochsendorf, 2006). Alignment of assessment items to target science ideas is achieved by using a set of review criteria in the

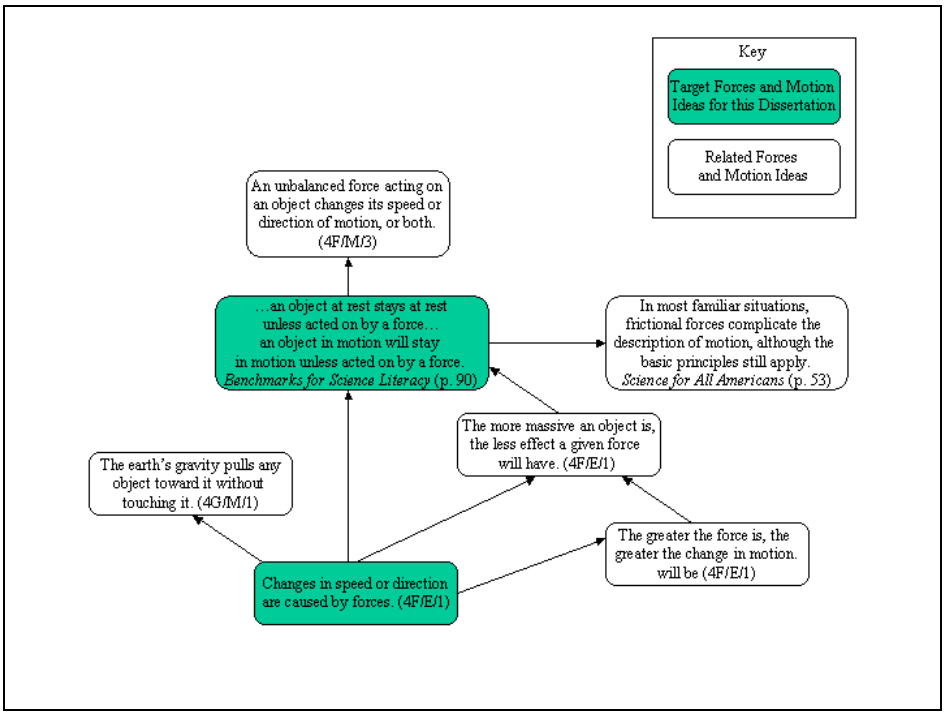


Figure 3. Assessment map of forces and motion target ideas used by SCALE-uP in development of MFA.

development process that are designed to carefully consider whether ideas in a content standard are *needed* to complete an assessment task successfully or if the task can be completed in some other way, and whether those ideas are *enough* by themselves or if other ideas and skills are needed (DeBoer, 2005). The Project 2061 assessment analysis procedure involves the following major components, which include a separate set of specific review criteria (not shown here) for each component:

1. Content Analysis. Evaluates the alignment of the task to the targeted learning goal (i.e., science idea).
    a. *Necessity*. If any part of the knowledge that is needed to complete the task . . . then the learning goal is necessary.

b. *Sufficiency*. If any part of the learning goal is adequate by itself to complete the task, then the learning goal is sufficient.

2. Confounding Factors. The focus is on features of the assessment task that could reduce the likelihood that students' responses will provide information about what they know and do not know.

    a. *Comprehensibility*. Addresses whether students are likely to understand the task statement, diagrams, and symbols [and] should look for unfamiliar general vocabulary and language that may not be familiar to poor readers or students whose first language is not English.

    b. Appropriateness of Task Context. A task context sets up particular situations involving specific organisms, objects, or events, including real-world occurrences or idealized conditions. The context should be understandable, interesting, and sensible to students.

    c. Resistance to Test Wiseness. The criterion addresses characteristics of the task that might (1) allow students to make a satisfactory response using only general test-taking skills without understanding the idea being tested or (2) mislead students into choosing an incorrect answer.

3. Cost Effectiveness. Whether the task seems to require a great deal of time and effort for students to complete, relative to other items that probe student understanding of the idea.

The review criteria are an important first step in minimizing the level of construct-irrelevant variance in the assessment tasks. For both Project 2061 and SCALE-uP, these

review criteria serve as one component in the entire item development procedure. Other important steps in the overall development process include clarification of the target ideas, piloting of assessment items with small groups of students, and interviewing students. The use of an assessment map (similar to the one shown in Figure 3) was helpful for assessment developers to consider the entire set of forces and motion ideas, which assisted in clarifying which ideas were and were not being targeted by a given assessment item.

SCALE-uP's assessment development team included science content experts, science educators, science teachers, and assessment specialists. The dissertation researcher was involved in the assessment development process and played an integral role in pulling together feedback from the initial development team, which ultimately lead to the final MFA instrument. Additionally, over the course of four years (after initial MFA development), the dissertation researcher was involved in some decisions to make modifications to the MFA based mostly on psychometric data as a result of its use with thousands of sixth graders, but also based on feedback from teachers, administrators, and researchers close to the project.

The assessment development procedure focused on the strict alignment of items with content standards and involves criteria aimed at reducing the number of construct irrelevant factors. The use of plain language and simple illustrations allow the MFA to be read and understood by a wide range of students with different English language abilities (Pyke & Ochsendorf, 2004). Illustrations for items are meant to provide a simple but rich context in which to consider the motion of objects and the forces acting. Although the goal was to develop an instrument that was accessible to as many students as possible, the

effectiveness of the MFA instrument at detecting students' ideas about forces and motion admittedly depends on a basic level of English proficiency. Students who are learning English are helped by the simple language and detailed diagrams that are used in the MFA. The assessment map in Figure 3 shows the organization of target forces and motion ideas for this dissertation study and other related forces and motion ideas. The map shown in Figure 3 is narrower in scope than a more detailed assessment map used by SCALE-uP in the development of the MFA. The SCALE-uP assessment map was adapted from a similar map on these ideas published by Project 2061 in *Atlas of Science Literacy* (AAAS, 2001).

The MFA consists of 10 items (6 constructed response items and 4 selected response items) that assess students' understanding of forces and motion. Items and their illustrations are situated in four task contexts: pushing a couch, deceleration of full and empty dump trucks, dropping and throwing a ball, and a rolling ball. Illustrations used in MFA were carefully developed to be accessible and engaging to as many students as possible. The final set of 10 items was obtained after piloting and evaluating the measurement qualities of a larger number of items for various phenomena. Minor modifications have been made to the MFA by SCALE-uP since its initial use in 2004.

The SCALE-uP assessment development team, led by Dr. Curtis Pyke, made some assumptions about students' interactions with and responses to assessment items. The MFA instrument was designed by SCALE-uP as a paper and pencil assessment test because it would be used with large numbers of students. In addition, because the instrument was to be used in a quasi-experimental curriculum study with data analyses procedures focused at the student level, it was necessary to arrive at a scaled score (0-

100) for each individual student. Individual items were weighted to take into account item statistics such as difficulty and discrimination.

Trained raters evaluated student responses to the constructed response items. Coding categories were determined by the SCALE-uP development team's analysis of pilot data of the MFA collected during 2003. Coding categories for the written assessments were further developed and refined using a sample of student responses from subsequent years. Science educators and science graduate students who were trained by SCALE-uP to use the codes were used to rate student responses to the assessment items. The dissertation researcher coordinated the training of assessment raters for four years of SCALE-uP research (2003-2007). During each year of its use, approximately 5% of the MFA assessments (~150) were rated by multiple raters to estimate the instrument's reliability. Inter-rater reliability Kappa statistics for pairs of raters for each item have been between .66 and .95, and average Kappa coefficients for each item across all raters have ranged from .77 to .91 (see, for example, Pyke, Lynch, Kuipers, Szesze, & Watson, 2006). A Kappa coefficient (which takes into account raters agreeing by chance) above .70 is considered acceptable for purposes of reliability.

In addition to the 10 items that comprised the original MFA, SCALE-uP researchers added five additional selected response items during the 2005-2006 study year. These items were added to the MFA as part of a SCALE-uP effort to explore issues of alignment and validity. These items had been developed independently by AAAS Project 2061 staff and judged to be aligned to the same target ideas. These five items involve real-world contexts, including a moving skateboard, pedaling a bike, a metal ball and crane, a toy car, and a rolling shopping cart. Table 4 shows all 15 items used by

SCALE-uP during 2005-2006 as part of the study of *M&F*. Table 4 also shows the intended alignment of each of the items with its target idea.

Each of the 14 students involved in this dissertation study, as well as a random sample of students from the 10 schools involved in the 2005-2006 study (n=294), were administered the MFA by the classroom teacher as a posttest upon completion of the *M&F* curriculum unit. The written assessment responses were rated at the item level by five SCALE-uP-trained graduate students (from the Department of Forensic Science), and a score from 0 to 100 was assigned to each student using a weighted scoring scheme developed by SCALE-uP (Pyke & Hanson, 2005). SCALE-uP used an analytic judgment method (Cizek, 2001) to arrive at four levels of understanding for the MFA. The four levels of understanding for the MFA are described as follows:

> Level IV. Flexible Evidence of Understanding (Scores of 70-100): Student shows a flexible understanding of, and commitment to, scientifically accurate ideas of forces and motion, with few misconceptions.
>
> Level III. Some Fluency of Understanding (Scores of 50-70): Student shows some fluency with the forces and motion ideas, but has misconceptions in certain contexts.
>
> Level II. Some Evidence of Understanding (Scores of 20-50): Student shows some evidence of understanding forces and motion in specific contexts.

Table 4. *MFA and Project 2061 Assessment Items with Intended Target Ideas.*

	Item	Target Idea
SCALE-uP (MFA Item)	Pushing a Couch-SR	The greater the force the greater the change in motion will be.
	Pushing a Couch-CR	The greater the force the greater the change in motion will be.
	Dump Truck-CR	The more massive an object is, the less effect a given force will have.
	Dropping a Ball-SR*	Changes in speed or direction are caused by forces.
	Dropping a Ball-CR1*	Changes in speed or direction are caused by forces.
	Dropping a Ball-CR2	Changes in speed or direction are caused by forces.
	Dropping a Ball-CR3	The greater the force the greater the change in motion will be.
	A Rolling Ball-SR*	Objects in motion stay in motion unless acted on by a force.
	A Rolling Ball-CR1*	Objects in motion stay in motion unless acted on by a force.
	Rolling Ball-CR2	Changes in speed or direction are caused by forces.
Project 2061 Item	Carly's Skateboard-SR*	Changes in speed or direction are caused by forces.
	A Toy Car-SR	Changes in speed or direction are caused by forces.
	A Metal Ball-SR*	Objects at rest stay at rest unless acted on by a force.
	Shopping Cart-SR	Objects in motion stay in motion unless acted on by a force.
	Pedaling a Bike-SR*	Objects in motion stay in motion unless acted on by a force.

SR=Selected Response; CR=Constructed Response; *Denotes items included in the interview protocol. Cronbach's Alpha for 10 MFA items = 0.54; for all 15 items = 0.54.

> Level I. No Evidence of Understanding (Scores of 0-20): Student shows no evidence of understanding forces and motion ideas (Pyke & Hanson, 2005).

Each student in the SCALE-uP study (and the students in this dissertation study) received a weighted scale score of 0-100, which was used as an indicator of their understanding of ideas related to forces and motion. The calculation of the weighted scale score balanced the contribution of selected response and constructed response items and the difficulty

and discrimination of each item as well as distributed the contribution of each of the target ideas (for more discussion, see Pyke, et al., 2006). Assessments were provided to the ethnographic research team to be incorporated with other *M&F* student work that was collected under SCALE-uP's IRB permissions. The dissertation researcher, however, did not look at these student assessments and would not know how a given student performed on the MFA.

Descriptions of Assessment Items for this Study

For reasons discussed below, not all of the items developed by SCALE-uP or Project 2061 were involved in this dissertation study. This section describes in detail the features of the seven assessment items that were used in this study. Items are discussed in terms of relevant features as delineated by SCALE-uP during item development (Pyke & Ochsendorf, 2006). For example, item contexts can include the following item features: (a) different kinds of motion (e.g., sliding, rolling, falling, flying); (b) the state of the forces acting (balanced vs. unbalanced); (c) the constancy of mass and force; (d) the type of force involved (e.g., push, pull, friction, gravity); (e) whether the item depicts objects starting at rest or in motion; and (f) whether the context involves an item changing direction. Features of items also place different cognitive demands on students by asking students to, for example, explain or predict phenomena, choose appropriate explanations, or compare representations (Pyke & Ochsendorf, 2006). The seven items in five contexts are discussed below and include *A Rolling Ball* (two items), *Dropping and Throwing a Ball* (two items), *Carly's Skateboard*, *Pedaling a Bike*, and *A Metal Ball*.

A Rolling Ball includes a selected response and a constructed response item (Figure 4). These two items involve the physical phenomenon of a ball rolling across a

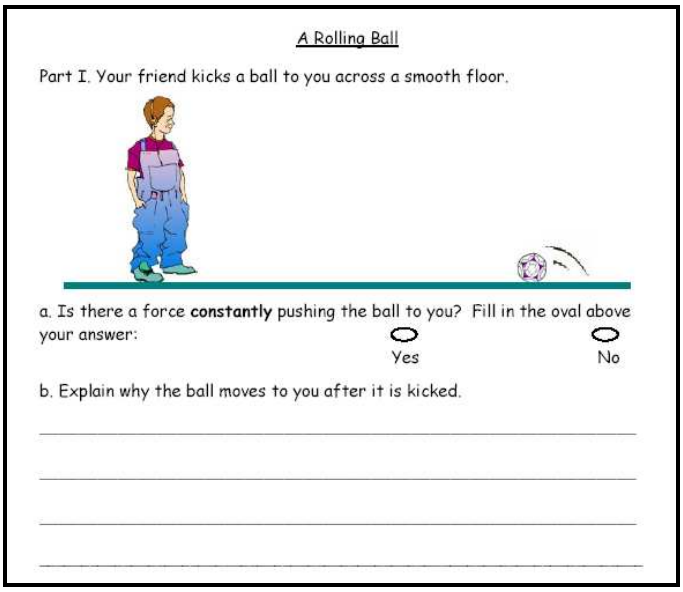


Figure 4. A Rolling Ball (two items).

smooth floor. The item features a context in which neither the mass of the object nor the force acting on it change in magnitude. There is an instantaneous initial force of a given magnitude that acts on the ball, while frictional forces between the ball and the surface are assumed constant throughout its motion. The expectation is that the task will cause students to consider the motion of the ball starting from an "in motion" state and mentally observe an object in motion whose direction of motion does not change. As determined by SCALE-uP, the items are intended to align with the target idea, objects in motion will stay in motion unless acted on by a force. In addition, the expectation for this task requires that students explain a phenomenon (i.e., "Explain why the ball moves to you after it is kicked"). The item is intended to elicit students' understanding and possible misconceptions around the idea that objects in motion will stay in motion without a force acting, which is a complicated idea for many middle school students. The item has the potential to evoke student ideas about the motion of rolling objects and the forces that

may or may not be acting on rolling objects while they are moving. It also has the potential to elicit students' understanding about the nature of forces, although this is not explicitly a target idea for this item. The item is somewhat limited in that it does not propose a frictionless situation in which to consider the possibility that objects in motion might actually stay in motion and not slow down and stop.

Dropping and Throwing a Ball includes a selected response item (What happens to the speed of the tennis ball as it falls?) and a constructed response item (What caused the speed of the tennis ball to increase, decrease, or stay the same?) (Figure 5). These two items involve the physical phenomenon of an object dropped from a given height. The item features a context in which neither the mass of the object (tennis ball) nor the force acting on it (gravity) change in magnitude, and the ball is starting from an "at-rest" state and moving downward such that the direction of the ball's motion does not change. As determined by SCALE- uP, the items are intended to align with the target idea, changes in speed or direction are caused by forces. In addition, this task requires that students justify their prediction about the speed of a falling object. The item has the potential to evoke student ideas about the motion of falling objects and the forces that may or may not be acting on falling objects. It also has the potential to elicit students' understanding about the nature of gravity, although this is not the target idea for this item.

Carly's Skateboard is a selected response item, which asks students to consider the phenomenon of a rolling object that comes to a stop (e.g., Why does the skateboard come to a stop?) (Figure 6). This item features a context and phenomenon in which neither the mass of the object in motion (skateboard) nor the force acting on it (a push)

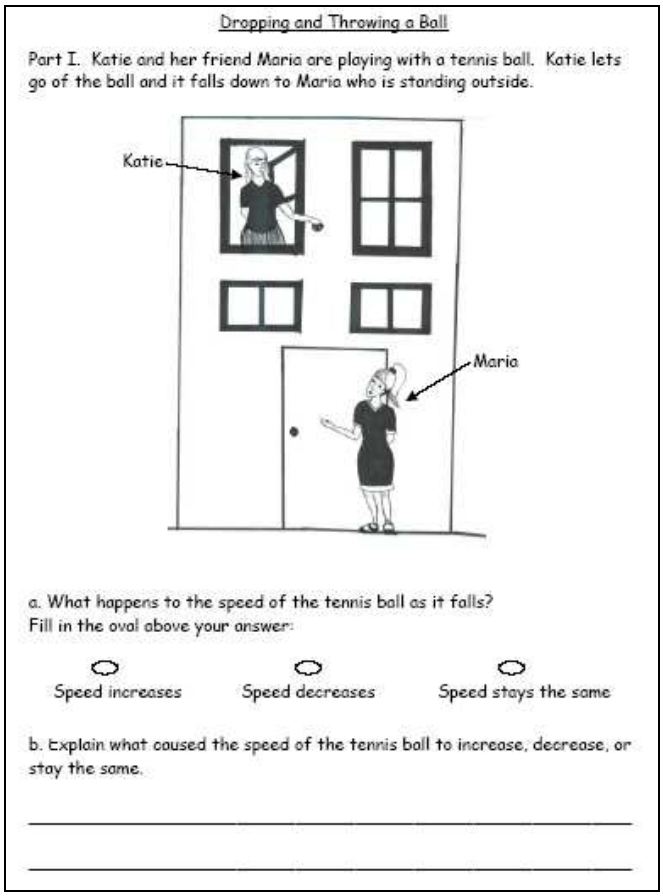


Figure 5. Dropping and Throwing a Ball (two items).

changes in magnitude. The expectation is that this task will cause students to consider why rolling objects, which receive a single push, will eventually slow down and stop. In addition, the task requires students to recognize and choose the appropriate explanation from among four possible explanations for the phenomenon. This item has the potential to evoke student ideas about the forces that may or may not be acting on rolling objects as they slow down. The item is intended to align with the target idea, changes in speed or direction are caused by forces. Therefore, the item can detect students' understandings and possible misconceptions about (a) what causes moving objects to slow down and stop and (b) the idea that objects change speed due to forces rather than a force running out.

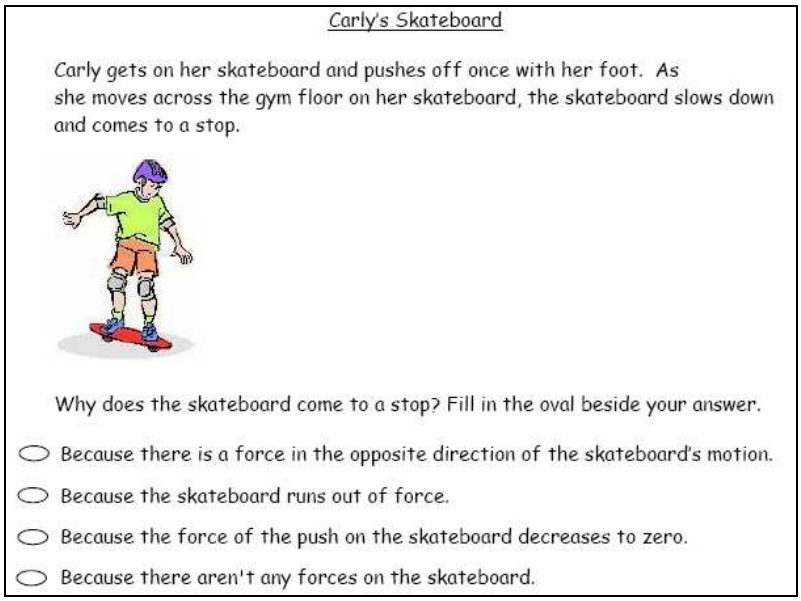
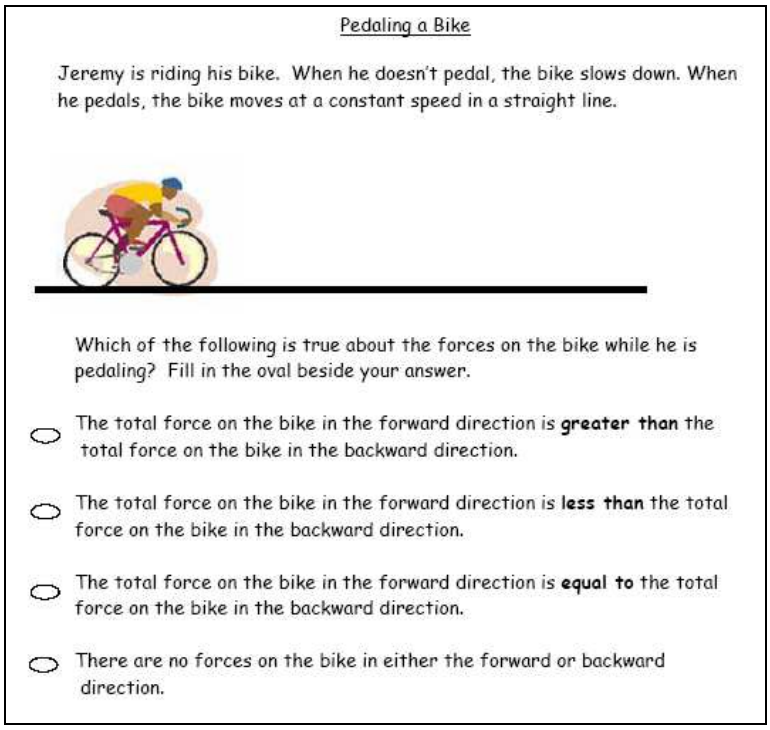


Figure 6. Carly's Skateboard.

Because of the selected response nature of the item, students will need to consider the possibilities that moving objects slow down and stop because they run out of force, that forces decrease to zero, that forces get used up, or that there are no forces acting on the moving object. Some students, when explaining how they answered this question during an interview, may discuss the concept of friction and the impact it has on rolling objects. Demonstrating an understanding of friction while responding to this item is not considered to be a target idea; and, while it may be noted in discussions of student responses, it was not coded for in the data analysis. Clinical interviews with students can be particularly useful on such a selected response item because more information can be collected about student thinking when students are asked to explain their answers rather than simply "filling in the oval" beside an answer choice.

Pedaling a Bike includes one selected response item (Figure 7). This item involves the phenomenon of a bike moving at a constant speed. The item features a



Pedaling a Bike

Jeremy is riding his bike. When he doesn't pedal, the bike slows down. When he pedals, the bike moves at a constant speed in a straight line.

Which of the following is true about the forces on the bike while he is pedaling? Fill in the oval beside your answer.

- The total force on the bike in the forward direction is **greater than** the total force on the bike in the backward direction.
- The total force on the bike in the forward direction is **less than** the total force on the bike in the backward direction.
- The total force on the bike in the forward direction is **equal to** the total force on the bike in the backward direction.
- There are no forces on the bike in either the forward or backward direction.

Figure 7. Pedaling a Bike.

context in which neither the mass of the object (bike and rider) or the forces acting on it (wind, friction, pedaling force) change in magnitude. The task asks students to consider the total forces that are acting on the bike in the forward direction and the backward direction if the bike is moving at a constant speed. Students need to consider how the magnitudes of the two forces (forward and backward) relate to one another. This item is intended to align with the target idea, objects in motion will stay in motion unless acted on by a force. This task requires that students choose the explanation that most accurately describes the state of the two forces acting on the bike while it is moving at a constant speed. The item has the potential to evoke student ideas about the forces acting on an object in constant motion. Specifically, the item attempts to determine if students understand the idea that if an object is in motion (at a constant speed) the forces on the

object in forward and backward directions are equal. Students likely have an easier time understanding the forward forces involved in a bike's motion (pedaling), but are perhaps less familiar with the forces working against the forward motion (rolling friction and wind resistance). This item provides students an opportunity to demonstrate that the know that such opposing forces are involved.

A Metal Ball includes one selected response item (Figure 8). This item involves the physical phenomenon of a ball hanging stationary from a crane. The item features a context in which neither the mass of the object (metal ball) or the forces acting on it change in magnitude. The task asks students to consider the two forces acting on the metal ball (upward and downward) and how the magnitudes of the two forces relate to one another. This item is intended to align with the target idea, objects at rest will stay at rest unless acted on by a force. This task requires that students choose the appropriate explanation for why the metal ball is stationary. The item has the potential to evoke student ideas about the forces acting on an object at rest. Specifically, the item attempts to determine if students understand the idea that the upward and downward forces on a stationary object are equal. While many students likely will have no trouble recognizing the downward force of gravity on the ball, some may not be as familiar with the upward force the crane exerts on the ball (and that it is equal to the downward force). This will be explored further in the study.

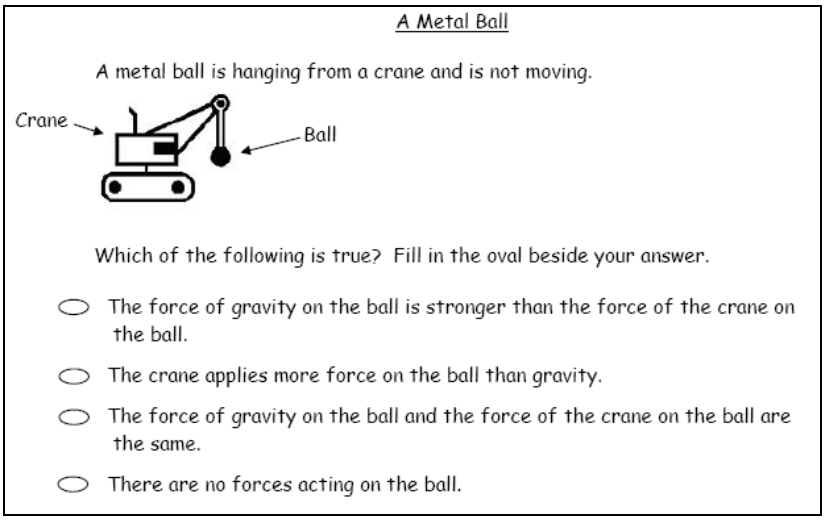


Figure 8. A Metal Ball.

Interview Protocol and Data Collection

This section describes the oral interview assessment protocol development and the interview data collection procedures performed by SCALE-uP staff during 2005-2006 that led to the establishment of the video database of student interviews used in the dissertation study. The interview protocol was developed specifically for the seven assessment items described above.

Interview Protocol for MFA

The interview protocol used to conduct the assessment interviews was developed by SCALE-uP research staff as part of their efforts around assessment validation (Appendix A). The protocol was developed specifically for the MFA items described above, and the development process involved careful consideration of the features and goals of the MFA instrument. During protocol development, the SCALE-uP research team evaluated the extent to which a given targeted forces and motion idea was both

necessary and sufficient to respond to a given item. If students responded correctly to an item during the oral interview (as determined by the MFA scoring guides developed for the paper and pencil tests), the interviewer asked follow-up probing questions to determine that the student had actually used the targeted science idea while responding (as opposed to guessing, using test-wiseness or using other knowledge to answer the question). If the student responded incorrectly to an item, the interviewer would ask probing questions to clarify the reasons for the student's incorrect response. The protocol sought to establish whether the items were comprehensible to the students and that students knew what they were supposed to do to respond. If an item was not comprehensible to students (i.e., language or vocabulary issues), the protocol was designed to determine the nature of the incomprehensibility. If the item was comprehensible to students, the protocol was designed to probe for the presence of the target idea the item was attempting to measure. If a student was able to demonstrate an understanding of the idea in follow-up probing, but not in an initial response, this could indicate a problem for the comprehensibility of the item.

To keep the interview length reasonable, not all of the MFA items were incorporated into the interview protocol. Only the seven items described in the previous section were chosen for interviewing by the SCALE-uP staff. The items used during the SCALE-uP interviews were selected by SCALE-uP staff (including the dissertation researcher) to provide coverage of a range of the target ideas. Items were also chosen because the context could be physically modeled by the interviewer for the students during the interview using everyday objects (i.e., by rolling a ball on a smooth surface). As noted in the previous section, five of the items were selected response items, and two

of the items were constructed response items. Four items (two from the dropping and throwing a ball context and two from the rolling ball context) of the original 10 MFA items were chosen for inclusion in the protocol. Students were provided an opportunity to work with physical models for some of the MFA task contexts. The physical models (for four items) included a tennis ball rolling across a table, a skateboard rolling across a table, and a toy crane holding a stationary ball. In each case, the phenomenon was demonstrated for the students in a way that mirrored the written assessment task, and students were then queried for their understanding and explanations of the phenomena. Three of the five Project 2061-developed assessment items were chosen for inclusion in the protocol (*Carly's Skateboard*, *Pedaling a Bike,* and *A Metal Ball*). Table 4 (earlier in this chapter) details the seven items used for the interview and the intended target idea for each item. The target ideas to be covered by the seven items included in the protocol are:

1. Forces cause changes in speed or direction are caused by forces (3 items).
2. Objects in motion stay in motion unless acted on by a force (3 items).
3. Objects at rest stay at rest unless acted on by a force (1 item).

The interview protocol was designed by SCALE-uP so that a student first read an assessment item aloud and then provided a written answer to it on his/her own. After the student had responded in writing to an item and then read their response aloud, the interview protocol is designed to (a) elicit confirmatory evidence of the student's understanding of the target idea, (b) assess the student's comprehension of the tasks/diagrams, and (c) provide other opportunities for the students to express their understanding of the target ideas (i.e., by using three-dimensional physical models). (See Appendix A for the complete interview protocol.)

Interviews and Data Preparation

Two SCALE-uP researchers who were involved in the development of the interview protocol for the MFA conducted all 14 student interviews. Interviewers were doctoral students in curriculum and instruction with considerable experience in both science teaching and science education research (referred to from now on as David and Ann). Each interviewer completed one pilot interview on-campus at GWU with a volunteer middle-school-aged student (not in this study) approximately two weeks prior to conducting the actual interviews. Each of the pilot interviews was video recorded and critiqued by several of the SCALE-uP staff (including the dissertation researcher) involved in the protocol development. The goal of the critique was to arrive at agreement about how the interviewer should approach various kinds of responses from students and to ensure that interviewers did not interject elements of bias into the data by introducing vocabulary or ideas that the students did not first initiate. Using feedback from these critiques, the protocol was adjusted to minimize any confusion about how the interview was to be conducted and to clarify how questions were to be asked to reveal student thinking.

The 14 interviews were conducted the same day the posttests were administered (N=5) or in the two days immediately following posttest administration (N=9). All 14 students who initially were selected by the teacher to participate agreed to be interviewed by the SCALE-uP staff. Students were interviewed for 20-25 minutes either during their science class or during their lunch period. Interviews took place either in the science preparation room adjacent to the students' science classroom or in the science classroom, if it was not being used for class. For each interview, the interviewer and student were

accompanied only by a video camera, which the interviewer set to begin filming when a subject entered the room. One of the interviewers had been present in the classroom prior to the interviews and was familiar to students. However, the interviewers did not have enough time to develop a strong rapport with the students prior to the interviews. During the interviews, individual students were seated at a table across from the interviewer to facilitate the manipulation of objects in the assessment items and to provide the student with space to write. Most of the students (13 of 14) were able to respond to five assessment items during the interview. One student was able to respond to only four items due to time constraints in the interview setting. Most of the students (12 of 14) were able to respond to six of the seven assessment items. (See Table 1 in Chapter I for more details.)

After digitizing and storing the video data on a central disk in the George Washington University's Discourse Lab, each video file was loaded into a visual qualitative data analysis software program, Atlas.TI©, for segmentation, transcription, and coding. Segmentation of the video file is a feature of the Atlas.TI software that allows the researcher to break a longer segment of video into shorter "clips" that are more manageable for transcription purposes. In these interview data, a segment typically represented a "turn-at-talk" for the interviewer or the interviewee. A turn-at-talk can be a short statement that the student or interviewer makes (i.e., "Yes" or "I don't know"), or it can be a longer statement such as a few complete sentences. A GWU undergraduate student, trained by SCALE-uP researchers to use Atlas.TI, completed the segmenting and transcribing of the interviews.

Data Analysis

This study was an analysis of the videotaped and transcribed interview data from these 14 students in relation to their responses to the seven paper and pencil MFA test questions. This section discusses the procedures used to analyze the videotaped interview data collected to address the research questions in this dissertation. The written MFA assessment data had previously been analyzed at both the item and test levels and stored by SCALE-uP researchers. The dissertation researcher was blind to the written results for these 14 students, both at the item and whole test level. The student ideas to be uncovered from analysis of the interview data were organized by assessment item and the coding of the data involved an analysis of student responses relative to each of the assessment item contexts (e.g., rolling ball or skateboard). This allowed the author to arrive at judgments regarding student understanding relative to each of the individual ideas that were targeted by a given assessment item and context. In addition, this allowed for conclusions to be drawn about the nature of student understanding relative to the particular target idea of interest across the set of items. By reducing and organizing the data in this way, it was possible to make comparisons between judgments about students' understanding on a given item as determined by the student's responses on the written assessment and judgments made about students' understanding on the same item from the videotaped interviews. Thus, a major goal of this dissertation was to arrive at an inference from the interview data for each student regarding their understanding of a given idea in a particular context.

Phase I: Code Category Development

Based on the review of research of student learning in this area (directly from Chapter II), codes were proposed that represented known, documented findings of student ideas about forces and motion (see Table 3 in Chapter II). These 13 common student ideas shown in Table 3 represent statements that summarize common research findings about ideas known to be held by middle school students regarding forces and motion. However, given the scope and limitations of the seven items used for this study, many of these assessment items were not designed to uncover student thinking about some of the common ideas presented in Table 3 (e.g., gravity). Therefore, it was necessary to arrive at a more focused set of student ideas to be used to characterize and categorize each of the 14 students' responses to the assessment items. This more focused set of student ideas represented the codes to be applied to each of the interview transcripts.

Although the student ideas presented in Table 3 (Chapter II) are known to be commonly held by many middle school students, the nature of the seven assessment items in this study did not allow for reliable detection of many of them. For example, in this dissertation study, none of the assessment items is explicitly targeting students' understanding of the nature of gravity. Although students may commonly discuss their ideas about gravity during the interviews (i.e., while responding to *Dropping and Throwing a Ball*), the coding procedures did not explicitly seek to categorize student thinking based on misconceptions about gravity. Instead, the coding procedures sought to determine the nature of student thinking around the given target idea (e.g., changes in speed or direction are caused by forces). Therefore, several common student ideas from

Table 3 that explicitly referred to students' ideas about gravity were not coded for in the data analysis (#8, #9, #10, #11, #12). However, the conclusions of this dissertation include a discussion of how an assessment context that involves gravity affects students' ability to demonstrate understanding of the different forces and motion target ideas. In addition, idea #13 (force as an acquired property of objects that move) was combined with idea #6 (objects in motion have a force in the direction of their motion), as it would be difficult to distinguish these two ideas.

Therefore, Table 5 shows the relevant common student ideas and the target ideas for the contexts represented in the MFA assessment items under investigation in this study (i.e., falling objects, rolling objects with friction, and objects at rest). These codes represent ideas that were known from the literature to be held by middle school students and seemed likely to be evident in transcripts of student responses to the seven assessment items that are part of the interview protocol. The list also includes the relevant target ideas (C_a, C_b, and C_c) to which the items are intended to be aligned. These 10 ideas are now referred to as "Student Idea Codes."

In an effort to arrive at an acceptable level of reliability for the coders, a first round of coding involved randomly selecting 7 of the 14 interviews to be coded. As part of this study, two individuals coded these seven interviews (the dissertation researcher and an education doctoral student at GWU with an advanced degree in physics and prior familiarity with the assessment items). The interviews were coded by each coder reading an individual interview transcript. For each transcript, the coder first determined which statements made by students were *Relevant* and which were *Not Relevant*.

Table 5. *Student Idea Codes Used for Initial Coding of Interview Transcripts.*

Code	Student Idea
1	Inanimate objects try to stop (or keep going) on their own accord.
2	Students associate amount of motion with amount of force.
3	Constant force results in constant motion
4	Decreasing motion (slowing down) is the result of decreased force or force running out as opposed to a force (i.e., friction) in the opposite direction.
5	Forces must act if objects are to stay in motion.
6	Objects in motion *have* a force in the direction of their motion (forces are acquired property).
7	If a body is not moving, there is no force acting on it.
C_a	Changes in speed or direction are caused by forces.
C_b	Objects in motion stay in motion unless acted on by a force.
C_c	Objects at rest stay at rest unless acted on by a force.

Note: Codes 1-7 are misconceptions and were used in data coding; C_{a-c} are the three target ideas.

Relevant statements are the statements students make that provide evidence of a student's understanding or not understanding the ideas of forces and motion under investigation. *Not Relevant* statements are statements that are unrelated to a student's understanding or not understanding of the target ideas of forces and motion. These include procedural statements, questions about the interview, or the reading aloud of assessment items. In addition, some statements could not be placed in either category because they were not interpretable by the coder. The next step for the coders was to arrive at judgments that corresponded to Student Idea Codes for the students' *Relevant* statements. Student Idea Codes were then independently assigned to as many of the *Relevant* statements as possible according to each coder's judgment.

The ten Student Idea Codes were used for a first round of coding by the two interview coders. All coding of interviews was completed using only the interview

transcripts (the video data will be discussed below). Appendix B shows tables for each assessment item that were used to code student responses. The interview coders consulted the tables in Appendix B for guidance on how to code particular interview data for a given assessment item. Appendix B includes details about the kind of *Evidence Statements* that the interview coder looked for in the interview transcript. Evidence statements represented statements students made during the interview that could be interpreted as evidence of students' ideas (both accurate and inaccurate) about forces and motion. Student Idea Codes were assigned, based on the Evidence Statements that were observed in each student's transcript. The assigning of codes was done for each student's transcript. When Student Idea Codes were applied to a specific portion of an interview transcript, the interview coder highlighted the portion of the transcript as being an Evidence Statement for that Student Idea Code. At the end of coding for each interview, the coder listed all Evidence Statements and Student Idea Codes for each item.

For each item, a student's evidence statements were coded in one of four of the following ways:

1. Single Student Idea Code. The coding of student responses resulted in the same code (either for target ideas or for misconceptions); then the corresponding Student Idea Code that accompanied the responses was applied.
2. Multiple Student Idea Codes. The coding of student responses resulted in different codes such that a student appeared to hold multiple misconceptions. Multiple Student Idea Codes were applied.
3. Correct and Incorrect Student Idea Codes. The coding resulted in a profile of codes that suggested the student holds conflicting correct and incorrect ideas

for a single item. In these cases, the dissertation researcher resolved the conflict in a way that considered all evidence to arrive at a conclusion about the student's ideas based on which expressed ideas were strongest.

4. No Code. If a student was unable to provide any evidence consistent with any of the Student Idea Codes, then no code was applied.

After the coding of the initial seven interviews was completed, the two raters reconciled with 100% agreement differences encountered in the assigning of Student Idea Codes for the seven interviews. This reconciliation process proceeded smoothly. There was general agreement between the raters when students had demonstrated an accurate understanding of science ideas, with only minor differences occurring in their decisions regarding the coding of inaccurate science ideas. The results from the coding and reconciliation are presented in detail in Chapter IV.

Based upon revisions to the coding categories, the two coders completed a second round of coding on a different set of seven randomly selected interviews. Four of these new seven interviews overlapped with the first seven interviews that were selected. (See Figure 9 for more details about interview sampling for data analysis.) Thus, three of the interviews were newly selected to ensure that the coding categories could be applied to new data that had not yet been considered. Based on the second round of coding, inter-rater reliability was checked to insure consistency of code usage across the two coders. A measure of percentage agreement was calculated, and it was determined that the coders agreed on 55 of the 61 possible codes applied across all seven interviews (90% agreement). Based on this second round of coding, it was determined that the Student

Idea Codes were functioning sufficiently well to justify moving to the final round of coding.

Phase II: Final Coding of Interview Data

During this final phase of the data analysis, the Student Idea Codes finalized from the iterative process in Phase I were applied to all 14 interview transcriptions. The interviews were coded by the same two individuals involved in Phase I of the data analysis. Each interview coder coded 10 of the 14 interviews such that six interviews overlapped and were coded by both raters (see Figure 9). These six interviews

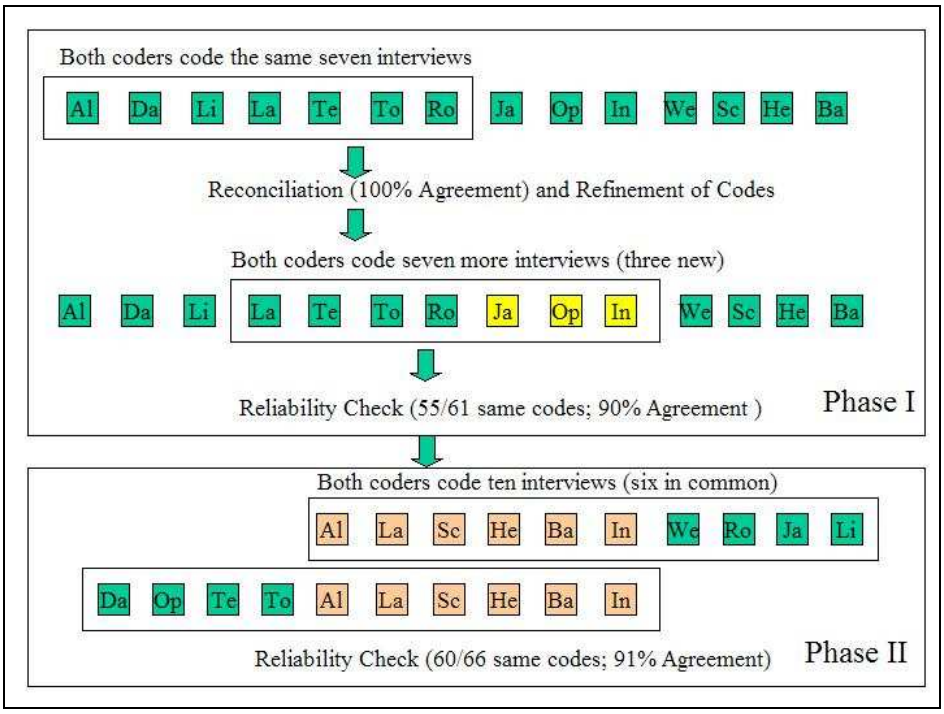


Figure 9. Sampling of interview transcripts for coding.

were used to calculate final inter-rater reliability (91%). The two coders agreed in the coding of 60 of the 66 evidence statements on these six transcripts. Upon completion of

coding of all 14 transcripts, the coding data was compiled into table format that allowed for analysis and interpretation of the oral interview data relative to the inferences made by SCALE-uP's written MFA assessment (see Table 8 in Chapter IV).

The Coding of Transcripts for Research Question #1

The first research question asked whether judgments made from the written assessment items accurately reflect what was revealed during the student interviews when students were probed about their responses to those same written items. As a first step to answering this question, it was necessary to compile the codes that were assigned for the interview data and the judgments that were made by SCALE-uP for the written MFA instrument. This was done for each of the items across all 14 students. To allow simple comparisons between student responses for each item, SCALE-uP's MFA posttest item-level ratings for correctness and incorrectness were paired to the final version of the Student Idea Codes. Once this was accomplished, it was possible to assemble the separate judgments for each of the two data sources (written posttest and oral interview) for each item and for each student. This information was entered into Table 9 (see Chapter IV).

This allowed for determination of the extent to which student responses to written items matched student thinking in interviews. Given that not all items have the same number of students responding, written and interview data were judged to be in agreement if there was agreement for at least 13 of the 14 students, i.e., less than 10% disagreement for a given item. Mismatches between written and interview data included false positives (students who correctly responded to assessment item on initial written assessment but did not appear to understand the target idea during the interview) and false negatives (students who incorrectly responded to item on written assessment but

showed evidence of understanding the target idea during the interview). In other words, if the evidence from the written assessment and the evidence from the interview did not align in terms of the misconception elicited for a given student, this was not considered evidence of disagreement for the sake of judging item effectiveness. So, evidence for disagreement only included those cases where the MFA written item provided an inference that was different from the students' understanding in terms of the correctness or incorrectness of the student's response. A measure of disagreement was computed for each of the seven assessment items. This analysis allowed for a response to the first research question regarding how well each written assessment item revealed student understanding of the target ideas when compared to independent judgments made from interview data.

Holistic Judgments to Address Research Question #2

The second research question was interested in the consistency with which students demonstrate their understanding of the same forces and motion ideas targeted by a variety of assessment task contexts in both a written assessment format and during a videotaped interview. To address the second research question, each student's responses were examined for their consistency across the different assessment item contexts.

In order to address this research question, a holistic judgment method was used with the videotaped interviews. The two science educators who conducted the interviews for SCALE-uP (David and Ann) viewed video recordings of entire interviews and holistically arrived at a judgment about the level of a given student's understanding across all of the item contexts. Each interviewer watched the interviews that the other conducted to minimize bias in coding. They holistically judged the video evidence of

each student's understanding of the forces and motion ideas and categorized that evidence as follows:

1. *Student Mostly Understands.* Evidence from the videotaped interview indicated that the student, in general, understood most of the forces and motion ideas (at least two of these targeted ideas) across each of the item contexts. A student who mostly understands may show some uncertainty about their ideas, especially when probed during an interview or when attempting to explain the ideas in a difficult context. However, these students are mostly accurate in their use of the forces and motion ideas to explain the various phenomena.
2. *Student Somewhat Understands.* Evidence from the videotaped interview indicated that the student understood some of the forces and motion ideas (at least one) in some of the item contexts. A student who somewhat understands demonstrates evidence that they are beginning to use at least one of the forces and motion ideas to explain phenomena in a way that that is scientifically accurate.
3. *Student Does Not Understand.* Evidence from the videotaped interview indicated that the student understood none of the forces and motion ideas in any of the item contexts. Students who do not understand are unable to provide evidence that they are able to use the forces and motion ideas to explain any of the phenomena.

Interviewers viewed the videotaped interviews and assigned each student to one of the above categories and then provided a brief rationale for their decision. Once the

interviewers assigned each student to one of the above categories based on their judgment of the student's level of understanding during the interview, this was then compared with a given student's level of performance on the written MFA posttest. (All students were scored on the MFA's four levels of understanding; see page 82). By using these two holistic judgments, it was possible to arrive at a conclusion regarding each student's ability to consistently demonstrate their understanding of forces and motion understanding across a variety of contexts during both the oral interview and the written MFA posttest, and it was possible to compare the holistic judgments arrived at by using written responses and interview data.

Statistically, a raw score calculated from the seven items involved in the interview protocol was strongly correlated to a student's overall performance on the 10-item written MFA ($r = 0.70$, $p < 0.01$). Therefore, a given student's performance during the interview was a good indicator of performance on the written MFA, thus justifying the above comparison. Of particular interest were cases in which students were consistently characterized by the two judgment processes as demonstrating evidence of understanding the ideas or as not demonstrating evidence of understanding.

It was hypothesized that students characterized as high performers according to the written MFA instrument would also be characterized as high performers by the holistic judgments made by the interviewers. Conversely, the researcher expected that students characterized as low performers according to the written MFA instrument would also be characterized as low performers by the holistic judgments made by the interviewers.

Contributions to Literature Through Research Question 3

To address the third research question about whether the ideas students revealed in this study have already been reported in the current research literature, the author considered any new Student Idea Codes that needed to be added during early phases of data analysis. These new codes represent instances in which evidence collected during the interview about students' ideas was not currently captured by the review of the literature presented in Chapter II. The analysis focuses on how ideas uncovered from this diverse sample of middle school student thinking were or were not different from the findings in the current literature. In addition, this analysis comments on what was learned from answering the second research question. Specifically, the dissertation discusses how analysis from the second research question adds to what is known about the role of item contexts in interacting with students' ideas about forces and motion during responses to aligned assessment items.

Human Subjects and Ethics Precautions

This study posed minimal risks to the subjects involved. As part of the SCALE-uP research process, all students and their parents signed informed consent forms authorizing the video recording of classroom events as well as providing permission for students to be interviewed by SCALE-uP research staff. As part of this study, an expedited Internal Review Board (IRB) review of the proposed research was completed. Because video data were used, it was possible to identify students. Thus, precautions were taken to ensure that students were not referred to or described in any way that could reveal their identities. This study used the SCALE-uP assigned pseudonyms to refer to students. In addition, SCALE-uP data sources that were consulted (including student posttests) had

been de-identified as to not allow for personal information about students to be ascertained. As per MCPS policy, when student work (including assessments) was released to SCALE-uP researchers, such data were completely anonymized. In all writing and analysis for this study (dissertation and beyond), only the students' pseudonyms were used.

CHAPTER IV: RESULTS

The study explored the extent to which paper and pencil assessment items aligned with specific middle school level forces and motion ideas accurately measured those ideas. Analysis of videotaped interviews with a sample of sixth grade students determined how the ideas these middle school students used when responding orally to items from an assessment instrument compared to the ideas they used when responding in writing to the same items on a posttest.

Organization of the Chapter

This chapter presents the results of the transcript coding and data analysis. The chapter begins with a presentation of summary statistics for the transcribed student interviews and then discusses the how data coding proceeded. A sample of a coded interview is presented along with the final frequency data for the set of student idea codes used. The results from the coded data are then compared to each student's written responses to answer the first research question. Results from the holistic judgments are then presented for the second research question followed by the presentation of new student idea codes to address the third research question.

Characteristics of the Video Data

Videotaped interviews of 14 students totaled 347 total minutes of video data. The average interview lasted 24 minutes and 46 seconds; the longest interview 34 minutes; and the shortest interview lasted 16 minutes. The 14 students spoke an average of 1383 words per interview with a total of 19,355 total words spoken by students across all interviews. Students made a total of 1182 statements during the interviews with the average student making 84 statements during an interview. The interview coders

determined that 643 of the 1182 total student statements were *relevant*, and 539 were considered n*ot relevant.* (See Chapter III.) These data for each student are shown in Table 6.

Table 6. *Characteristics of Videotaped Interviews for All Students.*

Student	Interview Length (min)	Words spoken	Total Statements	Relevant Statements	Not Relevant Statements
Barry	24	1344	72	44	28
Alan	17	952	59	30	29
Wendy	27	1566	113	63	50
Tom	24	1320	84	47	37
Teddy	34	1904	111	57	54
Scott	21	1176	93	51	42
Robin	27	1485	85	40	45
Opal	27	1512	63	27	36
Lisa	16	896	44	20	24
Larry	27	1458	92	42	50
Janet	19	1064	73	49	24
Ingrid	26	1430	95	60	35
Heidi	30	1680	111	69	42
Darlene	28	1568	87	44	43
Total	347	19355	1182	643	539
Average (SD)	24:46	1383 (282.19)	84 (20.57)	46 (13.93)	39 (9.74)

Overview of Data Coding

As discussed in Chapter III, data analysis of the first research question proceeded in two phases that were designed to achieve consistency and reliability in the use of Student Idea Codes. After the coding of the initial seven interviews was completed, the two raters were able to reconcile with 100% agreement differences identified in the assignments of Student Idea Codes. There was general agreement about when students

had demonstrated an accurate understanding of science ideas. There were minor differences in decisions about inaccurate science ideas and the corresponding Student Idea Code. These differences were attributable to different interpretations of the Student Idea Code statements. This resulted in clarification of some code statements to improve the clarity of each code's intention. For example, because Student Idea Code 4 included the phrase "decreased force or force running out," the coders could apply this code when other codes could have been used such as Code 6 (forces must act if objects are to stay in motion) or Code 2 (students associate amount of force with amount of motion). Therefore, Code 4 was revised to read, "Forces in the opposite direction of motion are not the cause of slowing down." In addition, two new codes (Codes 8 and 9) related to force and friction were added to the existing pool of Student Idea Codes based on evidence from the first round of coding. These new codes were added because both raters agreed that specific interview evidence did not fit into one of the existing code categories, and a new category was needed to capture a clearly stated student idea that was emerging in the transcripts. The final table of Student Idea Codes is presented in Table 7.

During the final data analysis, the Student Idea Codes finalized from the iterative process in Phase 1 were used to code student statements in all 14 videotaped, transcribed interviews. The interviews were coded by the same two individuals involved in Phase 1 of the data analysis. Each interview coder coded 10 of the 14 interviews so that six interviews were coded by both raters. These six interviews were used to calculate the final inter-rater reliability statistics. The two interview coders agreed on 60 of the

Table 7. *Final Student Idea Codes Used to Code Interview Transcripts.*

Code	Student Idea

1	Inanimate objects try to stop (or keep going) on their own accord.
2	Students associate amount of motion with amount of force.
3	Constant force results in constant motion
4	Forces in the opposite direction of motion are not the cause of slowing down.
5	Forces must act if objects are to stay in motion.
6	Objects in motion *have* a force in the direction of their motion (forces are acquired property).
7	If a body is not moving, there is no force acting on it.
8	Friction causes a force to run out (i.e., with rolling objects).
9	Friction does not act in a given direction OR friction is not a force.
C_a	Changes in speed or direction are caused by forces.
C_b	Objects in motion stay in motion unless acted on by a force.
C_c	Objects at rest stay at rest unless acted on by a force.

Note: Codes 1-9 are misconceptions and were used in data coding; C_{a-c} are the three target ideas.

possible 66 (91%) statements to be coded in these six transcripts. On a few occasions, one of the coders applied a Student Idea Code and the other coder did not apply a code. On a few other occasions, there was disagreement in the Student Idea Code that was applied.

Sample of Coded Interview

The following section presents a short sample of a coded interview. This depicts how interview transcripts of student statements were first judged to be relevant or not relevant and how evidence statements were considered to be consistent with specific Student Idea Codes. This short segment represents a portion of the interview conducted with Barry during the skateboard assessment item:

> David: Alright, this is the third one. It's a little different because it is multiple choice. So if you could read that to me and all the answer possibilities.
>
> Barry: <R Carly gets on her skateboard and pushes off once with her foot. As she moves across the gym floor on her skateboard the skateboard slows down and comes to a stop. Why does the skateboard come to a stop? Fill in the oval beside

your answer. Because there is a force in the opposite direction of the skateboard's motion, because the skateboard runs out of force, because the force of the push on the skateboard decreases to zero, because there aren't any forces on the skateboard R>.

David: Okay. So, take a minute, and choose the best answer. Fill in the oval next to it.

COMMENT: "((BARRY TWIRLS PENCIL AS READING))"

COMMENT: "((BARRY MARKS ANSWER))

Barry: Okay.

Not Relevant Statements

David: Okay, which one did you choose?

Student Idea Code C_a –Changes in speed or direction are caused by forces.

Barry: I did, <R because there is a force in the opposite direction of the skateboard's motion R>.

Evidence Statement

David: Okay. Um, so let's, again, talk a little bit about the question and a little bit about your answer. Can you explain what this question is about?

Barry: It was asking when a force was applied to a skateboard, what caused it to stop?

Evidence Statement

David: Okay. And can you explain why you chose the answer you did?

Barry: I chose that because we learned in the experiments that if you apply force it, it doesn't really stop until something else stops it. So it wasn't because the skateboard runs out of force it was because of the friction, the force in the opposite direction.

Student Idea Code C_a –Changes in speed or direction are caused by forces.

David: Okay, and what is friction?

Barry: Friction is when two objects touch and they just create friction when they touch. It's like another force.

Relevant Statements

David: Okay.

The partial transcript from Barry's interview shows how verbal report data were first judged by the coders to be *relevant* or *not relevant* and then judged to be consistent with a Student Idea Code. The coding procedures produced a number of Student Idea Codes

for each student that were assigned to statements made during the interview. These data are presented in Table 8 to show the number of Student Idea Codes assigned to statements made by each student during the interview.

Table 8 shows the frequency and variety of Student Idea Codes assigned to the evidence collected during each student's interview. The frequency of Student Idea Codes varies considerably, which indicates that some student ideas were more prevalent in this sample of students. The frequency of Student Idea Codes is also dependent on the amount

Table 8. *Frequency of Student Idea Codes for Each Student Based on Interview Data.*

Student	Student Idea Codes											
	1	2	3	4	5	6	7	8	9	C_a	C_b	C_c
Barry		1	3		1	2				6	2	
Alan					4			1		2		2
Wendy					3	2				4		4
Tom			1		1				1		5	
Teddy		1			2	1				3		
Scott		2								3	2	2
Robin		2			2					3		2
Opal		2						1		1	1	
Lisa			1		2	2						
Larry					4					3	2	
Janet										7	1	
Ingrid					3	1			1	1	4	
Heidi					5		2			3		
Darlene		1								2	1	
Total	0	9	5	0	26	8	2	2	2	38	18	10

of data collected for a given item that targeted a specific idea. For example, fewer students responded to the "Metal Ball" assessment item. Therefore, it is not surprising that there was relatively less evidence of students understanding target idea C_c (objects at

rest stay at rest unless acted on by a force) across all students. As depicted below, Figure 10 shows the frequency of Student Idea Codes for all students during the oral interviews.

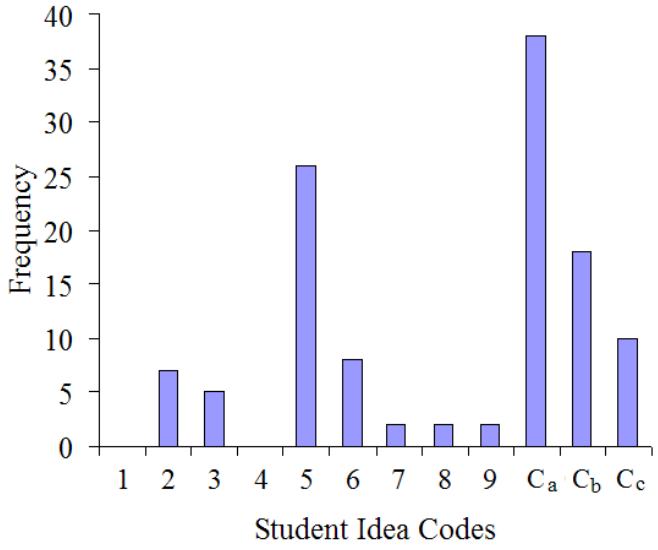


Figure 10. Frequency of student idea codes assigned across all students.

Research Question 1

This section of Chapter IV presents data to be used to answer the first research question:

> To what extent do the inferences at posttest from each of seven paper and pencil science assessment items (aligned by expert judges using specific alignment criteria to particular forces and motion ideas) agree with the inferences about students' understanding of those ideas as determined during videotaped oral interviews at posttest?

This section of Chapter IV addresses each of the seven assessment items under investigation in this dissertation. The results from the data coding procedures are presented in Table 9 for each student on an item-by-item basis. The data show the results of the final coding of interview transcripts (shown as "oral" in Table 9) along with the students' item-level written MFA posttest results as coded by SCALE-uP researchers (shown as "write" in Table 9). In Table 9, a "C" represents an instance of a student's written response rated by SCALE-uP on the written MFA posttest as consistent with evidence of understanding motion and forces ideas. In Table 9, an "IC" represents an instance of a student's written response rated by SCALE-uP on the MFA posttest as "incorrect" relative to an understanding of forces and motion ideas. All numbers (1-9) shown in Table 9 correspond to Student Idea Codes (Table 7) as determined by coders who looked for evidence in the interview transcripts. The occurrence of C_a, C_b, and C_c in Table 9 represent instances in which students' responses to specific assessment items during the interviews were considered as evidence of understanding one of the three target ideas of this study.

Table 9 also presents the level of disagreement for each item across all students. The shaded areas in Table 9 represent instances in which written MFA posttest evidence collected and rated by SCALE-uP researchers disagreed with the oral interview evidence collected and coded during this study. These disagreements only highlight the instances in which the inference made by SCALE-uP researchers on the written MFA posttest about whether or not students understand a specific target idea does not agree with the inference made about students after coding transcripts of the oral interview data. The

following sections discuss the evidence collected for each of the assessment items under investigation in this study.

Rolling Ball—Selected Response

The first item is the selected response portion of the rolling ball assessment context. For this item, students were asked about a ball that was kicked once, "Is there a force constantly pushing the ball to you?" Students could select between "yes" and "no" answer choices. All 14 students responded to this item during the written MFA posttest and during the videotaped interview. Eight of the students answered correctly during the written MFA posttest (selecting "no") while six of the students responded incorrectly (selecting "yes") on the posttest. Of the eight students who did not believe there was a force constantly pushing the ball during the written MFA posttest, seven students responded in the same way during the interview, and further follow-up questioning by the interviewers revealed the same student ideas. Of the six students who believed there was a force constantly pushing the ball during the written MFA posttest, five responded similarly during the interview and during follow-up questioning. This led the coders to conclude that these five students thought there was a force constantly pushing the ball as it rolled.

There were two instances of disagreement for the rolling ball selected response item as evidence from the written MFA posttest did not agree with the coded interview data. One student, Robin, responded on the written MFA posttest that there was not a force constantly pushing the ball as it rolled but then responded during the interview in the following way:

Ann: And what was your answer?

Table 9. *Judgments Made for All Students Interviewed per Item During SCALE-uP's Written MFA Posttest and Oral Interview.*

Student	Rolling Ball (SR)		Rolling Ball (CR)		Falling Ball (SR)		Falling Ball (CR)		Carly's Skateboard (SR)		A Metal Ball(SR)		Pedaling a Bike (SR)	
	Write	Oral	Write	Oral	Write	Oral .	Write	Oral	Write	Oral	Write	Oral	Write	Oral
Alan	IC	5	IC	5	C_a	C_a	C_a	C_a	IC	5,8	C_c	C_c	-	-
Barry	C_b	C_b	C_b	2,5	IC	3	IC	3	C_a	C_a	-`	-	IC	6
Darlene	C_b	C_b	IC	No code	C_a	C_a	IC	2	C_a	C_a	-	-	-	-
Heidi	IC	2	IC	5	C_a	C_a	C_a	C_a	IC	5	C_c	7	-	-
Ingrid	C_b	C_b	C_b	C_b	C_a	C_a	C_a	C_a	IC	5,9	-	-	C_b	6
Janet	C_b	C_b	C_b	C_b	C_a	C_a	C_a	C_a	C_a	C_a	-	-	IC	C_b
Larry	C_b	C_b	IC	5	C_a	C_a	C_a	C_a	IC	5	IC	No code	-	-
Lisa	IC	5	IC	5,6	IC	3	IC	No code	-	-	-	-	-	-
Opal	C_b	C_b	IC	No code	IC	3	IC	No code	IC	2,8	-	-	-	-
Robin	C_b	5	IC	5	C_a	C_a	IC	C_a	IC	2	C_c	C_c	-	-
Scott	IC	C_b	IC	C_b	C_a	C_a	IC	2	C_a	C_a	C_c	C_c	IC	No code
Teddy	IC	5	IC	5	C_a	C_a	IC	2	C_a	C_a	-	-	C_b	6
Tom	C_b	C_b	C_b	C_b	IC	3	IC	3	C_a	5,9	-	-	IC	C_b
Wendy	IC	5	IC	5,6	C_a	C_a	IC	2	C_a	C_a	C_c	C_c	-	-
Total Disagreement	2/14		2/14		0/14		1/14		1/13		1/6		4/6	

**SR=Selected Response; CR=Constructed Response. Shaded cells denote instances of disagreement between evidence sources (written MFA and oral interview). C = Correct understanding; IC = Incorrect understanding. C_a, C_b, and C_c denote evidence of understanding each of the motion and forces target ideas. No Code= No student statements were able to be coded.*

Robin: Yes.

Ann: And you said yes there was a force constantly pushing the ball, and why was that?

Robin: Because it made sense to me because if there wasn't a force constantly pushing it, then the ball wouldn't have made it towards me.

Robin responded on a few other occasions during the oral interview that she believed there was a force pushing the ball. This is in contrast to her written MFA posttest response in which she circled "no," that there was not a force pushing the ball as it rolled. Another student, Scott, responded on the written MFA posttest that there was a force constantly pushing the ball by circling "yes" but then during the interview selected a different answer, as presented in the following statements:

Ann: And how did you know not to answer yes?

Scott: I didn't answer "yes" because the force is not pushing, constantly pushing the ball to me because friction is pushing the ball the opposite direction of me and force is pushing the ball to me when the person kicks it, and not further than that.

Ann: Not further?

Scott: Not further than when after the person kicks it.

Scott's response during the interview contradicts how he responded during the written MFA posttest. Therefore, for Scott and Robin, inferences from SCALE-uP's written MFA posttest responses did not agree with the inferences from the transcribed interview data.

For the rest of the students' responses, there was agreement between the written MFA posttest and the coded interview data. These 12 students' written MFA posttest responses of "yes" and "no" to this selected response item matched data obtained during the videotaped interviews. In these cases, the inferences made from SCALE-uP's written

MFA posttest item were in agreement with the inferences made from the transcribed interview data.

In conclusion, for this rolling ball selected response item, there were two instances of disagreement when comparing the inferences from the written MFA posttest with the inferences from the interviews. In Robin's case, she probably guessed correctly on the written item (there were only two answer choices, yes and no) but clearly revealed a lack of understanding during the interview. The lack of agreement in Scott's case is harder to explain. Perhaps his knowledge is not yet secure enough to answer consistently from testing to testing. There are some hints of this in his interview statements as he discusses friction, which is not relevant to this idea. In all, this represented 14% disagreement (2 out of 14) between the two data sources for the students with this item. This level of disagreement is higher than the 10% disagreement cutoff for this study. However, a certain level of disagreement between data sources should perhaps be expected for an assessment item that provides two answer choices ("yes" and "no").

Rolling Ball Constructed Response

The next question from the rolling ball assessment context asked students to "explain why the ball moves to you after it is kicked." This was one of the two constructed response items in this study. During the written MFA posttest, 4 of the 14 students provided statements that were coded as evidence of understanding the target idea (objects in motion stay in motion unless acted on by a force) by SCALE-uP researchers. Of these four students, three provided statements in the interviews that were coded as evidence of understanding the target idea. Statements made by these three students during the written MFA posttest and the oral interview included:

> Janet (written posttest): (after selecting "No") The ball moves to me because there is a force that kicked it toward me.
> (interview): (after selecting "No") There is another force, except it's not pushing it towards me. There is another force called friction pushing the opposite way.
>
> Tom (written posttest): (after selecting "No") It moves because the force of the kick that makes it move forward.
> (interview): (after selecting "No") No because after it's kicked there is no more force pushing it to get to me. (later) ...there is no other force pushing it along, there is just the force from that kick.
>
> Ingrid (written posttest): (after selecting "No") It moves because it is increasing speed from the force applied to the ball to make it move.
> Ingrid (interview): (after selecting "No") It moves because there is a force that pushed it and it moves until it recently comes to a stop. (later) … there isn't like a force constantly pushing it, because after you kicked it, you don't, you, eh, it just rolls to the person until it comes to a stop.

The fourth student, Barry, provided written evidence of understanding on the MFA posttest, which was in disagreement with the evidence from his written posttest response. During the written MFA posttest, Barry first selected "no" that there was not a force constantly pushing the ball and then wrote:

> Barry (written posttest): The ball moves to you after it is kicked because force was applied to the ball which made it move.

This was considered by SCALE-uP's assessment raters as evidence that Barry understood the target idea. However, during the interview, Barry was unable to provide consistent evidence of understanding the target idea, as he offered several statements that conflicted with prior statements. Barry's ideas in this context are revealed in the following exchanges:

> Barry: The ball moves toward me because force was applied to the ball to make it move.
>
> *A few seconds later...*
>
> David: Okay, why is there no constant force pushing?

> Barry: Because there was just one kick, there is no force still kicking it towards the person.

This was considered to be evidence of understanding by the interview coder. However, when asked another follow-up question from the interview protocol about how the ball keeps moving, Barry responds in the following way:

> David: So, if there is no force constantly pushing the ball, how does it keep moving?
>
> Barry: Because the force was applied to it. And the force that was applied to the ball hasn't worn off yet.

This led the interview coder to conclude that the statement was consistent with Student Idea Code 5 (forces must act if objects are to stay in motion). Later, Barry made a few statements about how the motion of the ball is dependent on the amount of force that is making the ball move. Both coders agreed that these statements were evidence of Student Idea Code 2 (amount of motion related to amount of force). Therefore, the evidence collected during the interview was inconclusive.

The remaining nine students did not provide evidence of understanding during SCALE-uP's written MFA posttest. Many of their written responses were coded as "high inference" by SCALE-uP because the raters were unable to make definitive inferences from the students' written response. During interviews, these students were unable to provide evidence of understanding of the target ideas. One exception (Scott) will be discussed below. Other students (Alan, Wendy, Heidi, and Larry) provided written statements on the written MFA posttest that were more clearly evidence of a common misconception and they were coded as such. In these cases, the students made similar statements that were inferred to be evidence of analogous misconceptions during the

interviews (i.e., forces must act if objects are to stay in motion). Two examples of these student ideas are:

> Alan (written postest): (after selecting "Yes") It is because the force that kicked it was aiming at you. The ball can't move on its own so some kind of force has to move it.
> (interview): So the ball moves to me after it is kicked, because objects don't move on their own, so obviously there is a force pushing it to me.
>
> Wendy (written posttest): (after selecting "Yes") The ball moves because there is more force pushing the ball forward than pushing the ball backward.
> (interview): Because if the force didn't keep going then the ball would slow down. So I thought….the ball keeps going and if the force stops being applied then it will slow down.

Therefore, these responses were considered to be evidence of Student Idea Code 5 (forces must act if objects are to stay in motion). This Student Idea Code corresponded well with SCALE-uP's classification of these students' written responses at posttest. Both Alan and Wendy's written posttest responses were classified by SCALE-uP as an "alternative conception of force."

Larry provided some initial evidence of understanding the target idea but, during follow-up probes, offered conflicting evidence in the following exchange:

> Ann: If there is not a force constantly pushing the ball, what is causing it to move?
>
> Larry: ….hmm….now I'm trying to think towards yes. Because if it's moving, then it should be applied with force all the time.
>
> A few seconds later….
>
> Ann: ….is there something always pushing the ball?
>
> Larry: Um, yes.
>
> Ann: What is that?
>
> Larry: Force.

> Ann: So, what is that force, can you describe it?
>
> Larry: I think it's like. Um, when someone pushed something, where like causes it to move, they use like strength, and I guess of the amount of strength is like certain amount of force used to it to make it go a certain distance.

It was determined that several of Larry's statements in the interview were consistent with Student Idea Code 5 (forces must act if objects are to stay in motion). Larry had previously selected "no" that there was not a force constantly acting, but then changed his answer during the interview. Larry's responses during this assessment context may indicate that he is struggling to understand these ideas in this rolling ball context. This was not in disagreement with Larry's written MFA posttest response in that SCALE-uP assessment raters were unable to make sense of his response. He wrote (after circling "no"), "there is force applied but there is no constantly moving because of friction." He seemed to be using the word "constantly" in this instance to describe the motion of the ball as opposed to the force acting. This written statement on the MFA posttest was coded as "high inference," as SCALE-uP raters were unable to definitively code the written evidence. Therefore, the coding of the transcript data did not disagree with the inferences from Larry's written MFA posttest responses. Larry was unable to provide definitive evidence that he understood the target idea.

Of the nine students who did not provide evidence of understanding the target idea on SCALE-uP's written MFA posttest, eight were also unable to provide evidence of understanding during the interviews. Oral responses by two students during the interview (Opal, Darlene) could not be definitively categorized by the interview coders. Both Opal and Darlene reported some confusion about the context of the item. Darlene mistook the question to be about the constant *motion* of the ball as opposed to whether or not there

was constant *force* acting on the ball. However, even after clarification of the assessment item, Darlene persisted in her discussion of the constant motion of the ball and was unable to provide evidence that demonstrated an understanding of the target idea. Opal also misunderstood the context and believed it was asking if the ball was applying a force to move itself. She correctly concluded that the ball does not apply a force to itself as it moves and provided some evidence consistent with a belief that there is some outside force acting on the ball as it rolls. However, the coders were unable to find clear evidence statements that corresponded to Student Idea Codes during this portion of Opal's interview.

There was one additional instance of disagreement for this item between the two evidence sources. In this instance, Scott did not provide evidence of understanding on the written MFA posttest but was able to provide evidence of understanding for the idea during the interview. During Scott's written MFA posttest response, he selected "yes" in response to a force constantly pushing the ball and then wrote:

> Scott (written posttest): The ball moves to me after it is kicked because a pushing force was applied to it which made it roll to me and since the ball is round it makes it hard for friction to stop the ball.

While Scott demonstrated a solid understanding of how friction acts, his response was coded by SCALE-uP raters as being incorrect (i.e., a misconception about the nature of forces). Scott replied differently during the interview, in which he responded that there is not a force constantly pushing the ball as it rolls by selecting "no" on the first part of this rolling ball context. Scott then stated that the force was only acting when the kick occurred and not "further than that" (as quoted previously). He later states that the ball rolls because "its round shape….and the force already applied" (statements which are

similar to his written MFA posttest response). During coding, these statements were considered to be evidence of Scott's understanding the target idea (objects in motion stay in motion unless acted on by a force). Although Scott's two responses appear to be similar, this dissertation study came to a different conclusion than SCALE-uP's written posttest regarding the inference about Scott's understanding.

In conclusion, for this rolling ball constructed response item, there were two instances of disagreement (Barry and Scott) when comparing the inferences from the written MFA with the inferences from the interviews. Barry's reasoning about the forces acting on the rolling ball seemed to breakdown while responding to probing questions during the interview and Scott's similar responses were coded differently during transcript analysis for this study when compared to the conclusion drawn by the written MFA posttest. This represented 14% disagreement between the two data sources with this item. This level of disagreement is higher than the 10% disagreement cutoff for this study. It is worth noting that similar student misconceptions were revealed for several students (Alan, Wendy and Heidi) during both assessment situations (written MFA posttest and the oral interview).

Dropping and Throwing a Ball—Selected Response

This context asks students to consider the motion of a ball that has been dropped by a girl from a window, and to answer the question, "What happens to the speed of the tennis ball as it falls?" Students are then prompted to choose between "speed increases," "speed decreases," or "speed stays the same." The question is intended to determine students' understanding of the idea that "changes in speed (or direction) are caused by forces." (A discussion of the specific case of gravity as a force will be given more

attention in detail in Chapter V.) All 14 students responded to this question on the written MFA posttest and also responded to this question during the videotaped interview.

Of the 14 students who responded during the written MFA posttest, ten students responded correctly that the speed of the ball increases as it falls. Four of the students responded that the "speed stays the same." In each of these 14 cases, the evidence collected during the interviews led to inferences that matched the students' written responses. (See Table 9.) Students who believed that the speed of the ball would stay the same as it fell (Opal, Lisa, Tom, Barry) responded in the same way during the interviews. (Key aspects of students' reasoning will be discussed in more detail in the presentation of the data for the constructed response portion of this assessment context.) For this assessment item, there was no disagreement between the two data sources for the students.

Dropping and Throwing a Ball—Constructed Response

The second part of the dropping and throwing a ball context asks students to "explain what caused the speed of the tennis ball to increase, decrease, or stay the same." The item is intended to determine if students understand the idea that "changes in speed (or direction) are caused by forces." All 14 students responded to this question on the written MFA posttest and during the interview. Five of the students provided evidence during the written MFA posttest that led assessment raters to infer that the students understood the target idea. In each case, the students (Alan, Larry, Janet, Ingrid, Heidi) provided evidence during the interview that led interviewer coders to infer that they understood the target idea. Statements made by some of these students during the written MFA posttest and during the interview included:

Larry (written posttest): The ball will speed up because gravity is pulling it down.
Larry (interview): I think it increases in speed because gravity is constantly pulling it down making it speed up as it falls. …I chose speed increases because I think, I remember usually if I drop something it will gradually get faster, and when I was in science we learned about gravity and how it pulls it down.
Ingrid: I said that the speed would increase and then I said since gravity is pulling it down and there is nothing to slow the ball down….it would continue falling until Maria caught it.

Janet (written posttest): The speed of the tennis ball increases because gravity is pulling down on the weight of the ball.
Janet (interview): It increases because gravity is pulling down on the weight of the ball.

Alan (written posttest): Gravity caused the speed of the tennis ball to increase. It's because as the ball goes down by gravity, the ball goes faster because there is no force that could stop it and gravity is pulling it faster.
Alan (interview): Because….because gravity pulls it out and as its going down and gravity is pulling, it, it along like harder, and since there is no friction, it will increase.

The other nine students could not provide written evidence on the written MFA posttest that they understood the target idea. Many of these students' written MFA posttest responses were coded as "alternative conceptions of gravity" by SCALE-uP assessment raters. These included responses that indicated misconceptions about the force of gravity (i.e., force of gravity changes, gravity pushes instead of pulls, gravity is constant so speed stays the same). Examples of these students' written responses on the MFA written posttest include:

Barry (written posttest): The speed of the tennis ball stays the same because gravity is pulling it down always at the same speed which means the tennis ball won't speed up or slow down.

Opal (written posttest): The tennis ball stayed the same because when you drop the ball it stays in one constant speed.

Lisa (written posttest): I think that the speed stays the same because of the the force she put to the tennis ball.

> Teddy (written posttest): The cause of the tennis ball to increase was gravity, as the ball was falling gravity grew stronger and made it go faster.
>
> Wendy (written posttest): The ball's speed increased because gravity was pulling it down and as the ball was dropping the force got stronger.
>
> Scott (written posttest): The speed of the tennis ball increases because at first when Katie let's go, gravity pull down the tennis ball slowly but as the tennis ball get's closer to the gravity starts to pull the ball to the ground harder and harder.

Another student, Darlene, wrote about the ball as the reason the ball's speed increased but did not mention a force as the cause of the increased speed.

Of these nine students, eight (Darlene, Lisa, Tom, Barry, Opal, Teddy, Wendy, Scott) were unable to provide consistent evidence of understanding the target idea in this context during the oral interview. This occurred for a variety of reasons. Darlene discussed the ball when describing why the speed of the ball increases as it falls when she states, "you can tell like the ball is going to pick up speed when it's falling." She later reveals her belief that smaller objects (i.e., an egg) fall faster than heavier objects (i.e., a watermelon) because smaller objects have less weight. (This idea will be discussed in more detail later.)

Lisa appears to misunderstand the question and believes that the tennis ball has been thrown as opposed to dropped. She states during the interview that she believes the speed will stay the same "because she didn't throw it that hard so the speed doesn't increase or decrease" and "because of the force put to the ball." Lisa's statements during this portion of the interview could not be coded in any of the categories. Tom also appears to misunderstand the question because he believes that the context involves comparing the speed of consecutive drops when he selected "speed stays the same" during the interview:

> Tom (interview): Because in science class we also did another experiment on like the same type of situation. And like, my teacher dropped a ball from like the counter and said, and he dropped it more than once, and it went about the same as it did before. (later)… because for like a moment I thought that gravity could pull something down faster that it can the second time, but it can't because I remember the little situation we had on the counter.

After a few attempts by the interviewer to clarify the context and describe the scenario as a single drop from a given height, Tom concluded that a ball dropped from higher up would "stay at a constant speed" in the following exchange:

> David: If that ball was dropped from really high up, do you think gravity would pull it faster then?
>
> Tom: No.
>
> David: No, Okay.
>
> Tom: It would stay at a constant speed. It would probably stay at a constant speed.
>
> David: Okay. And so you've been using this word gravity. Can you explain what you mean by that?
>
> Tom: Gravity is a force that makes things stay to the ground….it just pulls stuff to the ground.

In this exchange, Tom reveals his belief that a dropped ball's speed will remain constant because the speed of gravity is constant. This statement was coded as being consistent with Student Idea Code 3, or constant force results in constant motion.

As Barry did on the written MFA posttest, he again concluded that the "speed stays the same" because "gravity is constantly pulling the ball to the ground at the same strength, meaning the pull of gravity never gets stronger or weaker, it stays the same." This was coded as "constant force results in constant motion" (Code 3) by the interview coders. Opal also selected that the speed would stay the same during the interview:

> Opal (interview): As Katie drops the ball it just goes straight down, it does pick up speed, but only a little bit because when you drop the ball it just goes right down it doesn't stop or anything so I just chose it stays the same.

During the interview, Opal does not mention the word force or gravity as a possible reason for the ball falling. The interview coders were unable to assign a Student Idea Code for Opal for this assessment item. Therefore, no code was assigned because Opal was unable to clearly articulate a description of any force acting on the ball to make it fall. Opal's response during the interview is consistent with her response on the written MFA posttest:

> Opal (written posttest): The tennis ball stayed the same because when you drop the ball it stays in one constant speed.

Opal's ideas about the falling ball were consistent in both the interview and written posttest settings.

Teddy's response during the interview is similar to his response on SCALE-uP's written MFA posttest. He states that the speed of the tennis ball will increase "because as the tennis ball is falling, the mass of the object gained which gravity pulled the object faster making it increase." Later he states that "in order for it to fall faster the mass has to be, it has to weigh more…so that it will go a lot faster." Teddy appeared to associate the speed of the falling object with the object gaining weight as it fell. This is an idea that did not specifically correspond to any of the Student Idea Codes. One coder assigned Student Idea Code 2, students' association of the amount of motion with amount of force. This will be discussed later in this chapter.

The interview responses of Wendy and Scott were similar to their written posttest responses. Each believed that the force of gravity changed as the ball came closer to the ground. Samples of their interview responses were:

> Wendy (interview): I think the speed increases because the gravity is pulling the ball more and more (and later) the force of the gravity…so it's like getting stronger.
>
> Scott (interview): I said gravity caused the speed of the tennis ball to increase because the further the ball got down to the ground the harder gravity pulls it (and later) because as the ball gets closer, further to the ground, the faster the gravity can pull it and the faster it goes.

In each case, the interview coders decided that the student's response was evidence of Student Idea Code 2, or students' association of the amount of motion with amount of force. Both Wendy and Scott believed that as the force of gravity increased, so did the ball's speed. Therefore, by using the coding categories, coders were able to definitively characterize Wendy and Scott's ideas.

Each of the other eight students (Darlene, Lisa, Tom, Barry, Opal, Teddy, Wendy, Scott) was unable to provide consistent evidence of understanding the target idea during both the written MFA posttest and the interview. In these instances, the inferences from both data sources, the written MFA posttest and interview, seem to be consistent.

For this item, there was one case of disagreement between the two evidence sources. Robin's written response was coded as not providing evidence of understanding on the MFA posttest, but her response during the interview was coded as providing evidence of understanding the target idea. Robin's response on the written MFA posttest was, "The force of it droping pushing it downward makes it speed up." SCALE-uP assessment raters categorized Robin's response in the category of "property of the ball" because her response was judged to be consistent with the idea that something about the ball itself that was pushing it downward as opposed to the downward pull of gravity. Robin's interview responses were similar to her response during the written MFA

posttest, although her responses were coded differently. Statements made by Robin during the interview included:

> Robin (interview): Gravity…what I meant by gravity is basically that the ball is basically being pushed down, or like, sometimes people say its being like, almost sticking like sticking down to the floor and it's being pulled down, or it could be a push down. (and later) …the balls' speed increases because it's a push and it is a pull so if you add the push and the pull together it's really, it goes down like fast together.

While Robin appears to have some misconception about the way gravity functions in relation to falling objects (i.e., a push), her response was coded as evidence of understanding the target idea. It was determined by the interview coders that her response represented evidence that she understood the target idea (C_a), "changes in speed or direction are caused by forces." That she may not understand the exact nature of the force (gravity) was not considered relevant to understanding the target idea in this situation.

In conclusion, for this *Dropping and Throwing a Ball* constructed response item, there was one instance of disagreement (Robin) when comparing the inferences from the written MFA posttest with the inferences from the interview. Therefore, there was 7% (1 out of 14) disagreement between the two data sources with this item. This low level of disagreement is acceptable and meets the predetermined acceptable level of disagreement put forth in this study. In addition, for some students (Wendy, Scott, Barry, Darlene) who revealed common misconceptions during the written MFA posttest, the same misconceptions were revealed during the interview.

Carly's Skateboard—Selected Response

This context asks students to consider the motion of a skateboard that has been given a single push across a smooth gymnasium floor and then slows and comes to a

stop. The question asks, “Why does the skateboard come to a stop?” Students are provided with four possible answer choices: there is a force in the opposite direction of the skateboard’s motion, the skateboard runs out of force, the force of the push on the skateboard decreases to zero, and there are no forces on the skateboard. The question is intended to determine if students understand the idea that changes in speed or direction are caused by forces. All 14 students responded to this question on the written MFA posttest but only 13 were able to respond to this question during the interview.

During the written MFA posttest, 6 of the 13 students selected a distracter and seven answered correctly. Opal and Larry selected “because the skateboard runs out of force.” Robin, Ingrid, and Heidi selected “because the force of the push on the skateboard decreases to zero.” Alan selected “because there aren’t any forces on the skateboard.” Each of these students was unable to provide evidence of understanding the target idea during the interview. Larry, after choosing the same response option from the written posttest, responds to probing questions during the interview:

> Ann: So what does it mean when it says that it runs out of force?
>
> Larry: That means that there is no more force being applied to it.
>
> Ann: Okay. So, are there any other answers that you almost chose?
>
> Larry: I almost…well I was thinking about “because there aren’t any forces on the skateboard,” but then it wouldn’t move.

Larry’s responses were coded as consistent with Student Idea Code 5 (forces must act if objects are to stay in motion).

Opal chose a different distracter during the interview by selecting “because the force of the push on the skateboard decreases to zero.” Opal justifies her selection by stating:

> Opal (interview): I chose this answer because when you push the skateboard it doesn't go forever, because there is friction to stop it. So after, um, first it goes fast and then slows down. (later)…I just thought that um, that when she pushes the force of the push on the skateboard would decrease because the friction was slowing it down.

The second portion of Opal's response was coded as evidence of "friction causes a force to run out," or Student Idea Code 8. Opal's selections on this item during the interview and during the written MFA posttest did not match. In neither case did Opal provide evidence of understanding the target idea. This is not considered to be disagreement, for the purposes outlined in this study.

Alan selected the same answer choice during both the written MFA posttest and during the oral interview, "because there aren't any forces on the skateboard." He explained his selection in the following way during the interview:

> Alan (interview): Because, now um, now there is um there isn't any force like um, like um, pushing it or pulling it. And so the skateboard will eventually stop.

Alan's interview response was coded as "forces must act if objects are to stay in motion," Student Idea Code 5. The inferences from both data sources were in agreement for this case.

Heidi selected the same answer choice during both the written MFA posttest and during the interview, "because the force of the push on the skateboard decreases to zero." She explained her selection during the interview in this way:

> Heidi (interview): I think it stopped because like there wasn't anything else pushing her, so then the skateboard like didn't have anything else to like have it pushed. And then it stopped slowly.

Heidi's interview response was coded as "forces must act of objects are to stay in motion," Student Idea Code 5. The inferences from both data sources were in agreement for Heidi's ideas about motion and forces.

Robin also selected the same answer choice during both the written MFA posttest and during the interview, "because the force of the push on the skateboard decreases to zero." She explained her selection during the interview:

> Robin (interview): I selected that one because the…if you think about it, and you're going like on a flat surface, eventually the, whatever is rolling has to come to a stop. So when it says decreases to zero, basically it's saying that it comes to a complete stop.

Robin's interview response was coded as "associating the amount of motion with amount of force," Student Idea Code 2. The inferences from both data sources were in agreement for Robin's ideas about motion and forces.

During the written MFA posttest, Ingrid selected "because the force of the push on the skateboard decreases to zero" and then during the interview she selected "because the skateboard runs out of force." Ingrid explained her selection during the interview:

> Ingrid (interview): I just looked back at the beginning when it said that she pushed off once with her foot, and as she moved it said that she came to a stop. I said that because she only pushed once, then it ran out of force from her push.

This response was coded as Student Idea Code 5, forces must act if objects are to stay in motion. Later in the interview, Ingrid discusses how friction acts to slow rolling objects. When the interviewer asks Ingrid if "there is a force in the opposite direction of the skateboard's motion," Ingrid replies that there is not. This was coded as Student Idea Code 9, friction does not act in a given direction *or* friction is not a force, because Ingrid appeared to understand some aspects of friction. Ingrid either does not recognize that friction is a force or that friction acts in a given direction. Inferences from Ingrid's responses during both the written MFA posttest and the oral interview were in agreement that she was unable to provide evidence of understanding at either time. Ingrid's answer

choices during both assessment situations were quite similar in that "the force running out" is similar in nature to "the force decreasing to zero."

During the written MFA posttest, seven students responded correctly. Six of these students (Barry, Wendy, Teddy, Scott, Janet, Darlene) were also able to provide evidence of understanding the target idea in this context during the interview. Examples of coded evidence statements found in the interview transcripts were:

> Barry (interview): It doesn't really stop until something else stops it. So it wasn't because the skateboard runs out of force it was because of the friction, the force in the opposite direction.
>
> Wendy (interview): Because when we were doing the speed cart, it was slowing down we found it was because of friction. And friction is between the skateboard and the ground and it's like the other direction.
>
> Teddy (interview): The first response was there is a force in the opposite direction of the skateboard's motion. That kind of brings to my attention that that might be friction. So, and friction is the force that stops the object. So since as I read this thing up top it said that it slows down and comes to a stop, and then that made me think of friction.
>
> Scott (interview): I picked that because friction is forcing the skateboard in the opposite direction of which it is going, which makes it to slow down and stop.
>
> Janet (interview): Friction. While the skateboard is going forward friction is pulling it backward.
>
> Darlene (interview): Because if Carly is on the skateboard there is friction with the floor and the, and the wheels of the skateboard. Like, like the two objects are like uh, pressing together and that's making it slow down.

Each of these statements was determined to be evidence of understanding the target idea by the interview coders. In all these cases, the inference from the written MFA posttest matched the inference from the coded evidence in the interview transcripts.

However, Tom's correct response during the written MFA posttest did not match his response during the interview in which he selected the answer choice, "because the

force of the push on the skateboard decreases to zero." The following lengthy exchange occurs between the interviewer and Tom, as he explained why he did not select the correct answer choice (because there is a force in the opposite direction of the skateboard's motion) during the interview:

> Tom: I almost chose the first one.
>
> David: Okay. Why?
>
> Tom: Because I thought that like, friction was the, because I thought friction was another force in the opposite direction of the skateboard but I just knew that friction was stopping it, like, when two objects like to come together, that's friction. And so, since it couldn't come this way, since, friction can't come at you, it comes like slowly, like it slows you down.
>
> David: Why did you think the third one was better?
>
> Tom: Because well, as long as it , because, since there is no opposite force, I just chose the third one because the push, because the push on the skateboard, I mean the push, the push of your leg can't go too far, so I just chose, push, I mean the third one.
>
> David: So, are there any forces in the opposite direction of the way the skateboard is moving?
>
> Tom: (after a pause) No.
>
> David: Can you describe how the force of the push decreasing to zero causes the skateboard to stop?
>
> Tom: Because if there's not more fush…I mean if there's no more push, if there is no more pushing anything, friction is going to eventually slow it down.

Tom understood that friction involves two objects rubbing together and that friction slows rolling objects. However, Tom was unsure about the directional aspect of friction in relation to an object's direction of motion. Tom's response about friction not "coming at you" was therefore determined to be evidence of Student Idea Code 9, friction does not act in a given direction. Thus, the evidence for this assessment item collected

from Tom during both the written MFA posttest and the interview were not in agreement because he responded correctly at time of posttest but did not respond correctly in the oral interview.

Of the 13 students who responded during both the written MFA posttest and during the interview, twelve students' responses were not in disagreement. Therefore, there was 8% disagreement (1 out of 13) between the two data sources for these students in this rolling skateboard assessment context.

A Metal Ball—Selected Response

This item asks students to consider the forces acting on a metal ball that is hanging from a crane. The ball is not moving. The question asks students if the force of gravity on the ball is stronger than the force of the crane on the ball, if the crane applies more force than gravity, if the force of gravity and the force of the crane are the same, or if there are no forces acting on the ball. The question is intended to determine if students understand the idea that "objects at rest will stay at rest unless acted on by a force" (C_c). All 14 students responded to this question on the written MFA posttest but only six (Alan, Wendy, Scott, Robin, Larry, Heidi) were able to respond to this question during the interview.

Of these six students, five of them chose the correct answer on the written MFA posttest and one student (Larry) did not. During both the written MFA posttest and the interview, Larry selected the answer choice, "the crane applies more force on the ball than gravity." Larry explained his answer choice during the interview in the following exchange:

> Larry: If it was the same I think it wouldn't, it would just stay the same and the crane couldn't move down or up because they have the same force. And it couldn't be gravity was stronger because then the ball would just go down.
>
> Ann: Okay, so if the gravity was stronger, you said the ball would just go down. If the forces were the same, what would happen?
>
> Larry: They would just stay the same and the crane wouldn't be able to move up or down.
>
> Ann: Is the crane moving at all? Like is the ball moving at all up or down in that question?
>
> Larry: No, but I think in the controls you could move the crane up and down, and it would have to be stronger in order to do that.

After several follow-up questions, Larry persists in his belief that the force of the crane must be stronger in order for the ball *to be able to* move. It appears that Larry may have misunderstood the question but neither interview coder was able to find transcript evidence consistent with any of the Student Idea Codes. Therefore, no code was applied for Larry's responses to this assessment item.

Of the five students who selected the correct answer on the written MFA posttest, four students (Alan, Wendy, Scott, Robin) also selected the correct answer and provided evidence of understanding in their comments during the interview. During the interview, these four students selected the response, "The force of gravity on the ball and the force of the crane on the ball are the same." Examples of statements made in response to follow-up probes during the interview included:

> Alan (interview): Because the force of gravity…if, if it was more, it would be pulling down, if it was less it will be pulling up.
>
> Wendy (interview): But this one, it's different, because the two, it makes it go up or down if one gets stronger, and so if they're the same then they can't go either way.

> Scott (interview): I selected that because gravity is pulling the ball down. And if it was pulling more than the crane then it would fall to the ground, and the force of the crane is the same as gravity because if it was a bigger force than gravity then it would go up in the air.
>
> Robin (interview): …both of them are about the same because the ball would be moving if they were two different. But we know that the metal ball is not moving, so it had to be number three because they are pushing towards each other which makes the ball stand still.

These statements, combined with the students selecting the correct answer choice, led the interview coders to conclude that the students provided evidence of understanding the target idea, "Objects at rest stay at rest unless acted upon by a force" (C_c).

Another student, Heidi, selected the correct answer on the written MFA posttest, "The force of gravity on the ball and the force of the crane on the ball are the same." However, during the interview Heidi selected the answer choice, "There are no forces acting on the ball." The following exchange then occurred:

> Ann: And what do you think is happening to the metal ball?
>
> Heidi: That it doesn't have anything pushing it so the ball stays, is staying, stays, it's not moving.
>
> (following a few more queries by Ann)
>
> Ann: No, right now while it's hanging there. Can you tell me the forces that are acting on the ball?
>
> Heidi: Nothing. It's just staying.
>
> Ann: Okay. No forces are acting on the ball?
>
> Heidi: Yeah.

Heidi's response was considered to be consistent with Student Idea Code 7 (if a body is not moving, there is no force acting on it) by both interview coders. Therefore, the

inference regarding Heidi's interview response was not consistent with the inference from the written MFA posttest.

Of the six students who responded to this item during the written MFA, evidence from five students' responses was in agreement with evidence collected for those students during the interview. Therefore, there was 17% disagreement (1 out of 6) between the two data sources for these students in this stationary metal ball assessment context.

Pedaling a Bike—Selected Response

This item asks students to consider the motion of a bicycle that moves at a constant speed when a rider pedals the bicycle. The question asks students about the forces on the bike while the rider is pedaling the bike and moving at a constant speed. Students are provided with four answer choices that ask about the relationship between the magnitude of the forward and backward forces acting on the bike. The item is intended to determine if students understand the idea that "objects in motion stay in motion unless acted on by a force." All 14 students responded to this question on the written MFA posttest but only six were interviewed on this item.

Five students (Barry, Tom, Scott, Janet, Ingrid) did not choose the correct answer during the written MFA and one student (Teddy) did. Responses during the interviews varied, and will be discussed individually.

During both the written MFA and during the interview, Barry selected the answer choice, "The total force on the bike in the forward direction is greater than the total force on the bike in the backward direction." Barry explained his answer choice in the following way:

> Barry (interview): I knew that the total force on the bike in the backward direction because it had to be greater to at least move forward, and to keep it in a straight line, to keep it at a constant speed it just has to be greater than the backward direction.

Barry's response was considered by both interview coders to be evidence consistent with Student Idea Code 6, objects in motion have a force in the direction of their motion. Therefore, there was agreement between the two inferences (written posttest and interview) for Barry's responses to this item.

Tom also selected the answer choice, "The total force on the bike in the forward direction is greater than the total force on the bike in the backward direction," during the written MFA posttest. However, during the interview, Tom provided evidence of understanding by selecting the answer choice, "The total force on the bike in the forward direction is equal to the total force on the bike in the backward direction." He explains his answer choice:

> Tom (interview): Since the bike is going to move at a constant speed, it can't be greater than the, uh, since the bike is moving at a constant speed the total force on the bike can't be moving, the total force on the bike in the forward direction can't be greater than the total force on the bike in the backward direction. And since it's moving at a constant speed it can't be less than, so it had to be equal to.

This evidence was considered by the interview coder to be consistent with understanding the target idea in this context. Tom's response during the interview was therefore not in agreement with his previous written response on the written MFA posttest.

During the written MFA posttest, Ingrid selected the correct answer choice, "The total force on the bike in the forward direction is equal to the total force on the bike in the backward direction." However, during the interview, Ingrid selected the answer choice, "The total force on the bike in the forward direction is greater than the total force on the bike in the backward direction." Ingrid appeared to understand the context as she

successfully restated what would happen to the bike's speed when there was pedaling and no pedaling. However, she may have misunderstood the intent of the question and she also appeared to confuse force with speed when she stated:

> Ingrid (interview): I selected it because since the force in the forward direction is greater than total force on the bike in the backward direction, he can, because he can move at a constant speed when he pedals, and it would be a greater speed, but then since it slows down when he doesn't pedal it would be a slower speed.

After several follow-up questions by the interviewer, Ingrid persists in her belief that the forward force is greater than the backward force. Therefore, Ingrid's responses during the interview were coded as being consistent with Student Idea Code 6, objects in motion have a force in the direction of their motion. This inference made during the interview was not in agreement with her previous correct answer choice selection on the MFA posttest.

During the written MFA posttest, Teddy selected the correct answer choice, "the total force on the bike in the forward direction is equal to the total force on the bike in the backward direction." However, during the interview, Teddy selected the answer choice, "the total force on the bike in the forward direction is greater than the total force on the bike in the backward direction." Teddy reported difficulty in understanding the assessment item and explained his answer choice:

> David: Was there anything confusing about it?
>
> Teddy: A little bit with the answers.
>
> David: How do you mean?
>
> Teddy: Um, cause it says, when he doesn't pedal the bike slows down, when he pedals the bike goes at a constant speed in a straight line. And then like I really didn't know what to choose for the bike, because it really didn't seem like any of these fit along the, uh, sentence up top.

> David: Okay. Can you describe how you went about answering the question?
>
> Teddy: Well I went about answering the question by, I thought which one fit into the sentence up top, which was more like it made sense to it, so that's why I picked the first one.

Teddy describes the specific forward force acting on the bike ("how he's pedaling") and the specific backward force acting on the bike ("friction"). He later explains how he eliminated the correct answer choice when the interviewer asks, "What happens when the two forces are equal?" Teddy responds that the bike would "come to a complete stop." Therefore, Teddy's answer choice and responses were considered evidence of Student Idea Code 6, objects in motion have a force in the direction of their motion. The inference from the oral interview is therefore not in agreement with his previous response on the written MFA posttest.

During the written MFA posttest, Scott selected the answer choice, "The total force on the bike in the forward direction is less than the total force on the bike in the backward direction." He selected the same answer choice during the interview. After reading the question aloud, Scott takes 90 seconds to respond and changes his answer twice. He then has difficulty explaining his answer choice:

> Scott (interview): I said this because for the bike to start moving it needs, uh, it needs a f-, a, it needs to be pedaled and it needs a strong force, but if he doesn't pedal then it slows down the bike.

After follow-up questioning, Scott is unable to explain why the forward forces would be less than the backward forces. Therefore, neither interview coder was able to find evidence consistent with any of the Student Idea Codes. Inferences from the two sources of evidence (written posttest and oral interview) were therefore not in disagreement.

During the written MFA posttest, Janet selected the answer choice, “The total force on the bike in the forward direction is greater than the total force on the bike in the backward direction.” During the interview, however, Janet selected the answer choice, “The total force on the bike in the forward direction is less than the total force on the bike in the backward direction.” She explains her answer choice during the interview:

> Janet (interview): I picked that answer because when he stops pedaling it slows down, so then the force that is making him go forward is less than friction because friction is first making it slow down.

Based on this evidence, Janet appeared to misunderstand the question by thinking she was being asked about the forces when the bike is slowing down. After some clarification by the interviewer, Janet changes her mind and states that she would have chosen “the third one” which states that the forces are equal when the bike moves at a constant speed in a straight line. This occurred in the following exchange:

> Ann: Okay good. Now, it does say here, the bike is moving at a constant speed in a straight line. What does that mean? Constant speed in a straight line?
> Janet: It's going at the same distance in the same time.
>
> Ann: And so because he's going at a constant speed. What do you think that means about the forces?
>
> Janet: I think they're equal.
>
> Ann: So which one of those responses kind of shows that they're equal?
>
> Janet: The third one.

Therefore, Janet’s response to this item could be definitively coded because there was enough evidence to suggest that she understood the target idea. Thus, the inferences from the written posttest and the interview are in disagreement as Janet was able to demonstrate understanding during the interview but not during the written MFA posttest.

Of the six students that responded to this item during the written MFA posttest, evidence from two students' responses (Barry and Scott) was not in disagreement with evidence collected during the interview. However, evidence from the other four students (Tom, Teddy, Ingrid, and Janet) was in disagreement. Therefore, there was 67% disagreement (four out of six) between the two data sources for these students in this bicycle assessment context. This level of disagreement exceeds the level of disagreement deemed acceptable for this study.

In conclusion, for Research Question 1, Table 9 shows the results of the transcript coding and data analysis procedure. Several of the items showed relatively low levels of disagreement between the two evidence sources while one of the items showed a high level of disagreement. This will be discussed in more depth in Chapter V.

Research Question 2

The next section of Chapter IV will present the data collected to address the following research question:

> During both a written posttest and a videotaped oral interview, to what extent are middle school students in this study able to consistently demonstrate their understanding of the same forces and motion ideas targeted by a variety of assessment task contexts?

To address this research question, a holistic judgment method was used with the videotaped interviews. The two assessment interviewers viewed video recordings of entire interviews and holistically arrived at a judgment about the level of a given student's understanding across all the item contexts. After watching a videotaped

interview, each interviewer holistically judged the evidence in the video about each student's understanding of the forces and motion ideas and categorized it into one of three categories (student mostly understands, student somewhat understands, or student does not understand). Each interviewer watched and judged seven interviews. Each interview was coded one time. The interviewers did not watch and judge students they had interviewed.

This holistic judgment made by the interviewers is compared to the performance levels as developed by SCALE-uP as part of a standard setting procedure. As discussed in Chapter III, SCALE-uP used an analytic judgment method to arrive at four levels of understanding for the MFA (Pyke & Hanson, 2005):

1. No evidence of understanding (0-20).
2. Some evidence of understanding (20-50).
3. Some fluency of understanding (50-70).
4. Flexible evidence of understanding (70-100).

Each student in the SCALE-uP study (including the students in this study) received a score of 0-100 that corresponded to one of the above levels of understanding of forces and motion ideas across all the items on the written MFA instrument. These levels were used as an indicator of their understanding of and commitment to motion and forces ideas while responding to the assessment items on the MFA instrument.

In this study, it was hypothesized that students characterized as high performers according to SCALE-uP's written MFA instrument (i.e., flexible evidence of understanding) would also be characterized as high performers by the holistic judgments made by the interviewers. Conversely, the researcher expected that students characterized

as low performers according to the written MFA instrument (i.e., no evidence of understanding) would also be characterized as low performers by the holistic judgments made by the interviewers. The results from these holistic judgments are presented in Table 10.

As shown in Table 10, four students were characterized on both scales as understanding the motion and forces ideas (Teddy, Scott, Janet, Barry) and three were characterized with low levels of understanding (Lisa, Heidi, Robin). By closely examining the above judgments and reviewing the evidence collected in both the written and oral interview settings, it was possible to gain a better understanding of each student's individual performance and how consistently they were able to apply their ideas about motion and forces across all the assessment contexts during the interview and written posttest.

Heidi was holistically judged as *student does not understand* during the interview and scored in the second level on the written MFA posttest (*Level II: Some evidence of understanding*). The interviewer commented that it was difficult to hear Heidi and it appeared that she may have been learning English. Heidi was able to respond to six of the seven assessment items, and the interview lasted 30 minutes (average interview was approximately 24 minutes). During the interview, she was only able to demonstrate evidence of understanding the first forces and motion idea (changes in speed or direction are caused by forces) in the *Dropping and Throwing a Ball* assessment context. She was unable to provide evidence of understanding the same idea in the rolling ball or skateboard contexts when forces are acting to slow the objects' motion. While she provided some written evidence at posttest of understanding in the metal ball context, she

Table 10. *Interviewed Students' Level of Understanding as Judged by Two Methods.*

Student	Student Understanding of Forces and Motion Ideas			
	SCALE-uP's MFA Level of Understanding (1, 2, 3, or 4)		Interviewer's Holistic Judgment (1, 2, or 3)	
Heidi	Some evidence of understanding	2	Student does not understand	1
Lisa	Some evidence of understanding		Student does not understand	
Robin	Some evidence of understanding		Student does not understand	
Opal	Some evidence of understanding		Student somewhat understands	2
Wendy	Some fluency of understanding	3	Student somewhat understands	
Tom	Some fluency of understanding		Student somewhat understands	
Alan	Some fluency of understanding		Student somewhat understands	
Darlene	Flexible evidence of understanding	4	Student somewhat understands	
Larry	Flexible evidence of understanding		Student somewhat understands	
Ingrid	Flexible evidence of understanding		Student somewhat understands	
Barry	Flexible evidence of understanding		Student mostly understands	3
Teddy	Flexible evidence of understanding		Student mostly understands	
Scott	Flexible evidence of understanding		Student mostly understands	
Janet	Flexible evidence of understanding		Student mostly understands	

*A polychoric correlation (test for ordinal data) between the Interviewer's Holistic Judgment and the MFA Level of Understanding was statistically significant (0.70).

was unable to demonstrate this understanding during the interview when she concluded that there was no force acting on a metal ball that is not moving. Overall judgments about Heidi's understanding of motion and forces seem to be consistent across all the evidence.

Lisa was also holistically judged as *student does not understand* during the interview and she scored in the second level on the written MFA posttest (*Level II: Some evidence of understanding*). Lisa's interview was the shortest of all the interviews (16 minutes) and she only responded to four items. She spoke very softly during the interview and was frequently difficult to hear. Lisa was unable to demonstrate understanding on any of the assessment items during the interview and her understanding of motion and forces ideas during the written MFA posttest is demonstrated on other items not included in the interview protocol. As discussed in the previous section, Lisa misunderstood one of the assessment contexts and this lack of understanding the context could have limited her performance during both the written MFA posttest and the interview.

Robin was judged as *student does not understand* during the interview and scored in the second level on the written MFA posttest (*Level II. Some evidence of understanding*). Robin seemed comfortable in the interview setting and spoke confidently. She responded to six of the seven assessment items and her interview lasted 27 minutes. During the interview, Robin was unable to demonstrate understanding of "objects in motion stay in motion" (C_b) in the rolling ball context and "changes in speed are caused by forces" in the skateboard context. She was able to demonstrate understanding of "changes in speed are caused by forces" in the *Dropping and Throwing a Ball* context. She also provided evidence of understanding "objects at rest stay at rest" in the metal ball context. The holistic judgment about Robin's level of understanding seems somewhat inconsistent with her performance on the written MFA posttest.

However, the interviewer commented that Robin almost scored higher on the holistic judgment and considered the option of assigning her as *student somewhat understands*.

Opal was also judged as *student somewhat understands* during the interview and she scored in the second level on the written MFA posttest (*Level II. Some evidence of understanding*). The interviewer commented on the difficulty hearing Opal because she spoke softly. Opal responded to five of the seven assessment items during the interview which lasted 27 minutes. Opal did not recognize that a force (gravity) acting on a falling ball would change its motion (i.e., cause it to increase its speed). She provided some evidence of understanding "objects in motion staying in motion" when she said that a ball kicked toward her could have gone on forever if something had not stopped it. She also demonstrated evidence of understanding that rolling objects (i.e., a ball and skateboard) slow down because of friction. However, she did not provide evidence that she understood friction was a force that could act in a given direction. In conclusion, Opal's demonstrated understanding on the written MFA posttest was in agreement with her demonstrated understanding during the interview.

Wendy was judged as *student somewhat understands* during the interview and she scored in the third level on the written MFA posttest (*Level III: Some fluency of understanding*). She was able to respond to six of the seven assessment items. She was unable to demonstrate evidence of understanding in the rolling ball assessment context because she believed that a force must act to keep objects in motion while they are moving. However, she demonstrates an understanding of an object slowing down being attributed to a force (i.e., friction) in the opposite direction of motion. She also demonstrates an understanding of forces being balanced and resulting in at-rest position.

In conclusion, Wendy has some understanding of some motion and forces ideas across the assessment contexts. This was demonstrated during both the written MFA posttest and the interview.

Tom was judged as *student somewhat understands* during the interview and he scored in the third level on the written MFA posttest (*Level III: Some fluency of understanding*). Tom spoke confidently during the interview, seemed comfortable in the setting, and responded to six of the seven assessment items. He was able to demonstrate his understanding of forces and motion during the *Rolling Ball* context in that he consistently concluded that forces do not need to act in the forward direction for objects to be in motion. However, he consistently applied the idea that constant force results in constant velocity in the *Dropping and Throwing a Ball* context. Interestingly, two of Tom's responses to selected response items (*Carly's Skateboard* and *Pedaling a Bike*) were not consistent during the written MFA posttest and the interview. While recognizing friction as a force that slows objects motion, Tom did not appear to understand that friction could have a directional component while responding in the skateboard context. However, this was not the case in the bicycle context where Tom demonstrated evidence of understanding forward and backward forces acting on a bicycle in motion. Tom's varied use of the motion and forces ideas across the assessment contexts was in agreement with the holistic judgment from the interview data and with his score on the written MFA posttest.

Alan was judged as *student somewhat understands* during the interview and he scored in the third level on the written MFA posttest (*Level III: Some fluency of understanding*). Alan spoke quietly during the interview, seemed nervous, and stuttered

throughout the interview. He responded to six of the seven assessment items during the interview. Alan was unable at any time to demonstrate an understanding of "objects in motion staying motion" during the rolling ball assessment context. In this context, he consistently used the idea that forces must act if objects are to stay in motion. He also used the same idea in the skateboard context to describe how forces must act to keep the skateboard in motion. Alan did demonstrate understanding of forces changing the speed of objects in the dropping ball context. Alan also demonstrated his understanding of "objects staying at rest" in the metal ball assessment context. Alan's use of correct and incorrect motion and forces ideas was observed in both the interview and on the written MFA posttest. Alan's demonstrated understanding of motion and forces ideas seems well characterized by both the holistic judgment and his written MFA posttest score.

Darlene was judged as *student somewhat understands* during the interview but she scored in the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the written MFA posttest. Darlene was able to respond to five of the seven assessment items during the interview. Darlene provided some evidence of understanding on the selected response items in the falling and rolling ball contexts but, when asked to explain why the ball's speed increases as it falls, is unable to attribute that increase in speed to a force. She was, however, able to demonstrate understanding of "a change in speed caused by a force" during the skateboard context when she described friction as a force acting in the opposite direction of the skateboard's motion. Darlene's use of the motion and forces ideas seemed inconsistent and her score of 74 on the written posttest may represent an overestimation of her understanding of the force and motion ideas.

Larry was judged as *student somewhat understands* during the interview but he scored in the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the MFA posttest. He responded to six of the seven assessment items during the interview that lasted 27 minutes. Larry seemed somewhat nervous at the beginning of the interview and often sought confirmation for his answers. He deliberated at length when asked to describe what he meant by "force" and eventually did not provide an answer. He may have been concerned about being wrong during the interview. When pressed later, he described force as "strength." On at least one occasion, he changed his answer in response to probing questions by the interviewer. For example, after describing how a force does not constantly push a rolling ball, he later changed his response during the interview. He then applied the same idea during the skateboard context when he described how a force must act if objects are to stay in motion. Larry seemed to apply his ideas about forces and motion inconsistently during the interview and his high score of 84 on the written posttest may represent an overestimation of his understanding. Larry may have felt more comfortable during the written test format when he was not required to explain or defend his answers. The follow-up questions seemed to prompt Larry to doubt his own answers about the contexts.

Ingrid was judged as *student somewhat understands* during the interview but she scored in the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the written MFA posttest. Ingrid spoke confidently during the interview and seemed sure of her responses. During the interview, Ingrid reveals some clear misconceptions about the nature of gravity that were intentionally not coded in the interview transcripts. Ingrid also revealed a few misconceptions about the nature of

forces in the rolling skateboard and bicycle contexts. Ingrid believed that forces must act to keep objects in motion in the skateboard context and that the bicycle moving forward must have a force in the direction of its motion. However, neither of these assessment contexts was included in the calculation of the MFA posttest score. It appears that the items and contexts included in the interview revealed a few of Ingrid's ideas about forces and motion that were not captured by the written MFA posttest. In conclusion, Ingrid's use of the force and motion ideas across the contexts was inconsistent and perhaps more accurately captured during the interview than during the written MFA posttest. It is likely that the MFA instrument was an overestimation of Ingrid's understanding of motion and forces.

Barry was judged as *student mostly understands* during the interview and he scored in the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the written MFA posttest. Barry spoke quietly but confidently during the interview. Barry reveals his misconception that constant force (gravity) results in constant velocity in the falling ball context and did not understand that a force applied results in acceleration. He also revealed a misconception in the bicycle context that objects moving at a constant speed have a net force in the direction of their motion. He did appear to understand that a force (friction) causes a rolling object (skateboard and bicycle) to slow down. However, although the holistic judgment of *student mostly understands* does appear to be consistent with his performance on the written posttest, the coded transcript data does not seem consistent with his item-by-item performance during the interview.

Teddy was judged as *student mostly understands* during the interview and he scored just within the lower limit of the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the MFA posttest. Teddy seemed comfortable in the interview setting and was careful but confident in his responses. On a few occasions, he confessed "I don't know" to difficult probing questions. He reveals his idea that objects need a force to keep them moving during the rolling ball context. However, he does not use this idea while responding in the skateboard context. Instead, he used his understanding of friction (as a force in the opposite direction) to explain why the skateboard comes to a stop. In the dropping ball context, he provided evidence of understanding that gravity is a force that acts on falling objects but could not provide evidence of understanding of how gravity caused the motion of the ball to change. The holistic judgment about Teddy's understanding during the interview (*student mostly understands*) does appear to be consistent with his performance on the written posttest. However, this does not seem to be completely consistent with the coded transcript data and his item-by-item performance during the interview. The interviewer confessed that she would have assigned a three minus (3-) if this were an option. This indicated her recognition that Teddy's understanding of the ideas was perhaps between the second and third levels of the holistic judgment categories.

Scott was judged as *student mostly understands* during the interview and he also scored just within the lower limit of the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the written MFA posttest. Scott wrote and spoke slowly and his speech was often fragmented but understandable. He also erased and changed his answers on a few occasions. Scott responded to all seven assessment items and the

interview lasted 21 minutes. He frequently referenced his experiences with the class experiments to describe and justify his answers. He demonstrated his understanding of a "change in speed being caused by a force" during the skateboard and dropping ball contexts. However, he shows some evidence of a misconception during the dropping ball context as he associated the amount of motion with the amount of force. In the rolling ball context, Scott's responses during the interview revealed more evidence of understanding than during the written posttest. He was able to demonstrate an understanding of "objects at rest staying at rest" in the metal ball context as he recognized that active and passive forces involved resulted in no motion. The holistic judgment of Scott's understanding during the interview (*student mostly understands*) does appear to be consistent with his performance on the written posttest. However, the interviewer confessed that "it was close" and almost judged Scott to be *student somewhat understands*, which suggested that Scott's understanding of the ideas was perhaps between the second and third levels of the holistic judgment categories.

Janet was judged as *student mostly understands* during the interview and she scored just within the upper limit of the highest level of understanding (*Level IV: Flexible evidence of understanding*) on the written MFA posttest. Janet appeared somewhat timid during interview but she quietly and confidently responded to the interviewer's questions. Janet consistently demonstrated her understanding of the target ideas across each of the assessment contexts. The exception was the bicycle context in which she provides conflicting evidence by first concluding that the forward force is less than the backward force and then, after a few follow-up questions, changing her response to the correct response. The holistic judgment about Janet's understanding during the interview

(*student mostly understands*) is consistent with her performance on the written MFA posttest.

In many cases, the judgments made by the written MFA posttest and the holistic judgments made by the interviewer are in agreement. There are a few cases in which the MFA posttest score seems to represent an overestimation of a student's understanding (i.e., Darlene, Larry, Ingrid). This could be due to the difference in items used to calculate the MFA posttest score and those items used to collect data during the interview (i.e., the metal ball, skateboard, and bicycle items were not used to compute the MFA posttest score). In addition, it is likely that the difference in the items used to judge students' understanding of motion and forces in the two settings contributes to a mismatch in the conclusions drawn. These issues will be discussed in more detail in Chapter V.

Research Question 3

The following section of Chapter IV will present the data to address the third research question:

> Based on the elicited ideas of sixth graders in this study, what can this research add to the literature base (as presented in the study) about common ideas students have regarding forces and motion?

It was proposed that a presentation of new "Student Idea Codes" added during the coding of data would represent instances of potential contribution to the research literature by this study. Based on the specific ideas of the students in this study and the Student Idea Codes that were added, the idea that "friction causes a force to run out" when considering the motion of slowing and stopping objects was considered to be a student idea that was

previously unreported in the research literature. Evidence consistent with this idea was observed during the interview with Opal who states:

> Opal (interview): So I was thinking that if she like pushes her skateboard with her feet, um, I just thought that um, that when she pushes the force of the push on the skateboard would decrease because the friction was slowing it down.

Alan states during the interview that force on the skateboard will be used up because of friction in the following statement:

> Alan (interview): Because um, as you push it, the force becomes, like um like um it like um like, it gets like um, used up, because of friction.

In addition, a Student Idea Code was added to encompass a few students' unique belief that "friction does not act in a given direction" or that "friction is not a force." Evidence consistent with the idea that "friction does not act in a given direction" was observed on a few occasions. First, Tom states while explaining his response to the skateboard item:

> Tom (interview): …but I just knew that friction was stopping it, like, when two objects like to come together, that's friction. And so, since it couldn't come at you, it comes like slowly like it slows you down.

Tom later states that "there is no opposite force" on the skateboard's motion and when asked by the interviewer if there are any forces in the opposite direction of the way the skateboard is moving, Tom states simply, "no." Tom believes that friction does not act in a specific direction or that friction is not a force. He could hold both ideas as well.

Ingrid, when asked "if there is a force in the opposite direction of the skateboard's motion," replies "no." However, just prior to this interviewer probe, Ingrid mentioned friction as a cause for rolling objects slowing down and stopping. Clearly, she has learned about friction and is able to talk about friction in appropriate ways but is unable to reconcile this notion with her belief that rolling objects run out of force, which is her

response to the skateboard assessment item. While a few of the students in this study held these beliefs, the interview coders decided that these ideas were prevalent enough to code them separately in the transcripts.

A few other ideas emerged from the transcript analysis of the falling ball context. One student, Teddy, asserted that the mass of a falling object increases as it falls. He stated that the speed will increase because "as the tennis ball is falling, the mass of the object gained, which gravity pulled the object faster making it increase." Another idea emerged during the interview with Darlene, who revealed her belief that smaller objects (i.e., an egg) fall faster than heavier objects (i.e., a watermelon) because smaller objects have less weight. She stated this idea during the following exchange:

> Darlene: Like if you put a watermelon here and an egg here, the egg will probably go down faster than the watermelon.
>
> David: Why?
>
> Darlene: Because the watermelon has more weight than the egg.
>
> David: Okay. So... if it's heavier then what does that mean?
>
> Darlene: That if like, it falls down slower than the egg, because ….the egg has less weight.

These student ideas will be discussed in more detail in Chapter V. The next chapter will also include a discussion of how assessment of motion and forces ideas with middle school students can be improved. Specifically, discussion in Chapter V will focus on how the assessment contexts used in this study either allowed or did not allow students to demonstrate evidence of their understanding of the target ideas.

Summary

In conclusion, many of the assessment items showed relatively high levels of agreement between the inferences from the two evidence sources while one of the assessment items showed a low level of agreement. When considering students' overall understanding as determined by the written MFA posttest and by the interviewer, the judgments seemed to be in agreement. On a few occasions, the written MFA posttest seems to represent an overestimation of a student's understanding. As a result of transcript coding, several unique student ideas emerged from the data which will be discussed in more detail in Chapter V.

CHAPTER V: CONCLUSIONS AND RECOMMENDATIONS

This chapter presents the interpretations and conclusions for this study. The chapter includes a review of the study's purpose and significance, a restatement of the research questions, summary of the findings, discussion, implications, conclusions and recommendations, as well as a critique of the study.

Purpose and Significance of the Study

Educational assessment, broadly defined, is about reasoning from what students say, do, or make in specific situations to more general inferences about what they know, have accomplished, or are apt to do in the future (Mislevy, Steinberg, & Almond, 2003). In many assessment situations, the amount of evidence provided by students from which to make inferences about their performance is quite limited (e.g., the selection of an answer choice). Therefore, the quality of an assessment item and the data it provides are fundamental to the validity of an inference about student knowledge. Addressing this crucial area of educational assessment, this study sought to confirm the inferences provided by a written science assessment instrument by comparing them to alternative inferences based on student interview data. As depicted in the conceptual framework (Figure 2), this study aimed to determine how the inferences about student knowledge determined by a written assessment format compared to the inferences about student knowledge as determined via interviews with students.

Specifically, this study explored the extent to which paper and pencil assessment items, aligned with specific middle school level forces and motion ideas, accurately measure those ideas when compared with student interviews about the same ideas. The study used a qualitative coding methodology to assess students' understanding from a

videotaped interview and to compare that understanding to inferences made about students' understanding from a paper and pencil assessment instrument. This study makes an important contribution to the science education literature with insights about how aligned middle school science assessment tasks can ascertain sixth-grade students' understanding of forces and motion ideas. The study provides a framework for conducting assessment validation research that considers the gathering of evidence about students' ideas in relation to their responses to assessment tasks. The study also provides valuable insights into the design of assessment tasks and students' responses to them.

Summary and Interpretations of Findings

For most assessment items, the data suggest that the items provide high-quality, valid inferences of middle school students understanding of specific physical science ideas. When considering student performance across a set of items for test level inferences, this study concludes that an aggregate score on the written MFA posttest provides a reasonable estimate of student understanding of motion and forces when compared to holistic judgments made from in-depth interview data. On a few occasions, the written MFA posttest overestimated overall student understanding of the ideas compared to the interview data. This study also presents several previously undocumented student ideas that emerged from transcript coding.

Research Question #1

This study concludes that there were relatively high levels of item-level agreement between the inferences made based on the written MFA posttest and those made from coding the interviews. On 6 of the 7 items under investigation in this study, student responses on the written MFA were confirmed during the interview 83% to 93%

of the time. For one of the items, there was a low level of agreement (33%) between the two inferences. The following sections of this chapter will present and discuss the levels of agreement for each item. This section also includes discussion of the instances of disagreement, where appropriate. The three causes for disagreements were comprehensibility, guessing, and unstable student ideas across test settings. Table 11 shows the item-by-item inferences (correct and incorrect understanding) for each student in the study along with summary level agreement for each item and for each individual student.

For the rolling ball item, both the selected response and constructed response items had two instances of disagreement out of the 14 possible (Table 11). For the selected response item, with only two answer choices, this is encouraging evidence for the effectiveness of the item. Common assessment development recommendations encourage at least four answer choices (Crocker & Algina, 1986) because there is a strong likelihood of successful guessing by students unsure about how to respond, i.e., when they don't know the idea being tested. One way to lessen the guessability problem is to include plausible answer choices based on commonly held misconceptions so students have an answer choice that actually reflects what they think (DeBoer, 2005). This may explain why this two answer choice item did not have more false positive responses, i.e., the distracter choice ("Yes") is a strong misconception so students have a plausible answer choice to select and don't have to guess. This was, in fact, confirmed in this study in which 5 of the 6 students selecting the distracter on the written MFA posttest provided similar evidence of having that misconception during the interview (i.e., a force acts on a ball while it rolls in a given direction). In addition, 7 of the 8 students who

Table 11. *Judgments Made for Students During SCALE-uP's Written MFA Posttest and Oral Interview.*

Student	Rolling Ball (SR)		Rolling Ball (CR)		Falling Ball (SR)		Falling Ball (CR)		Skateboard (SR)		A Metal Ball (SR)		Bicycle (SR)		Level of Understanding	
	Write	Oral	Write	Oral	Write	Oral	Write	Oral	Write	Oral	Write	Oral	Write	Oral	MFA (SCALE-uP) (1-4)	Interview (1-3)
Alan	IC	IC	IC	IC	C	C	C	C	IC	IC	C	C	-	-	3	2
Barry	C	C	C	IC	IC	IC	IC	IC	C	C	-`	-	IC	IC	4	3
Darlene	C	C	IC	IC	C	C	IC	IC	C	C	-	-	-	-	4	2
Heidi	IC	IC	IC	IC	C	C	C	C	IC	IC	C	IC	-	-	2	1
Ingrid	C	C	C	C	C	C	C	C	IC	IC	-	-	C	IC	4	2
Janet	C	C	C	C	C	C	C	C	C	C	-	-	IC	C	4	3
Larry	C	C	IC	IC	C	C	C	C	IC	IC	IC	IC	-	-	4	2
Lisa	IC	IC	IC	IC	IC	IC	IC	IC	-	-	-	-	-	-	2	1
Opal	C	C	IC	IC	IC	IC	IC	IC	IC	IC	-	-	-	-	2	2
Robin	C	IC	IC	IC	C	C	IC	C	IC	IC	C	C	-	-	2	1
Scott	IC	C	IC	C	C	C	IC	IC	C	C	C	C	IC	IC	4	3
Teddy	IC	IC	IC	IC	C	C	IC	IC	C	C	-	-	C	IC	4	3
Tom	C	C	C	C	IC	IC	IC	IC	C	IC	-	-	IC	C	3	2
Wendy	IC	IC	IC	IC	C	C	IC	IC	C	C	C	C	-	-	3	2
Agree-ment	86%		86%		100%		93%		92%		83%		33%			

*Shaded cells denote instances of disagreement between evidence sources (written MFA and oral interview). C = Correct understanding; IC = Incorrect understanding. SCALE-uP MFA Levels of Understanding (1=No Understanding; 2=Some Understanding; 3=Some Fluency with Ideas; 4=Flexible Understanding). Interview Levels of Understanding (1=Does Not Understand; 2=Somewhat Understands; 3=Mostly Understands).

responded correctly on the written MFA posttest were able to provide evidence during the interview to confirm their understanding of the motion and forces idea. However, in the two instances of disagreement, one student may have guessed to arrive at the correct response on the written posttest, because she was unable to demonstrate understanding of motion and forces for this item at any point during the interview. The other student chose the incorrect answer choice on the written MFA posttest but was able to demonstrate understanding during the interview. A possible explanation for the disagreement in this student's two responses was his insecure knowledge across testing contexts, as discussed in Chapter IV. Overall, there was an 86% level of agreement between the written MFA and interview results. This suggests the item provides high quality inferences about student understanding of the targeted idea.

There were similar results for the constructed response portion of the rolling ball context. Of the 14 students who responded to this item, there were two instances of disagreement between the inferences provided by the written MFA posttest and the inference provided by the coded interview data. One student, who responded incorrectly during the written MFA posttest, was able to demonstrate understanding during the videotaped interview. While the two responses were similar, this dissertation came to a different conclusion regarding the student's understanding when compared to SCALE-uP's MFA judgment, which was based, in part, on the student's incorrect response to the first item in this context. Another student, who answered correctly on the written MFA, seemed to provide some initial evidence of understanding during the interview (including the idea that there is not a force constantly pushing the ball as it rolls), but upon further questioning by the interviewer revealed naïve thinking about the nature of forces (i.e.,

forces running out and wearing off). This student appeared to hold multiple ideas about the nature of the forces acting on a rolling ball. The incorrect ideas were elicited during the interview as a result of several probing questions by the interviewer but were not elicited by the written MFA. Besides providing an opportunity to confirm what students do and do not know, probing during an interview also has the potential to uncover other misconceptions. With only two instances of disagreement out of 14 students, the written item provides valid inferences about student knowledge 86% of the time.

In the falling ball test item context, there were also relatively high levels of agreement between the two sets of inferences. For the selected response portion of the item, there was complete agreement between the two sources of evidence across all students and indicated that the item provides valid inferences about student understanding. The ten students who selected the correct answer on the written MFA also provided evidence of understanding during the interview, and the four students who selected incorrect answers on the written MFA chose the same distracters during the interview to confirm their misconception. On the constructed response item, there was one instance of disagreement when a student responded incorrectly on the written MFA but was able to respond correctly during the interview. The student responses in both settings were somewhat similar but the interview allowed the student to elaborate in more detail and reveal her correct understanding, whereas on the written MFA her more limited response was rated by SCALE-uP as evidence of a misconception. However, this student did not express a strong understanding in either setting, which could represent a limitation in the coding procedures for responses that show weak understanding. Therefore, the interview provided more accurate information about her understanding

compared to the written MFA. Overall though, the 100% and 93% agreement between the two sets of inferences about student ideas suggest that these written items can be used to make valid inferences about student understanding of the specific forces and motion idea targeted.

In the rolling skateboard assessment item context, the answers of 12 of 13 students on the written test were confirmed during the interview. In the one instance of disagreement, a student responded correctly at the time of the written MFA by selecting the correct answer choice (there is a force in the opposite direction of the skateboard's motion). This student was then unable to provide evidence of understanding during the interview when he selected an incorrect answer choice (because the force of the push on the skateboard decreases to zero). He stated that he almost chose the correct answer choice but then stated, "I just knew friction was stopping it...and since it couldn't come this way, since, friction can't come at you." It is clear that the student was unclear about the directional aspect of the force of friction. He understood that friction was a cause of the object slowing down but did not have an understanding of how friction acted in the opposite direction of an object's motion. Since the correct answer choice does not include the term friction, students needed a more advanced understanding that friction is the force acting in the opposite direction of the skateboard's motion. His response on the written MFA posttest seemed to suggest that he understood this idea but upon further probing during the interview, it became clear that he did not understand the target idea in this context of forces causing objects to slow down. Based on this one instance of disagreement, it is difficult to recommend a revision to this item to include the term "friction" in the correct answer choice (e.g., friction is acting in the opposite direction of

the skateboard's motion). Such a change may result in more false positive responses as students may be predisposed to select the answer choice with the more technical vocabulary. In addition, such an addition may change the item to target the idea of friction, which is not the intent of the item. Overall for this particular item, the level of agreement between the sets of inferences was high (92%) and the written item seemed to be functioning well. The item appears to provide solid evidence from which to draw valid inferences about student understanding of the target idea.

In the hanging metal ball assessment context, there was also a high level of agreement between the sets of inferences. In only one instance (out of a total of six) was there a disagreement between inferences about student understanding. A student responded correctly on the written MFA that the forces acting on the hanging ball were equal but was then unable to demonstrate evidence of understanding of the target idea during the interview by saying that "it doesn't have anything pushing it so the ball stays...it's not moving." During two direct follow-up questions, she confirmed her belief that there were no forces acting on the ball. This is in contrast to her answer choice selection that stated that the forces acting on the metal ball were equal. It appears in this context that she may have been struggling to apply her ideas about equal and balanced forces consistently (i.e., in static, balanced situations are there no forces or equal forces?) in novel contexts after spending several weeks learning about the forces and motion ideas.

For the five students who chose the correct answer on the written MFA, four of them confirmed that they understood the idea being tested during the interview. For the one student who chose the incorrect answer, this student's misconception was confirmed

during the interview. This represents a small amount of disagreement between the inferences for this item. The item appears to provide valid inferences about students' understanding of the balanced nature of forces acting on a resting object. For students who indicated that there are no forces acting on the ball, the extent to which they may have an understanding of net forces (i.e., "no net force" being similar to "no forces" or "forces are equal") was unclear. Given the relative unpopularity of the "no forces" answer choice for the students in this study, more research is needed with students to explore this issue.

The final assessment context involves a bicycle and the forces acting on it while it is pedaled and moving at a constant speed and in a straight line. Six students responded to this item during the written MFA posttest and during the interview, but four of the six sets of inferences about student ideas were in disagreement. This occurred in two ways. Two students chose the correct answer during the written MFA but were then unable to demonstrate correct understanding during the interview. In one of these instances, a student openly confessed during the interview that he didn't understand the item and so he picked the answer choice that "fit into the sentence up top." Given this student's admitted guessing during the interview and a confessed lack of understanding of the item, it is likely that successful guessing occurred during the written MFA as well. The other student appeared to be insecure in her knowledge of the ideas because after responding correctly during the written MFA, she successfully summarized the item context but then provided a detailed and confident rationale for selecting the incorrect answer choice related to greater forward forces. Another two students who did not respond correctly during the written MFA were then able to demonstrate correct understanding during the

interview (false negatives). One student was confused by the item and context but upon clarification during the oral interview, she was able to provide clear evidence of correct understanding. The evidence from this student's interview suggests that the item has poor comprehensibility, which may inhibit a successful response. Another student provided an exceptionally concise explanation for why the forces on a bike moving at a constant speed "had to be equal" during the interview. This conflicted with his written MFA response and would seem to suggest his knowledge was not secure at time of posttest. Again, the interview was able to provide a more complete picture of this student's understanding on this item.

In light of the student data for this assessment item, it appears that the context may be complicated and cognitively taxing for students. As students process the item information, they are required to read and understand two situations (the bike's speed while pedaled and the bike's speed while not pedaled). The students are then asked about the forces acting on the bike while pedaled. It is possible that the task places a load on working memory such that information processing is constrained while attempting to produce a response. A revision to the item might include restricting the context to the bike's speed while pedaled. Given that only 33% of the responses to the written MFA were confirmed during the interview, this item (in its current form) should not be included in future versions of the MFA.

Given the well-documented misconception that there is a net force acting in the direction of an object's motion (Osborne and Freyberg, 1985; Palmer and Flanagan, 1997; Project 2061, 2006; Watts and Zylbersztajn, 1981), it is not surprising that only two of the six students knew, when probed during an interview, that the forces acting on a

constantly moving object are actually balanced. On the written MFA, 3 of the 6 students in this study thought that the force in the forward direction was greater, and this is consistent with previous findings in the literature.

This task also presents students with a familiar context such as riding a bicycle. In assessment development, this would seem to be a positive aspect of the item because the context is accessible and is not likely to advantage any group of students. However, this context could be confusing given the variety of forces involved in maintaining a constant speed while riding a bike. Although impossible to represent perfectly, there may be other contexts that present an object moving at a constant speed that more simply represent the phenomena of an object moving while the forces are balanced (e.g., ice or air hockey or marbles on smooth surfaces).

In summary, six of the items in this study produced high levels of agreement between the inferences from the written MFA posttest and the inferences from the coded interview data and one item produced low levels of agreement between inferences. This speaks well of the quality of the written items and the criteria that were used to ensure content alignment and construct validity during item development. But, the results also show that applying these criteria is not enough by itself. Follow-up in which the thinking of students can be confirmed is important as well. The findings also lend support for the use of commonly held ideas as distracters in item development as students' ideas were commonly confirmed during the interviews in this study. In addition, this practice may also be related to the relatively low incidence of false positive and false negative responses. Although, such connections would require more systematic exploration and go beyond the scope of this study.

Finally, there are commonalities across the different types of disagreement that occurred between evidence sources for the items. In several instances (especially in the bicycle context), students appeared to not understand the context, which suggests comprehensibility issues with the item. While the item review criteria provided an effective screen for issues of comprehensibility, it is clear that qualitative data collected during an interview study provide more student-centered information regarding the comprehensibility of assessment items. Based on these instances of disagreement, suggestions have been made for ways to improve the items. In other instances, disagreements were attributed to the insecure state of student ideas. Some level of this kind of error in assessment data is unavoidable. Error of this type is perhaps exacerbated in a study involving the measurement of 6th graders understanding of difficult science ideas having just completed a curriculum unit targeting those ideas. A different study involving students who had not immediately finished learning the ideas might produce results with fewer disagreements attributed to insecure knowledge. Finally, a few instances of disagreement could be attributed to students' guessing. In selected response items especially, guessing can be a source of error. It is possible that the inclusion of distracters based on common student ideas from the literature helped to reduce students' successful guessing. The use of common student ideas in the item development process has guided both SCALE-uP's and Project 2061's work in ways that may help to explain the relatively high levels of agreement between evidence sources in this study.

Research Question #2

The second research question concerns the extent to which students are able to consistently demonstrate their understanding of the forces and motion ideas during both the written MFA posttest and the interview. The question states:

> During both a written posttest and a videotaped oral interview, to what extent are middle school students in this study able to consistently demonstrate their understanding of the same forces and motion ideas targeted by a variety of assessment task contexts?

The results in Table 11 suggest that students vary in their understanding of forces and motion ideas. The data show one student unable to apply a correct understanding of motion and forces in any assessment context, nine students able to demonstrate a correct understanding in one or two contexts, and four students able to show understanding across three or more of the five contexts used in this study. This finding demonstrates the considerable variation with which students successfully use ideas about forces and motion in explaining phenomena. In addition, response patterns revealed that some contexts proved more accessible for students to demonstrate their understanding of the ideas (i.e., Falling Ball and A Metal Ball) while other contexts proved more difficult (i.e., Pedaling a Bike). In addition, while eight students correctly concluded that there was no force pushing a ball as it rolled in the Rolling Ball context, only four of those students were able to correctly respond to the question asking why the ball moves after it is kicked. From an assessment perspective, items and contexts that generate variability in student understanding of ideas are useful for measuring the range of understanding that exists within a single student and across many students. Assessments that provide such

variability in difficulty are more likely to result in inferences and score interpretations that more closely reflect students' knowledge and skills than assessments developed with items of similar difficulty levels (Crocker & Algina, 1986).

When comparing the judgment made by the written MFA and the holistic judgment based on the coded interview data, there appeared to be some level of disagreement between the two data sources of specific students (Table 11). These results indicated that, for three students in particular (Ingrid, Larry, and Darlene), overall judgments from the written MFA represent overestimations of students' understanding of the forces and motion ideas when compared to the interview results. This finding has implications for the standard setting procedure used for the MFA instrument. Specifically, the MFA levels of understanding (shown on page 82) may need to be revisited in light of the data presented in this study so that students with misconceptions are less likely to be characterized as having fluent understanding by the MFA instrument.

Research Question #3

The final research question addresses the discovery of unique student ideas identified during this study. The question states:

> Based on the elicited ideas of sixth graders in this study, what can this research add to the literature base (as presented in the study) about common ideas students have regarding forces and motion?

The discussion in Chapter IV presented several unique student ideas revealed in this study:

1. Friction causes a force to run out.
2. Friction does not act in a given direction.

3. Friction is not a force.
4. The mass of a falling object increases as it falls.
5. Smaller objects fall faster than heavier objects.

The first three ideas were included in the list of Student Idea Codes in this study because they were deemed as unique and sufficiently common in the transcript data. The other two student ideas (#4 and #5) were not as common as student responses but were considered unique because these ideas were not found in the literature review of research on middle school student ideas. These unique ideas will be discussed more now.

The idea that "friction causes a force to run out" was particularly interesting since many students understand the idea that friction will cause a rolling (or moving) object to slow down and eventually stop. The word friction is a relatively familiar term to students. Curriculum and instruction often explicitly introduce this idea and encourage students to incorporate this in their discussions of forces and motion. Several students in this study used the term, but did not appear to always understand its precise meaning. It may be that as students learn about new concepts such as friction and incorporate them into their existing frameworks about what causes changes in motion, they use opposing views about the nature of forces to explain motion. Some students appear to be assimilating this new idea of friction causing motion to change with their descriptions of forces running out in moving objects (an incomplete and naïve idea about force).

Other students in the study believed that friction was either not a force or that friction does not act in a given direction. Some students believed both ideas. First, the idea that friction is a force can be difficult for students to understand given that they typically recognize active forces and do not always appreciate the presence of passive

forces. Friction, in many ways, could be viewed as a passive force since it is not readily observable but acts to slow objects in motion. Perhaps one way to explicitly address this is to help students understand that friction does act in a given direction. If students come to understand that forces act in specific directions (often against motion), they may more readily accept the notion of friction as a force. This idea of directional forces (i.e., vectors) is somewhat more advanced and could go beyond the scope of middle school instruction.

One student believed that objects become heavier as they fall, and that this explained their increasing speed. This was a unique idea and observed once in this study. It is well known that many students believe that gravity pulls harder and harder (or the force of gravity becomes stronger) as an object falls. The idea that the object itself somehow gains mass (or weight) as it falls is a unique student idea identified in one interview transcript.

Another student expressed his belief that an object falls at a constant speed. This was based on a class activity in which a teacher demonstrated how two different objects fall at the same speed. The student may have believed that objects falling at the "same speed" meant that a given object maintains the "same speed" as it falls. Two different objects simultaneously falling at the same speed or an object falling at the same speed on consecutive drops should be distinguished from the constant pull of gravity that causes a falling object to increase speed. One student had difficulty distinguishing between two objects that fall at the same speed and a single object falling at the same speed in a single drop. While the specific case of gravity presents opportunities for demonstration and experimentation with acceleration of an object from the application of a constant force,

some students believed that objects fall at the "same speed." This confusion may be impeding their ability to demonstrate an understanding of the idea that a "change in speed is caused by a force." Therefore, the use of a falling object context, like the one presented on the MFA, may present a limitation in the assessment of this target idea. It may be important to couple this context with an additional assessment context (other than gravity) that represents a constant force (e.g., a pulled cart, accelerating car).

Finally, another student revealed in an interview that lighter objects fall faster than heavier objects (e.g., eggs vs. watermelons). This is the opposite of the well-documented misconception that heavier objects fall faster than objects that weigh less (Driver, et al., 1994). This unique idea may be related to the student's correct understanding that it is harder for heavier objects to be moved (i.e., it requires more force to push or pull a heavier object) while it is much easier for lighter objects to be moved. Since the force of gravity is constant, the lighter object will accelerate more quickly under a given force. It is somewhat surprising that more students do not hold this idea given that it maybe based on such an intuitive idea that could be considered consistent with students' lived experiences. However, the more likely experience with falling objects and the force of gravity is one in which lighter objects (e.g., leaves, paper) fall more slowly than heavier objects (e.g., rocks, books). In addition, many students may be able to successfully explain specific falling object phenomena based on their everyday experiences (e.g., objects' speed increasing as they fall) but this may not provide strong evidence about their understanding of the more general forces and motion ideas under study.

Discussion and Implications

This section will discuss the implications of the study in the context of the development and analysis of SCALE-uP assessment items. The section will discuss aspects of the Project 2061 Assessment Analysis procedure and some lessons learned in assessing middle school forces and motion content standards.

Project 2061 Assessment Analysis

As discussed in Chapter III, the Project 2061 Assessment Analysis uses specific criteria and a process of expert panels using the criteria to analyze and refine assessment items aligned to specific science content standards. Alignment of assessment items to target science ideas is achieved by using review criteria that carefully consider whether ideas in a content standard are needed to complete an assessment task successfully, if the task can be completed in another way, and whether those ideas are enough by themselves or if other ideas and skills are needed (DeBoer, 2005). The review criteria reduce the number of construct irrelevant factors that affect student responses. SCALE-uP used this procedure in assessment item development to encourage writers to consult the criteria while drafting items. Assessment items were also repeatedly analyzed and revised using the criteria to minimize the amount of variability from factors other than the students' use of the target ideas of interest. It is likely that the Project 2061 Assessment Analysis criteria played a critical role in developing items that produced evidence from which valid inferences can be made about student understanding. However, the present study showed that it is also important to hear the words of the students themselves when judging the validity of written test items. Based on the student interview data, there are a number of possible areas of improvement for the items used in this study.

The selected response item for the rolling ball context may provide an opportunity for "test-wiseness" in student responses given that there are only two possible answer choices. Users of the item should be aware of this possible limitation in light of the results of this study. Future revisions should carefully consider this aspect of the assessment item by revising to include additional answer choices. In addition, it appears from the interviews that the bicycle assessment context may be too complicated and difficult for students to process all of the information. The context may be overly complicated because it requires information processing resources that negatively affect students' ability to access long-term memory stores. In light of the Project 2061 Assessment Analysis criterion that addresses the *Appropriateness of Task Context,* this item is worth reviewing.

The use of such analysis criteria in item development and analysis was a critical component of SCALE-uP's development process. This aspect of SCALE-uP's assessment design work is conducted "in such a way that validity evidence emerges" (Mislevy, 2007, p. 467). The criteria allow teams of reviewers to carefully consider all aspects of an item that could elicit irrelevant knowledge or skills in a given student's response. It appears that, in many cases, the use of these criteria has led to items that provide valid inferences about students' knowledge of specific science ideas by preemptively identifying and reducing the presence of construct-irrelevant factors. Despite the apparent success of such criteria, this study demonstrated that even with analysis criteria being used by a panel of expert reviewers, items could still be developed and used which continue to include construct irrelevant factors, which affects students' ability to demonstrate their understanding of the science content.

Assessment of a Motion and Forces Standard

This study also offers important lessons about the assessment of "big ideas" in science education. The MFA represents a serious effort to create an instrument that shows educators and researchers how a middle school student can apply understanding of motion and forces to explain natural phenomena across a number of relevant contexts. From this study, it is evident that students can apply their understanding of "changes in speed or direction are caused by forces" with varying levels of success. Similar numbers of students (though not the same students) were able to demonstrate understanding of this idea in both the falling ball and skateboard contexts. However, these contexts represent two different kinds of motion (falling vs. rolling) and two different kinds of forces (gravity vs. friction) that cause the change in the motion. In fact, only one student in this study was able to demonstrate understanding in both the falling and rolling contexts during the interviews. For many students, applying this idea across different kinds of contexts is a challenge. Students' understanding of such an idea may be associated with a given context, such as the one in which it was learned. Prior to this study, students' exposure to the *Motion and Forces* curriculum materials and the opportunities afforded to learn these ideas may have been limited to the specific phenomena and activities included (e.g., sliding disks, rolling marbles, and accelerating carts). The materials provided few opportunities for students to apply their newly acquired ideas beyond the phenomena presented (Ochsendorf, et al., 2003). Thus, many of the students in this study may have found it difficult to demonstrate understanding across multiple, new contexts. Other research in cognitive science has also suggested that students' knowledge is context dependent and that transfer to other contexts presents a significant obstacle (Bransford,

Brown, & Cocking, 1999; NRC, 2001). This issue of transfer is particularly relevant in science education, where the goal of instruction is often to help students understand how specific concepts learned in a few limited contexts are relevant in explain natural phenomena more generally.

The specific case of gravity as a force presents interesting challenges for assessment development and evidence interpretations. In this study, many students appear to understand that gravity causes the speed of falling objects to increase but it is less clear that they understand the generalization of this idea (i.e., that forces cause change in speed). In development phases, SCALE-uP researchers considered the idea of "gravity as a force" to be a prerequisite idea which was supported by national standards documents (AAAS, 1993). However, it may be that students who understand the specific case (gravity pulls an object to the ground and that object's speed increases) may not understand the generalization (changes in speed are caused by forces). Based on the interview data in this study, it remains unclear if students are able to make such a connection to this more general idea. In this study, an understanding of the specific case of gravity and its effect on falling speed was seen as evidence of understanding the larger generalization. However, an understanding in a specific context does not always indicate that an individual understands the scientific generalization. This clearly represents a challenge for new kinds of assessment that attempt to determine if students can apply newly acquired knowledge across relevant contexts. This issue in this particular context requires further research.

The idea that "objects in motion stay in motion unless acted on by a force" is a difficult idea for students to understand because it is often counterintuitive to their lived

experiences and observations of moving objects (AAAS, 1993). The MFA assesses this idea in the rolling ball context and the bicycle context (developed by Project 2061). Only 4 of the 14 students were able to successfully demonstrate understanding of this idea in the rolling ball context during the interviews. Two of six students demonstrated understanding of the idea in the bicycle context, although these two students' correct interview responses did not agree with their incorrect written posttest responses. The bicycle context seems to be more complicated because students must consider the magnitude of forward and backward forces as the bicycle moves at a constant speed. It is unclear if a student who believes that there are balanced forces acting on objects in constant motion also believes that "objects in motion stay in motion." These may be slightly different ideas given the intent of the target idea.

The rolling ball items attempt to determine if students believe there is a force constantly acting on a rolling ball to make it move forward and why the ball moves after it is kicked. If students indicate that there is not a force constantly pushing the ball, SCALE-uP raters considered this as evidence that students understand "objects in motion will stay in motion." It is worth noting that in no situation during the interviews or written posttest did any student invoke Newton's First Law of Motion or make statements that closely mirrored the target idea of interest. SCALE-uP's assessment of this difficult idea was achieved through the rolling ball context and items. However, an additional assessment context that more directly elicits the idea that "objects in motion will stay in motion" is needed to definitively determine if students understand this idea. In this study, successful responses to the Rolling Ball items were sufficient for evidence of understanding the target idea.

Both assessment items (rolling ball and bicycle) make use of familiar contexts to assess an idea that is not readily observable in students' lives. In many instructional situations, it is often the goal to reduce or minimize the action of friction to move students toward an appreciation of the idea that objects in motion will stay in motion. While this is indeed the approach taken by the *Motion and Forces* materials used by students in this study, the assessment contexts in this study make no attempt to do so. It is likely that assessment contexts seeking to present such phenomena (i.e., ice or air hockey) use situations that are not familiar or accessible to all learners. This is perhaps a limitation of assessment in this domain. The students who responded successfully and consistently in the rolling ball context developed a more sophisticated view of the nature of forces (i.e., force is not synonymous with motion) and were able to recognize situations in which this is relevant. However, student beliefs that there was not a force constantly pushing a rolling ball forward or that the backward and forward forces acting on a bicycle in constant motion are equal may represent intermediaries to an understanding of the idea that objects in motion stay in motion.

The idea that "objects at rest stay at rest unless acted on by a force" is not as difficult for students to understand because this idea is more consistent with their lived experiences with forces and motion. However, as noted in Chapter II, it is often difficult for students to appreciate and understand the nature of passive forces (Gunstone & Watts, 1985). In this study, this idea was examined in the metal ball context, which asks students to consider the magnitude of the two forces that act on a stationary, hanging metal ball (gravity and a crane). Four of six students that responded to this item were able to demonstrate their understanding of the target idea. In this single context, students were

able to demonstrate an understanding that forces acting on a hanging object were balanced when the object was not moving. Therefore, students may have been demonstrating an understanding of a slightly more advanced idea that "objects at rest stay at rest unless acted on by a net force." Many students were aware of both the upward and downward forces although the object was at rest. In order to respond correctly to the item, students needed to understand that the magnitude of those two forces was equal and that there was no net force acting on the metal ball. This is perhaps a subtle but important distinction that is relevant to assessing student understanding of the idea that "objects at rest stay at rest unless acted on by a force."

Implications for Assessment Development

This study has implications for the processes used to develop assessments of students' science understanding. Considering the findings of this study, a thorough development and review process involving a comprehensive set of analysis criteria led to six items that provided solid inferences about student understanding. However, even when using such analysis criteria not all developed items provided valid inferences, as interview data for one of the items was not in agreement with the written posttest data. In this case, the disagreement came to light as a result of a systematic procedure for collecting interview data to confirm students' idea use while responding to test items. In the absence of such a validation effort, an ineffective assessment item would have continued to be used while providing misleading evidence of student understanding. Therefore, it is important that test development procedures include such validation efforts. Although the process of collecting, coding, and analyzing interview data is time consuming, assessment development should include such systematic efforts in the

piloting phases of the development process. By including such analyses in the early piloting phases of assessment development, researchers can be certain that direct evidence collected as students respond to the cognitive demands of assessment situations will be included in the development efforts, as called for by Leighton (2004) and Gorin (2006). Ideally, developers should engage in an iterative process of development, data collection, and revision based on a variety of evidence sources, which includes verbal reports from students as well as psychometric data. The findings from this study suggest that this process ought to occur prior to test use so that the resulting evidence can lead to more valid inferences regarding students' knowledge and skills. Subsequent analyses involving statistical psychometric validation procedures, which investigate the internal structure of the test or its relationship to other measures, should be utilized as well. Researchers can often gain additional insights from looking at data patterns from greater numbers of student responses.

Additionally, as researchers adopt existing tests for their own uses (with often different purposes), it is incumbent upon the researcher to confirm that the measurement construct that was intended by the instrument is in fact the construct that was enacted in the new setting (Gorin, 2006). Many researchers have viewed test validation as an ongoing process, which is interested in the links between test scores and test use (Mislevy, 2007; Messick, 1989). In such validation efforts, smaller scale qualitative studies that interrogate student idea use during assessment situations can reveal a great deal about the constructs that are enacted as student responses lead to score interpretations and decision-making in new testing situations.

Conclusions and Recommendations

Based on this study, the assessment items under investigation generally produced evidence that led to valid inferences about student understanding of a "big idea" in middle school science education. In addition, the methods and analysis techniques put forth in this study represented a practical and logical way to draw inferences about the existence of student ideas in the student population. Given that few research efforts have attempted to account for student ideas at this level of analysis in assessment validation, this study contributes a valuable approach to an important line of inquiry. The role of assessment and the quality of inferences they produce has become an increasingly high-stakes endeavor. Assessment development and validation efforts that fail to consider the ways in which student ideas are used during assessment do not provide students and educators with the most effective opportunity to demonstrate what has been taught and what has been learned.

This study was motivated, in part, by the premise that validation efforts in the assessment community do not take into account enough qualitative evidence collected from individual students who respond to assessment items. The lack of research was presented in Chapter II. This study attempted to address this deficit in validation research within a specific science domain, motion and forces, by using a targeted group of students. As a result, the study utilized a sensible methodology for other researchers to use in assessment development. The study responds to a call in the literature (Gorin, 2006; Leighton, 2004) for research to bridge a gap in assessment validation, uniting the fields of cognitive psychology and educational measurement. The study pays careful attention to students' cognitive activity and whether it matches the assumptions set forth

in assessment development. By doing so, the study illuminates some of the challenges faced by assessment developers and educators in the measurement of student science learning (NRC, 2001).

As views of learning have become more sophisticated, the expansion of conceptions of knowledge and skill about which inferences are desired has become increasingly apparent (Mislevy, 2007, p. 464). More traditionally, learning was defined as students' ability to demonstrate understanding by responding with declarative knowledge (i.e., facts, formulas, names, etc…). *Knowing What Students Know* (NRC, 2001) describes more contemporary views of learning that consider the cognitive, situative, and differential perspectives that describe learning. In addition, a more common perspective of learning is one that takes into account how students use specific ideas and in what contexts. By applying ideas in specific situations to explain relevant phenomena, students appropriately demonstrate their understanding of those ideas. These expanding views of learning and understanding have challenged assessment design experts to arrive at more sophisticated measurement strategies to determine if students can actually demonstrate the kind of learning required by national standards. This view of learning, in part, guided the assessment development efforts of SCALE-uP, the development of the assessment items in this study, and the design of the curriculum materials that students used prior to participating in interviews for this research. This contemporary view of learning and the challenges inherent in assessing such learning guided this study's methods for collecting validation evidence for an assessment designed to reflect such views. This evidence is qualitative in nature and can help educational measurement researchers make decisions

about whether or not assessment items are measuring the kind of learning expected of students in today's classrooms.

This approach has also collected validation evidence for the MFA instrument in a way that aligns an assessment item, the student's response to a given item, the subsequent inference made about the student's understanding, and the nature of the content targeted by the item. Thus, the study integrated the three components of the assessment triangle (observation, interpretation, cognition) (NRC, 2001) to produce convincing assessment validation evidence.

Critique of Study

This study relied on an interview methodology with a structured protocol to collect data about student ideas. Interviewing about student cognition is limited by the characteristics of the interview setting, the questions, and the ability of the student to accurately verbalize thought processes. While the ultimate interview data presented in this study yielded rich information about student ideas, the data analysis could have been enriched by an additional follow-up question or probing attempt. The interviews revealed a great deal about student understanding, but they were not perfectly conducted. No student is able to fully express his or her thought processes in the limited amount of time available. These were practical limitations of this study.

In addition, the students in this study had recently completed several weeks of instruction that focused explicitly on the target ideas in this study. It is likely that the state of their knowledge and understanding was in transition as they attempted to reconcile prior understandings with the ideas learned in the classroom. This scenario could

potentially limit the quality of the evidence because student ideas may be rapidly changing.

Another limitation of this study is related to language use by the participants in the study. Although interview data is often considered to be the best available evidence from which to make inferences about individuals' cognition, such data does not necessarily overcome the deficiencies and biases in the language used by both the interviewer and the student. Both the interviewer and the student's use of terms is influenced by basic literacy, common language use, culture, and even the use of specific terms emphasized in recent experiences, including exposure to curriculum materials. For this study, the curriculum materials frequently introduced scientific terms to students which may have had an influence on the nature of scientific terms used during the interviews. Thus, the variety of factors influencing the ways in which specific language was used or not used by participants in the study was not explored and likely influenced the findings.

The forces and motion ideas presented in this study represent an important but small part of a middle school physical science curriculum. Therefore, this study has important limitations within the content domain. The assessment items and relevant contexts limit the scope of discussion and narrow the implications of the findings. A study with interviews that use different items or different contexts with the same students would produce different results. Students' understanding tends to be highly context-dependent. Therefore, the phenomena presented to students could prove disadvantageous to some students who have learned in different contexts than the ones used in this study.

The issue of transfer of ideas is a challenge for assessment design and was a limitation in this study.

Appendix A: SCALE-uP Forces and Motion Student Interview Protocol

Typically in assessment validation work, researchers can collect evidence from four main areas: a) evidence based on test content, b) evidence based on internal structure, c) evidence based on relation to other variables, and d) evidence based on student response processes. In this interview protocol development work, the evidence being collected from student response processes is being used to confirm the decisions/inferences that are made by the test items regarding individual student knowledge structures. That is, if a student responds correctly to an item, has the student "used" the appropriate piece of knowledge that the item was interested in? If so, did they use additional knowledge to respond to the item? If not, how were they able to arrive at a correct response? Gathering evidence from students' response processes to address these kinds of questions can help to confirm that the assessment (and perhaps the framework that guided its development) is functioning as intended.

Conversely, if a student does not respond correctly to an item, we infer that the student does not possess the particular piece of knowledge that the item was interested in. Can we collect evidence to strengthen the nature of this inference? In other words, the interview is structured in such a way as to attempt to provide evidence to either support or refute that inference. If a student does not respond correctly, is it because they don't possess understanding or is it because they did not understand the item, or some other reason? If it is due to the comprehensibility of the item, can students subsequently demonstrate evidence of understanding on the item after clarification? If students needed additional knowledge to be able to respond correctly, what was that additional knowledge?

These are the kinds of the questions that guided the development of the interview protocol and the interview itself.

Goals of Student Interviews

This interview protocol was designed so that students will first read the question and then provide an answer on their own. After the student has responded in writing, the interview protocol is designed to then, a) elicit confirmatory evidence of the students' understanding of the target idea, b) assess students' comprehension of the tasks/diagrams, and c) provide other opportunities for students to express their understanding of the target ideas in other contexts (i.e., by using 3-dimensional models or other representations of the phenomena).

Introduction:

SAY: *Hello, my name is _____. I am interested in the ways that you think about what causes motion and how objects move. I would like to ask you some questions about your ideas. I am going to show you a few questions, give you time to answer the questions on your own, and then I will ask you about some of your ideas. Do you have any questions for me before we get started?*

Part I. A Rolling Ball

SAY: *Please read this question out loud.*
THEN SAY: *Take a minute to answer the question in writing.*
Student reads and answers question on own in writing. Interviewer watches student's response and asks the student to read his/her response.
SAY: *Can you read your response out loud?*

Interviewer evaluates response for evidence of understanding target idea based on the assessment item rating guide (objects in motion will stay in motion unless acted upon by a force). Based on student's response, the interviewer can proceed in one of two directions.

1. Rating Guide Correct (Circles "No" and Correct Response): If student's response reveals evidence of understanding the target idea, according to rating guide below, probe for necessity and sufficiency (i.e., did student use the target idea to get correct response?).

Probing Questions:
Can you explain what this question was about?
How did you go about answering this question?
How did you eliminate the "Yes" choice?
Why is there no force constantly pushing the ball?
If there is no force constantly pushing the ball, what is causing it to move?
OR *Why does the ball continue to move if there is no force constantly pushing it?*
If there were a force constantly pushing the ball, what would happen?
Was there anything else that helped you answer this question?
How did you know this?
Where did you learn this?
END TASK

There are two possibilities based on the responses to the probing questions:
Option 1A. Student used the target idea in their response, according to necessity criterion. If student uses the target idea and is successful in responding to the above probes, it is acceptable to move on to the next task.
Option 2B. Student did not use the target idea in correctly answering question, necessity is violated (i.e., test-wiseness, pre-requisite ideas used). If student does not use the target idea, responses to the above probes will provide information

necessary to determine how correct response was demonstrated without the use of the idea (necessity violation).

2. **Rating Guide Incorrect (Circles "Yes" or "No and an incorrect response"):** If student's response reveals no evidence of understanding the target idea, probing questions are then asked to assess whether student comprehends task (question, diagram, stem, context).

Probing Questions:
Can you explain what this question was about?
Were there any words or sentences that were hard to understand?
Does the picture help you answer the question? How?
Can you give an example of a smooth floor? What does it mean when it says constantly?

Based on student responses to these probes, interviewer needs to assess whether the student comprehends the task and then proceed in two possible directions:

Option 2A: If student does NOT comprehend task, END TASK. Above probes should be sufficient to collect evidence to help determine what about the task was incomprehensible to the student. Interviewer can move on to next task.

Option 2B: If there is evidence that student DOES comprehend task, ask the following questions to determine if the student can use the target idea with other ideas to correctly explain the phenomena or has a misconception:

Can you describe how you went about answering this question?
How did you select your answer for the first question? (i.e., Yes or No)
How did you eliminate the other option?
Why does ball move to you after it is kicked? OR *What makes the ball move after it is kicked?*
Can you explain to me what you mean by_____ (i.e., force, push, friction, motion)? (Ask student to further explain words they have used).
Is there anything else you'd like to know about the floor? the ball? Or the kick?
How did you know this?
Where did you learn this?
END TASK

i) If student used the target idea in their response, sufficiency is violated. Student responses to the above probes will provide information necessary to determine how sufficiency criterion was violated. Move to Next Task (For Analysis: what else did the student need to know to correctly respond to the item initially? In other words, student is able to use idea to explain phenomena but is unable to do so on assessment for some reason.)

ii) If student is unable to use the target idea to answer question, move to next question.

Part II. Dropping and Throwing a Ball

SAY: *Please read this question out loud.*

THEN SAY: *Take a minute to answer the question in writing.*

SAY: *Can you read your response out loud?*

Student reads and answers question on own in writing. Interviewer watches student's response.

Interviewer evaluates response for correctness (TARGET IDEA= *Changes in speed or direction are caused by forces* and SR Answer Key=*Speed increases*; CR Answer Key: *Gravity/Force cause the speed of the tennis ball to increase*.). Based on student's response, the interviewer can proceed in one of two directions.

1. Rating Guide Correct: ("Speed Increases" bubbled in and a Correct Written Response). If student's response reveals evidence of understanding the target idea, probe for necessity and sufficiency (i.e., did student use the target idea to get correct response?).

Probing Questions:
Can you explain what this question was about?
Can you explain why you selected that response? (selected response)
How did you eliminate the other options? (selected response)
Are there other answers you almost chose? Why? (selected response)
What do you mean by gravity/force? (Interviewer checks here for correct use of the target idea). *If you were to use another word what would it be?*
Why does the gravity/force/pull make the ball's speed increase?
If there were no gravity, what would happen when Katie let go of it?
Was there anything else that helped you answer this question?
How did you know this?
Where did you learn this?
END TASK

There are two possibilities here based on student responses to probing questions:
Option 1A. Student used the target idea in their response, according to necessity criterion. If student uses the target idea and is successful in responding to the above probes, it is acceptable to move on to the next task.
Option 1B. Student did not use the target idea in correctly answering question, necessity is violated (i.e., test-wiseness, pre-requisite ideas used). If student does not use the target idea, probing questions will help determine how correct response was demonstrated without use of target idea (necessity violation; guessing, test-wiseness, etc…). Can move on to next task.

2. Rating Guide Incorrect: (Any circle bubbled and an incorrect written response). If student's response reveals no evidence of understanding target idea, probing questions are then asked to assess whether student comprehends task (question, stem, diagram, context).

Probing Questions:
Can you explain what this question was about?
What is Katie doing? What is Maria doing?
Does the picture help you answer the question? How?
Were there any words or sentences that you did not understand?
What do you think "speed" means?

Based on student responses to these probes, interviewer needs to assess whether the student comprehends the task and then proceed in two possible directions:

Option 2A: If student does not comprehend task, END TASK. Above probes should be sufficient to collect evidence to help determine what about the task was incomprehensible to the student. Interviewer can move on to next task.

Option 2B: If there is evidence that student comprehends task, ask the following questions to determine if the student can use the target idea with other ideas to correctly explain the phenomena or has a misconception:

Can you describe how you went about answering this question?
Can you explain why you selected that response? (Selected response)
Are there other answers you almost chose? Why? (Selected response)
What do you mean by gravity/force/pull? (Interviewer checks here for correct use of the target idea).
Why does the ball's speed increase?
How did you know this?
Where did you learn this?
END TASK

i) If student used the target idea in their response, sufficiency is violated. Student responses to the above probes will provide information necessary to determine how sufficiency criterion was violated. Move to next task (For Analysis: what else did the student need to know to correctly respond to the item initially? In other words, student is able to use idea to explain phenomena but is unable to do so on assessment for some reason.)

ii) If student is unable to use the target idea to answer question, move to next task.

Part III. AAAS-P2061 Item 1 (Carly's Skateboard)

SAY: *Please read this question out loud.*

THEN SAY: *Take a minute to answer the question. Fill in the bubble.*

Student reads and answers question on own in writing. Interviewer watches student's response.

Interviewer evaluates response for correctness (Answer Key=1st response option). Based on student's response, the interviewer can proceed in one of two directions.

1. Rating Guide Correct: If student's response reveals evidence of understanding target idea, (Selects 1st response option) probe for necessity (i.e., did student use the target idea to get correct response?). (TARGET IDEA = Changes in speed or direction are caused by forces.)

Probing Questions:
Can you explain what this question was about?
Can you explain why you selected that response?
How did you eliminate the other answer choices?
Are there other answers you almost chose? Why?
What do you think it means by "force in the opposite direction"?
Why does this stop the motion of Carly's skateboard?
How did you know this?
Where did you learn this?
END TASK

There are two possibilities here based on student responses to probing questions:
Option 1A. Student used the target idea in their response, according to necessity criterion. If student uses the target idea and is successful in responding to the above probes, it is acceptable to move on to the next task.
Option 1B. Student did not use the target idea in correctly answering question, necessity is violated (i.e., test-wiseness, pre-requisite ideas used). If student does not use the target idea, above probes will provide evidence to determine how correct response was demonstrated without use of target idea (necessity violation; guessing, test-wiseness, etc…).

2. Rating Guide Incorrect: If student's response reveals no evidence of understanding target idea (student selects 2nd, 3rd, or 4th response), probing questions are then asked to assess whether student comprehends task (question, stem, context).

Probing Questions:
Can you explain what this question was about?
What is Carly doing?
Were there any words or sentences that you did not understand?
What do you think "force" means?

Based on student responses to these probes, interviewer needs to assess whether the student comprehends the task and then proceed in two possible directions:

Option 2A: If student does not comprehend task, above probes should be sufficient to collect evidence to help determine what about the task was incomprehensible to the student. Interviewer can move on to next task. END TASK.

Option 2B: If there is evidence that student comprehends task, ask the following questions to determine if the student can use the target idea with other ideas to correctly explain the phenomena or has a misconception:.

Can you describe how you went about answering this question?
Can you explain why you selected that response?
Are there other answers you almost chose? Why?
How did you eliminate the first answer?
If student chose the 2nd response, ASK, *Can you describe how the skateboard runs out of force, which causes the skateboard to stop?*
If student chose the 3rd response, ASK, *Can you describe how the force of the pushing decreasing to zero causes the skateboard to stop?*
If student chose the 4th response, ASK, *What does it mean if there are no forces acting on the skateboard?*
Are there any forces in the opposite direction of the skateboard's motion?
END TASK

i) If student used the target idea in their response, probing questions should provide evidence needed for analysis to determine how student was unable to demonstrate understanding on the item. Student is able to use idea to explain phenomenon during interview but is unable to do so on assessment. In other words, what else did the student need to know to correctly respond to the item initially?

ii) If student is unable to use the target idea to answer question, interviewer's prior probes should provide evidence for students' misconceptions around the target idea.

Part IV. AAAS-P2061 Item 2 (A Metal Ball)

SAY: *Please read this question out loud.*

THEN SAY: *Take a minute to answer the question in writing. Fill in the bubble.*

Student reads and answers question on own in writing. Interviewer watches and evaluates student's response.

Interviewer evaluates response for correctness (Answer Key=3rd response option). Based on student's response, the interviewer can proceed in one of two directions.

1. Rating Guide Correct: If student's response reveals evidence of understanding target idea, (Selects 3rd response option) probe for necessity (i.e., did student use the target idea to get correct response?). (TARGET IDEA = An object at rest will stay at rest unless acted upon by a force.)

Probing Questions:
Can you explain what this question was about?
Can you explain why you selected that response?
How did you eliminate the other answer choices?
Are there other answers you almost chose? Why?
What do you mean by 'force/gravity/pull'?
Can you describe how the forces on the metal ball are the same?
When objects are not moving, such as the hanging metal ball, what can you tell me about the forces that are acting on the ball?
How did you know this?
Where did you learn this?
END TASK

There are two possibilities here based on student responses to probing questions:
Option 1A. Student used the target idea in their response, according to necessity criterion. If student uses the target idea and is successful in responding to the above probes, it is acceptable to move on to the next task.
Option 1B. Student did not use the target idea in correctly answering question, necessity is violated (i.e., test-wiseness, pre-requisite ideas used). If student does not use the target idea, above probes will determine how correct response was demonstrated without use of target idea (necessity violation; guessing, test-wiseness, etc…).

2. Rating Guide Incorrect: If student's response reveals no evidence of understanding target idea (student selects 1st, 2nd, or 4th response), probing questions are then asked to assess whether student comprehends task (question, stem, context).

Probing Questions:
Can you explain what this question was about?
What is happening to the metal ball?
How was the picture helpful?
Were there any words or sentences that you did not understand?
What do you think "force" means?
What does gravity mean?

Based on student responses to these probes, interviewer needs to assess whether the student comprehends the task and then proceed in two possible directions:

Option 2A: If student does not comprehend task, above probes should be sufficient to collect evidence to help determine what about the task was

incomprehensible to the student. Interviewer can move on to next task. END TASK.

Option 2B: If there is evidence that student comprehends task, ask the following questions to determine if the student can use the target idea with other ideas to correctly explain the phenomena or has a misconception.

Can you describe how you went about answering this question?
Can you explain why you selected that response?
Are there other answers you almost chose? Why?
How did you eliminate the third answer? How did you eliminate the 3rd answer choice?
If student chose the 1st response, ASK, *Can you describe how the force of gravity on the ball is stronger than the force of the crane?*
If student chose the 2nd response, ASK, *Can you describe how the force of the crane on the ball is stronger than the force of gravity?*
If student chose the 4th response, ASK, *What does it mean if there are no forces acting on the ball?*
END TASK

i) If student used the target idea in their response, probing questions should provide evidence needed for analysis to determine how student was unable to demonstrate understanding on the item. Student is able to use idea to explain phenomenon during interview but is unable to do so on assessment. In other words, what else did the student need to know to correctly respond to the item initially?

ii) If student is unable to use the target idea to answer question, interviewer's prior probes should provide evidence for students' misconceptions around the target idea.

Part V. AAAS-P2061 Item 3 (Pedaling a Bike)

SAY: *Please read this question out loud.*

THEN SAY: *Take a minute to answer the question in writing. Fill in the bubble.*

Student reads and answers question on own in writing. Interviewer watches student's response.

Interviewer evaluates response for correctness (Answer Key=3rd response option). Based on student's response, the interviewer can proceed in one of two directions.

1. Rating Guide Correct: If student's response reveals evidence of understanding target idea, (Selects 3rd response option) probe for necessity (i.e., did student use the target idea to get correct response?). (TARGET IDEA = Objects in motion stay in motion unless acted upon by a force.)

Probing Questions:
Can you explain what this question was about?
Can you explain why you selected that response?
How did you eliminate the other answer choices?
Are there other answers you almost chose? Why?
What do you mean by 'force/total force'?
What are the forces that are acting in the forward and backward direction? Can you describe them?
How did you know this?
Where did you learn this?
END TASK

There are two possibilities here based on student responses to probing questions:

Option 1A. Student used the target idea in their response, according to necessity criterion. If student uses the target idea and is successful in responding to the above probes, it is acceptable to move on to the next task.
Option 1B. Student did not use the target idea in correctly answering question, necessity is violated (i.e., test-wiseness, pre-requisite ideas used). If student does not use the target idea, above probes will provide evidence to determine how correct response was demonstrated without use of target idea (necessity violation; guessing, test-wiseness, etc…).

2. Rating Guide Incorrect: If student's response reveals no evidence of understanding target idea (student selects 1st, 2nd, or 4th response), probing questions are then asked to assess whether student comprehends task (question, stem, context).

Probing Questions:
Can you explain what this question was about?
Were there any words or sentences that you did not understand?
How was the picture helpful?
What happens when Jeremy pedals his bike? What happens when Jeremy does not pedal?
What do you think "total force" means? What does a forward force mean? What does a backward force mean?

Based on student responses to these probes, interviewer needs to assess whether the student comprehends the task and then proceed in two possible directions:

Option 2A: If student does not comprehend task, above probes should be sufficient to collect evidence to help determine what about the task was incomprehensible to the student. Interviewer can move on to next task.
Option 2B: If there is evidence that student comprehends task, ask the following questions to determine if the student can use the target idea with other ideas to correctly explain the phenomena or has a misconception.

Can you describe how you went about answering this question?
Can you explain why you selected that response?
What happens when Jeremy pedals his bike?
What happens when Jeremy does not pedal?
What do you think "total force" means? What does a forward force mean? What does a backward force mean?
Are there other answers you almost chose? Why?
How did you eliminate the 3rd answer?

If student chose the 1st response, ASK, *Can you describe how the total force on the bike in the forward direction is greater than the total force on the bike in the backward direction? What are the forces that act in the forward direction? What are the forces that act in the backward direction?*
If student chose the 2nd response, ASK, *Can you describe how the total force on the bike in the forward direction is less than the total force on the bike in the backward direction? What are the forces that act in the forward direction? What are the forces that act in the backward direction?*
If student chose the 4th response, ASK, *What does it mean if there are no forces acting on the bike in either direction?*
END TASK

i) If student used the target idea in their response, probing questions should provide evidence needed for analysis to determine how student was unable to demonstrate understanding on the item. Student is able to use idea to explain phenomenon during interview but is unable to do so on assessment. In other words, what else did the student need to know to correctly respond to the item initially?

ii) If student is unable to use the target idea to answer question, interviewer's prior probes should provide evidence for students' misconceptions around the target idea.

Part VI. Modeling of Phenomena

Crane Item
SAY: *Now* we *are going to explore some of these ideas about forces and motion and using some objects.*

SAY: Here is a toy crane. A ball is hanging from the crane and not moving.

Probing questions:
Can you describe the forces acting on the ball when it is hanging?

Interviewer decides among following probes and potentially revisits previous conversations:

> If two forces:
> *What is the force pulling down?*
> *What is the force pulling up?*
> *Is the strength of the forces the same or different? How are they the same or different?*
> *If they are the same, how does this affect the motion of the ball?*
>
> If just gravity?
> *Why did the tennis ball fall and this ball is not falling?*
> *What keeps the ball from falling to the ground?*
> *How does the crane keep the ball from falling?*

Skateboard Phenomena
SAY: *Here is a skateboard. I'm going to give it a small push a few times and let's watch what happens.* (Push the skateboard with a small push so that it only moves part way across the table).

Probing questions:
Why does the skateboard slow down and stop?

Interviewer decides among following probes and potentially revisit previous conversations:

What do you mean by friction/force? How does that work to make the skateboard stop?
What would happen if I pushed it harder? Would it have stopped? Why?
Is there anything stopping the skateboard? OR Is there a force acting in the opposite direction?

Rolling Ball Phenomena
SAY: *Here is a soccer ball. I'm going to give it a push and let's watch what happens.* (Interviewer pushes soccer ball across the ball to the student. Push the ball hard enough so that it doesn't stop before it reaches the student).

Probing questions:
Why does the ball continue to move to you after I push it?
Is there a force constantly pushing the ball to you?

Appendix B: Evidence Statements and Student Idea Codes for Each Assessment Item

Item	Evidence Statements	Student Idea Codes
Rolling Ball—Selected Response. Is there a force constantly pushing the ball toward you?	Student selects "Yes" and describes: *The force hasn't run out yet and it is still acting.* *The force you applied is still acting on it but it's not constant.* *The force is constantly going.* *Objects don't move on their own so there is a force pushing it.* *The ball keeps going and if the force stops then it will slow down.*	5. Force must act if objects are to stay in motion
	The ball has the force from the kick.	6. Objects in motion *have* a force in the direction of their motion (forces are acquired property)
	Student selects "No" and describes: *There is no constant force pushing.* *There was just one kick.* *There is no force still kicking.*	Correct. Objects in motion stay in motion unless acted on by a force.
Interviewer Probes: How did you eliminate the "Yes/No" choice?	*It depends on how hard you kick it.*	2. Students associate amount of motion with amount of force
Why is there no force constantly pushing the ball? If there is no force constantly pushing the ball, what is causing it to move? Why does the ball continue to move if there is no force constantly pushing it?	*Force has or has not worn off.*	4. Decreasing motion (slowing down) is the result of decreased force or force running out as opposed to a force (i.e., friction) in the opposite direction.

Item	Evidence Statements	Student Idea Codes
Rolling Ball—Constructed Response. What is causing the ball to continue to roll toward you?	*There is no constant force pushing.* *There was just one kick.* *There is no force still kicking.*	Correct. Objects in motion stay in motion.
	The force you applied is still acting on it but it's not constant. *Objects don't move on their own so there is a force pushing it.* *The ball is moving so there must be a force constantly pushing it.*	5. Force must act if objects are to stay in motion
	It depends on how hard you kick it.	2. Students associate amount of motion with amount of force
	The force hasn't run out yet and it is still acting. *When the force runs out, the ball will stop rolling.* *Force hasn't worn off.*	4. Decreasing motion (slowing down) is the result of decreased force or force running out as opposed to a force (i.e., friction) in the opposite direction.
	The ball can move on its own.	1. Inanimate objects try to stop (or keep going) on their own accord
Interviewer Probes: Why is there no force constantly pushing the ball? If there is no force constantly pushing the ball, what is causing it to move? Why does the ball continue to move if there is no force constantly pushing it? Why does ball move to you after it is kicked? OR What makes the ball move after it is kicked?	Use same codes as above.	

Item	Evidence Statements	Student Idea Codes
Dropping and Throwing a Ball—Selected Response. Does the speed of the tennis ball increase, decrease or stay the same as it falls?	Selects "Increase" and describes: *Gravity.* *It will increase because of gravity.* *Gravity is pulling it along.*	Correct. Changes in speed are caused by forces.
	Selects "Decrease" and describes: *Gravity decreases as the ball falls to the ground.*	No Relevant Code (misconception not related to target idea)
	Selects "Stay the Same" and describes: *Pull of gravity is constant.* *Gravity does not get stronger or weaker.* *Gravity is constantly pulling.* *Ball falls at same speed.*	No Relevant Code (misconception not related to target idea)
Interviewer Probes: Can you explain why you selected that response? How did you eliminate the other options? Are there other answers you almost chose? Why?	*Gravity pulls faster and faster* *Gravity pulls more and more.* *It will increase because there is no friction.*	No Relevant Code (misconception not related to target idea)

Item	Evidence Statements	Student Idea Codes
Dropping and Throwing a Ball —Constructed Response. What causes the speed of the tennis ball to increase, decrease, or stay the same?	Student describes: *It will increase because of gravity.* *Gravity is pulling it down.*	Correct. Changes in speed or direction are caused by forces.
	Gravity pulls harder as the ball falls.	No Relevant Code (misconception not related to target idea)
	The ball possesses force.	6. Objects in motion have a force in the direction of their motion (forces are acquired property)
	The ball's weight or heaviness causes it to fall the way it does.	No Relevant Code (misconception not related to target idea)
Interviewer Probes *What do you mean by gravity/force? If you were to use another word what would it be?* *Why does the gravity/force/pull make the ball's speed increase?* *Why does the ball's speed increase?*	*It will increase because there is no friction.*	No Relevant Code (misconception not related to target idea)

Item	Evidence Statements	Student Idea Codes
Carly's Skateboard. Why does the skateboard come to a stop?	Student selects "Because there is a force in the opposite direction of the skateboard's motion" and describes:	
	It doesn't stop until something else stops it. *It was because of the friction.* *It was because of the force in the opposite direction.* *When the wheels touch the ground, it slows it down.*	Correct. Changes in speed or direction are caused by forces.
	Student selects "Because the skateboard runs out of force" and describes:	
	The force of Carly's push runs out and the skateboard stops.	4. Decreasing motion (slowing down) is the result of decreased force or force running out as opposed to a force (i.e., friction) in the opposite direction.
	Carly did not push hard enough. *If she had pushed harder, she would still be moving.*	2. Students associate amount of motion with amount of force
	Force of the push on the skateboard would decrease because the friction was slowing it down.	8. Friction causes a force to run out.
	Student selects "Because the force of the push on the skateboard decreases to zero" and describes:	
	The force gets used up because of friction.	4. Decreasing motion (slowing down) is the result of decreased force or force running out as opposed to a force (i.e., friction) in the opposite direction.
	There aren't any forces, so the skateboard will stop. *If you push something, it can't move forever.*	5. Force must act if objects are to stay in motion.
	Student selects "Because there aren't any forces on the skateboard" and describes:	
	There aren't any forces, so the skateboard will stop.	5. Force must act if objects are to stay in motion.
	Friction does not act in a specific direction (opposite).	No Relevant Code (misconception not related to target idea)

	The skateboard was trying to stop.	1. Inanimate objects try to stop (or keep going) on their own accord.
	Friction can't come at you or act in a given direction.	9. Friction does not act in a given direction.

Item	Evidence Statements	Student Idea Codes
Pedaling a Bike. When he pedals, the bike moves at a constant speed. Which of the following is true about the forces on the bike while he is pedaling?	Student selects "The total force on the bike in the forward direction is equal to the total force on the bike in the backward direction" and describes:	
	The forward force of his legs and the backward force of friction are the same. *The forces are equal.*	Correct. Objects in motion stay in motion unless acted on by a force.
	Student selects "Forces on the bike in the forward direction are greater than the forces in the backward direction" and describes:	
	That the force had to be greater to move the bike forward.	6. Objects in motion have a force in the direction of their motion (forces are acquired property)
	Student selects "Forces on the bike in the forward direction are less than the forces in the backward direction" and describes:	
	Undocumented	Possible unique explanations:
	Student selects "There are no forces on the bike in either the forward or backward direction" and describes:	
	Undocumented	Possible unique explanations:

Interviewer probes: How did you eliminate the other answer choices? Can you explain why you selected that response? How did you eliminate the other answer choices? Are there other answers you almost chose? Why? What do you mean by 'force/total force'? What are the forces that are acting in the forward and backward direction? Can you describe them?	*If it was equal, the bike wouldn't be moving.* *If there were no forces, the bike wouldn't be moving so there must be some forces.*	7. If a body is not moving, there is no force acting on it.

Item	Evidence Statements	Student Idea Codes
A Metal Ball. A metal ball is hanging from a crane and is not moving. Which of the following is true?	Student selects "The force of gravity on the ball and the force of the crane on the ball are the same" and describes: *If gravity were stronger it would move down and if the crane were stronger, it would go up.*	Correct. Objects at rest will stay at rest unless acted on by a force.
	Student selects: "The force of gravity on the ball is stronger than the force of the crane on the ball" and describes: *Gravity holding the ball up.*	Possible unique explanation:
	Student selects "The crane applies more force on the ball than gravity" and describes: *The crane is holding the ball up and keeping it from falling.*	Possible unique explanation:
	Student selects "There are no forces acting on the ball" and describes: *If the forces are the same, the ball can't go anywhere.* *If the ball is not moving, there are no forces acting.*	7. If a body is not moving, there is no force acting on it

Interviewer probes: Can you explain why you selected that response? How did you eliminate the other answer choices? Are there other answers you almost chose? Why? What do you mean by 'force/gravity/pull'? Can you describe how the forces on the metal ball are the same? When objects are not moving, such as the hanging metal ball, what can you tell me about the forces that are acting on the ball?	*If the force of gravity were greater, it would go down.* *If the force of the crane were greater, it would go up.* *Ball is not moving so it must be the same amount of force.*	Correct. Changes in speed or direction are caused by forces.

References

Abell, C. F., & DeBoer, G. E. (2007, April). *Probing middle school students' knowledge of thermal expansion and contraction through content-aligned assessment.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

American Association for the Advancement of Science [AAAS]. (1993). *Benchmarks for Scientific Literacy.* New York: Oxford University Press.

American Association for the Advancement of Science [AAAS]. (2001). *Atlas of science literacy*. New York: Oxford University Press.

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. (1999). *Standards for educational and psychological testing*. Washington, DC.

Anderson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York: W.H. Freeman.

Augustine, N. R. (2006). *Rising above the gathering storm: Energizing and employing America for a brighter economic future*. Washington, DC: National Academy Press.

Baddeley, A. (1986). *Working memory*. Oxford: Clarendon Press/Oxford University Press.

Baddeley, A. (1990). *Human memory: Theory and practice.* Boston, MA: Allyn & Bacon, Inc.

Baxter, G. P., & Glaser, R. (1998). The cognitive complexity of science performance assessments. *Educational Measurement: Issues and Practice, 17*, 37-45.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(1061-1071).

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. *Cognition, 9*, 117-123.

Carlsen, W. S. (2007). Language and science learning. In N. G. Lederman (Ed.), *Handbook of research on science education*. Mahwah, NJ: Lawrence Erlbaum.

Champagne, A. B., Klopfer, L. E., & Anderson, J. H. (1980). Factors influencing the learning of classical mechanics. *American Journal of Physics, 48*(12), 1074-1079.

Chi, M. T. H., Feltovich, P., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*, 121-152.

Cizek, G. J. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ.: Lawrence Erlbaum Associates.

Clement, J. (1982). Students' preconceptions in introductory mechanics. *American Journal of Physics, 50*(1), 66-71.

Clement, J. (2000). Analysis of clinical interviews: Foundations and model viability. In R. A. Lesh (Ed.), *Handbook of research and design in mathematics and science education*. Mahway, NJ: Lawrence Erlbaum.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning, 6*(2), 81-89.

Confrey, J., & Stohl, V. (2004). *On evaluating curricular effectiveness: Judging the quality of K-12 mathematics evaluations*. Washington, DC: National Academy Press.

Creswell, J. W. (1994). *Research design: Qualitative and quantitative approaches*. Thousand Oaks, CA: Sage Publications.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Harcourt Brace Jovanovich College Publishers.

Cronbach, L., & Meehl, P. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

DeBarger, A. H. (2006, April). *Validities of standards-based science inquiry assessments: Integrating studies of alignment, performance, and cognition.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

DeBoer, G., Abell, C. F., Gogos, A., Regan, T., Wilson, P., & Smith, S. (2007, April). *Assessment linked to science learning goals: Probing student thinking through assessment.* Paper presented at the Annual Meeting of the National Association of Research in Science Teaching, New Orleans, LA.

DeBoer, G. E. (2005). Standard-izing test items. *Science Scope, 28*(4), 10-11.

Denzin, N. K., & Lincoln, Y. S. (1994). *Handbook of qualitative research*. Thousand Oaks, CA: Sage Publications.

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

Donmoyer, R. (2001). Paradigm talk reconsidered. In V. Richardson (Ed.), *Handbook of research on teaching* (4th. ed., pp. 174-197). Washington, DC.: American Educational Research Association.

Driver, R., Guesne, E., & Tiberghien, A. (1985). *Children's ideas in science*. Philadelphia, PA: Open University Press.

Driver, R., Squires, A., Rushworth, P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas.* London: Routledge.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215-251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* (revised edition ed.). Cambridge, MA: MIT Press.

Eryilmaz, A. (2002). Effects of Conceptual Assignments and Conceptual Change Discussion on Students' Misconceptions and Achievement Regarding Force and Motion. *Journal of Research in Science Teaching, 39*(10), 1001-1015.

Flaherty, E. G. (1974). The thinking aloud technique and problem solving. *Journal of Educational Research, 68*, 223-225.

Gogos, A., & DeBoer, G. E. (2007, April). *Assessing students' understanding of controlling variables.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Gonzales, P., Guzman, J. C., Partelow, L., Pahlke, E., Jocelyn, L., Kastberg, D., & Williams, T. (2004). *Highlights from the Trends in International Mathematics and Science Study (TIMSS) 2003*. Washington, DC: National Center for Education Statistics.

Gonzalez, V. (1999). *Language and cognitive development in second language learning: Educational implications for children and adults*. Boston, MA: Allyn and Bacon, Inc.

Gorin, J.S. (2006). Test design with cognition in mind. *Educational Measurement: Issues and Practice, 25*(4), 21-35.

Gorsky, P., & Finegold, M. (1992). Using computer simulations to restructure students' conceptions of force. *Journal of Computers in Mathematics and Science Teaching, 11*(2), 163-178.

Greeno, J. G., Collins, A. M., & Resnick, L. B. (1997). Cognition and learning. In R. Calfee (Ed.), *Handbook of educational psychology* (pp. 15-47). New York, NY: Simon & Schuster Macmillan.

Griffin, S., & Case, R. (1997). Re-thinking the primary school math curriculum: An approach based on cognitive science. *Issues in Education, 3*(1), 1-49.

Gunstone, R., & Watts, M. (1985). Force and motion. In A. Tiberghien (Ed.), *Children's ideas in science* (pp. 85-104). Philadelphia, PA: Open University Press.

Hamilton, L. S., Nussbaum, E. M., & Snow, R. E. (1997). Interview procedures for validating science assessments. *Applied Measurement in Education, 10*, 181-200.

Hammer, D. (1996). More than misconceptions: Multiple perspectives on student knowledge and reasoning, and an appropriate role for education research. *American Journal of Physics, 64*(19), 1316-1325.

Harlow, A., & Jones, A. (2004). Why students answer TIMSS science test items the way they do. *Research in Science Education, 34*, 221-238.

Harvard-Smithsonian Center for Astrophysics. (2001). *ARIES: Exploring Motion and Forces: Speed, Acceleration, and Friction.* Watertown, MA: Charlesbridge Publishing.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force Concept Inventory. *The Physics Teacher, 30*(3), 141-158.

Huitt, W. (2006). *The cognitive system*. Educational Psychology Interactive. Retrieved August 10, 2007, from the World Wide Web: http://chiron.valdosta.edu/whuitt/col/cogsys/cogsys.html

Ioannides, C., & Vosniadou, S. (2002). The changing meanings of Force. *Cognitve Science Quarterly, 2*(1), 5-62.

Kesidou, S., & Roseman, J. E. (2002). How well do middle school science programs measure up? *Journal of Research in Science Teaching, 39*(6), 522-549.

Kibble, B. (2006). Understanding forces: What's the problem? *Physics Education, 41*(3), 228-231.

Kuipers, J. C., Viechnicki, G. B., Fugate-Brunino, E., Jones, L., & Wright, L. (2006). *A video ethnography of talking and writing about motion and forces in a diverse science classroom.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*(4), 6-15.

Lynch, S. (2007, April). *A model for fidelity of implementation in a study of a science curriculum unit: Evaluation based on program theory.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Lynch, S., Kuipers, J. C., Pyke, C., & Szesze, M. (2005). Examing the effects of a highly rated science curriculum unit on diverse students: Results from a planning grant. *Journal of Research in Science Teaching, 42*(8), 912-946.

Lynch, S., O'Donnell, C. L., Merchlinsky, S., Hatchuel, E., Rethinam, V., & Watson, W. (2006, April). *The role of the "comparison group" in quasi-experimental research designs: Unexpected results from an effectiveness study of middle school curriculum units.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lynch, S., Taymans, J., Watson, W., Ochsendorf, R., Pyke, C., & Szesze, M. (2007). Effectiveness of a highly-rated science curriculum unit for students with disabilities in general education classrooms. *Exceptional Children, 73*(2), 202-223.

Marioni, C. (1989). Aspects of students' understanding in classroom settings (age 10-17): Case study on motion and forces. *Physics Education, 24*, 273-277.

McCloskey, M. (1983). Intuitive physics. *Scientific American, 248*(4), 114-122.

McDermott, L. C. (1984). Research on conceptual understanding in mechanics. *Physics Today, 37*(7), 24-32.

McDermott, L. C., Heron, P. R. L., Shaffer, P. S., & Stetzer, M. R. (2006). Improving the preparation of K-12 teachers through physics education research. *American Journal of Physics, 74*(9), 763-767.

McDermott, L. C., & Redish, E. F. (1999). PER-1: Physics education research. *American Journal of Physics, 67*, 755-767.

Messick, S. (1989). Validity. In R. Linn (Ed), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education and MacMillan.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Minstrell, J. (1982). Explaining the 'at-rest' condition of an object. *The Physics Teacher, 20*, 10-20.

Mislevy, R. J. (2006). Cognitive psychology and educational assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 257-305). Westport, CT: American Council on Education/Praeger.

Mislevy, R.J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3-67.

Montanero, M., Suero, M. I., Perez, A. L., & Pardo, P. J. (2002). Implicit theories of static interactions between two bodies. *Physics Education, 37*(4), 318-323.

National Research Council (NRC). (1996). *National Science Education Standards*. Washington, DC: National Academy Press.

National Research Council (NRC). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academy Press.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231-259.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. New York, NY: Cambridge University Press.

Ochsendorf, R., Deprey, N., Lopez, R., Roudebush, D., Lynch, S., & Pyke, C. (2003). *Using the Project 2061 Curriculum Analysis to rate a middle school science curriculum unit: ARIES-Exploring Motion and Forces. SCALE-uP Report No. 10.* Washington, DC: George Washington University, SCALE-uP.

Osborne, R., & Freyberg, P. (1985). *Learning in science: The implications of children's science*. Auckland: Heinemann.

Osborne, R., & Gilbert, J. K. (1980). A technique for exploring students' views of the world. *Physics Education, 15*, 376-379.

Palmer, D. H., & Flanagan, R. B. (1997). Readiness to change the conception that "motion-implies-force": A comparison of 12-year old and 16-year old students. *Science Education, 81*(3), 317-331.

Park, J., & Han, S. (2002). Using deductive reasoning to promote the change of students' conceptions about force and motion. *International Journal of Science Education, 24*(6), 593-609.

Pfundt, H., & Duit, R. (1994). *Students' alternative frameworks and science education* (4th ed.). Kiel, Germany: Institute for Science Education.

Phillips, D. C. (1995). The good, the bad, and the ugly: The many faces of constructivism. *Educational Researcher, 24*(7), 5-12.

Piaget, J. (1928). *The child's conception of the world*. London: Routledge and Kegan Paul.

Piaget, J. (1952). *The child's conception of number*. London: Routledge and Kegan Paul.

Piaget, J. (1972). *Understanding causality*. New York: Norton.

Piaget, J. (1983). Piaget's theory. In P. Mussen (Ed.), *Handbook of Child Psychology* (4th ed., Vol. 1). New York: Wiley.

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

Project 2061. (2006). *School report on force and motion*. Washington, DC: American Association for the Advancement of Science, Project 2061.

Pyke, C., & Hanson, A. (2005). *Standard setting process: The setting of cut scores for the Seasons and Motion and Forces Assessments. SCALE-uP Report No. 14.* Washington, D.C.: The George Washington University, SCALE-uP.

Pyke, C., Lynch, S., Kuipers, J. C., Szesze, M., & Watson, W. (2006). *Implementation study of Exploring Motion and Forces (2005-2006): SCALE-uP Report No. 8.* Washington, DC: The George Washington University, SCALE-uP.

Pyke, C., & Ochsendorf, R. (2004). *Conservation of matter assessment manual.* Washington, DC: The George Washington University, SCALE-uP.

Pyke, C., & Ochsendorf, R. (2006, April). *Concept assessment in curriculum unit evaluation.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Reiner, M., Slotta, J. D., Chi, M. T. H., & Resnick, L. B. (2000). Naïve physics reasoning: A commitment to substance-based conceptions. *Cognition and Instruction, 18*(1), 1-34.

Scott, P., Asoko, H., & Leach, J. (2007). Student conceptions and conceptual learning in science. In N. G. Lederman (Ed.), *The Handbook of Research on Science Education*. Mahwah, NJ: Lawrence Erlbaum Associates.

Sfard, A. (1998). On two metaphors for learning and the dangers of choosing just one. *Educational Researcher, 27*(2), 4-13.

Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal of Research in Mathematics Education, 26*, 114-145.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1&2), 1-98.

Stern, L., & Ahlgren, A. (2002). Analysis of students' assessments in middle school curriculum materials: Aiming precisely at benchmarks and standards. *Journal of Research in Science Teaching, 39*(9), 889-910.

Strauss, A., & Corbin, J. (1990). *Basics of qualitative research: Grounded theory procedures and techniques*. Newbury Park, CA: Sage Publications.

Strike, K., & Posner, G. (1992). A revisionist theory of conceptual change. In R. J. Hamilton (Ed.), *Philosophy of science, cognitive psychology, and educational theory and practice* (pp. 147-176). Albany, NY: SUNY Press.

U.S. Congress. (2001). No Child Left Behind Act of 2001. Public Law 107-110, 107th Congress. Washington, DC: Government Printing Office.

Vosniadou, S., Ioannides, C., Dimitrakopoulou, A., & Papademetriou, E. (2001). Designing learning environments to promote conceptual change in science. *Learning and Instruction, 11*, 381-419.

Watson, W., Pyke, C., Lynch, S., & Ochsendorf, R. (2007, April). *Understanding the effectiveness of curriculum materials through replication.* Paper presented at the Annual Meeting of the National Association for Research in Science Teaching, New Orleans, LA.

Watson, W., Pyke, C., Lynch, S., & Ochsendorf, R. (submitted for publication). Three quasi-experiments that informed the decision to scale-up a middle school motion and forces unit: Strategic replication in evidence-based research.

Watts, D. M. (1982a). Gravity--Don't take it for granted! *Physics Education, 17*, 116-121.

Watts, D. M. (1982b). A study of schoolchildrens' alternative frameworks of the concept of force. *European Journal of Science Education, 5*(2), 217-230.

Watts, D. M., & Zylbersztajn, A. (1981). A survey of some children's ideas about force. *Physics Education, 16*, 360-365.

Wegner, D. M. (2002). *The illusion of conscious will*. Cambridge, MA: Bradford Books/MIT Press.

Weller, H. G. (1995). Diagnosing and altering three Aristotelian alternative conceptions in dynamics: Microcomputer simulations of scientific models. *Journal of Research in Science Teaching, 32*(3), 271-290.