UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

LEARNING PHYSICS CONCEPTS AS A FUNCTION OF COLLOQUIAL LANGUAGE USAGE

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILISOPHY

By

STEVEN J. MAIER
Norman, Oklahoma
2008

LEARNING PHYSICS CONCEPTS AS A FUNCTION OF COLLOQUIAL LANGUAGE USAGE

A DISSERTATION APPROVED FOR THE
DEPARTMENT OF INSTRUCTIONAL LEADERSHIP AND ACADEMIC CURRICULUM

BY

Edmund A. Marek, Chair

Jon E. Pedersen

John J. Chiodo

H. Michael Crowson

Bruce J. Ackerson

© Copyright by STEVEN J. MAIER 2008
All Rights Reserved.

ACKNOWLEDGEMENTS

I would first like to acknowledge my committee members, Edmund Marek, Jon Pedersen, John Chiodo, Mike Crowson and Bruce Ackerson. Collectively, their expectations, expertise, and counsel have made me not only a better researcher, but a better person. I am fortunate to have been able to work with a group of individuals so committed to producing graduates adept at quality and meaningful research. In particular, Edmund Marek's steadfast support as an advisor, colleague, and a friend has had a significant positive role in my development as a professional.

Without the love and support of my parents over the years, I would not have had the ambition or resolve to achieve lofty goals. I only hope I can return the favor toward my own children. Acknowledgements are due to my siblings and extended family. Their gentle prodding and care over the years has led me to where I am today and has helped me remain focused on what is most important in life.

To my colleagues at Northwestern Oklahoma State University, I give many thanks. The lasting friendships and camaraderie are what has enabled me to maintain a healthy balance between work, play, and study.

Finally, I need to extend my gratitude to my wife and children. Amber, for whom I have great respect for, has demonstrated the most genuine understanding, patience and encouragement during all of my graduate study endeavors. Our love for one another has supported me more than she may recognize. And of course, my children deserve thanks also. Their resilience, inventiveness, and unfading love over the years have helped me remain motivated while simultaneously providing purpose for pursuing inquiries in science education research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITION OF TERMS

Abstracting: The act of linking the meaning of one concept to other more general descriptors via specific terminology. For example, a child may make think of morning milk when presented an image of a cow, a farmer may intuitively think of assets.

Accommodation: The process of forming a new understanding or conception after having recognized that one's knowledge base does not account for novel or anomalous observations and experiences.

Alternative Conception: An understanding of a concept one has that is different from the accurate or accepted meaning. This term is generally preferred over misconception because it focuses less on learners being incorrect and more on the merits of already existing conceptions.

Assimilation: The cognitive act of accounting for observations and experiences based on what is already known.

Blending: The cognitive act of merging multiple mental models or understandings.

Cognitive Conflict or Cognitive Dissonance: A disequilibrium on the part of the learner when observations and/or experiences are in opposition with what is already known and understood.

Colloquial Language Usage: Everyday usage of terminology, typical of casual conversation outside the context of a classroom.

Concept Development: The process of forming a new understanding or conception.

Conceptual Change: The process of changing one's existing understanding or conception to account for new observations or experiences.

Developmental Stage: The cognitive state of an individual reflected by his or her quality of thought. The four stages of development described by Piaget are: sensorimotor, preoperations, concrete operations and formal operations.

FCI Gain: In raw form, this is a numerical value ranging from -1.00 to 1.00. Pre- and post-test values of individuals are required to determine FCI gain. This value may be interpreted to signify the percent change from the pre-test performance to the post-test performance. FCI gain is determined using the following algorithm:

$$G = \frac{postscore\% - prescore\%}{100 - prescore\%}$$

Disequilibrium: A mental state on the part of the learner initiated by not being able to account for observations and experiences based upon what is already known.

Formal Learning Environment: An instructional setting where the content and terminology of the material is technical and structured; most typically a classroom setting.

Mental Content: The culmination of accommodated conceptions one uses to assimilate observations and experiences.

Mental Functioning: A general term referring to thought processes: assimilation to disequilibrium then to accommodation and lastly organization.

Mental Model: A collection of accommodated conceptions one uses to assimilate related observations and experiences. A mental model could be described as a subset of one's mental content.

Misconception: An understanding of a concept one has that is different from the true or accepted meaning.

Natural Language Instrument: An instrument assessing conceptual knowledge designed to be as independent of technical terminology as possible. The Force Concept Inventory is an example of a natural language instrument.

Newtonian Mechanics: The vector-based portion of physics related to kinematics and dynamics; the description of the motion of objects following the application of an applied force and the description systems under the influence of applied forces, respectively. This domain of physics is most often studied by first semester introductory physics students.

Organization: The cognitive act of ordering newly accommodated concepts with other already accommodated concepts within one's mental content. This results in the formation of new mental structures.

Semantics: The meaning of terminology as a function of time, perspective and context of environment.

Technical Terminology: Terminology of a specific and unique meaning in a particular discipline.

Vernacular: A vocabulary in a discipline having specific meanings not necessarily the same as the meanings of the same words in everyday usage.

ABSTRACT

Data from two sections of college introductory, algebra-based physics courses ($n_1 = 139$, $n_2 = 91$) were collected using three separate instruments to investigate the relationships between reasoning ability, conceptual gain and colloquial language usage. To obtain a measure of reasoning ability, Lawson's Classroom Test of Scientific Reasoning Ability (TSR) was administered once near mid-term for each sample. The Force Concept Inventory (FCI) was administered at the beginning and at the end of the term for pre- and post-test measures. Pre- and post-test data from the Mechanics Language Usage instrument were also collected in conjunction with FCI data collection at the beginning and end of the term. The MLU was developed specifically for this study prior to data collection, and results of a pilot test to establish validity and reliability are reported.

T-tests were performed on the data collected to compare the means from each sample. In addition, correlations among the measures were investigated between the samples separately and combined. Results from these investigations served as justification for combining the samples into a single sample of 230 for performing further statistical analyses.

The primary objective of this study was to determine if scientific reasoning ability (a function of developmental stage) and conceptual gains in Newtonian mechanics predict students' usages of "force" as measured by the MLU. Regression analyses were performed to evaluate these mediated relationships among TSR and FCI performance as a predictor of MLU performance. Statistically significant correlations and relationships existed among several of the measures, which are discussed at length in the body of the narrative.

The findings of this research are that although there exists a discernable relationship between reasoning ability and conceptual change, more work needs to be done to establish improved quantitative measures of the role language usage has in developing understandings of course content.

CHAPTER 1

Introduction

Precursor to Study

The premise of this study is that there are two required mechanisms all individuals utilize as they develop understandings of concepts. Firstly, experiences guide the level of understanding an individual has of a concept. This is the case whether concepts are being refined from already existing conceptions or for entirely newly-developed concepts following an experience. These experiences may consist of attending an air show, comparing minerals and rocks in a personal collection, reading a passage in a book, or in involvement in a structured classroom activity, regardless of its format. Secondly, serving as the greater emphasis of this study, is the role language usage plays in learning. Daily spoken communication, communication between teacher and student, communication among students and finally, the usage of language in classroom materials each provide the means for an exchange of ideas. Of importance here is that each of these forms of communication are dependent upon the learner's working definitions of the terminology used.

The interplay between student language usage of technical terminology, how this language usage changes following instruction, students' reasoning abilities and the simultaneous development of concepts is the mainstay of this study. Therefore, the Background of the Study and the Literature Review that follow will provide relevant empirical, theoretical and philosophical evidence demonstrating how language usage, its change over time, reasoning ability and conceptual development are coupled to one another.

Background of Study

Over the past several decades, physics education research revealed much about the extent and depth of students' understandings of physics concepts prior to and following classroom instruction (McDermott and Redish, 1999). Measurement of how students' conceptual understandings change following instruction necessitated the development of numerous quantitative and qualitative assessment instruments (Maloney et. al., 2004). The implications and results of such endeavors directly impact how physics material is presented to physics students of various programs of study ranging from pre-service elementary education majors to calculus-based first year physics majors. A few examples of end-product curricula, specific to physics instruction, include Explorations in Physics (Jackson et. al., 2003), Inquiry Physics (Meador, 2001), Investigations in Natural Science: Physics (Renner et. al., 1985), Just in Time Teaching (Novak et. al., 1999), Minds on Physics (Leonard et. al., 2000), Models in Physics Instruction (Wells et. al., 1995), Physics by Inquiry (McDermott, 1996), Physics for Elementary Teachers (Goldberg et. al., 2006), Tutorials in Introductory Physics (McDermott et. al., 2002), and Workshop Physics (Laws, 1997). Whether based upon empirical evidence, learning theory or both, the designs of these curricula share a common goal: to promote an improved accuracy and quality of learned physics content for beginning physics students.

Many sound research-based curricula and newly-developed materials for instruction are explicitly directed strategies designed to address trends revealed by the analysis of empirical results of students' responses to conceptual questions. The motivation for these new curricula and instructional media arose out of students' poor performance on fundamental conceptual questions following instruction (Hestenes, et. al., 1992 and McDermott & Redish, 1999).

Interpreting students' written and oral responses to qualitative instruments requires that attention be paid to student usage of terminology. Likewise, the terminology selected to generate questions for an instrument designed to assess students' conceptual understandings needs to be done with great care. Whether one is using an instrument devoid of technical terminology to assess conceptual understandings of physics concepts or an instrument assessing rote knowledge of technical terminology describing physics concepts, the words read by the students carry an array of colloquial meanings. It is this dependence on language that has spurred the interest and guided the direction of this study.

Despite the context of language usage in assessment instruments, students access multiple working vocabularies as they interpret questions and select responses. Itza-Ortiz et. al. (2003), Hart (2002), and Clerk & Rutherford (2000), demonstrated that the use of multiple vocabularies by instructors and students leads to misconceptions in physics in addition to *evaluator* misinterpreting the results of conceptual assessment instruments. This suggests that discrepancies may arise in the assessment of understandings of concepts in part because the meanings of terminology used in a formal learning environment are quite different from usage in informal colloquial contexts. Although not yet formally acknowledged in physics education literature at the time, this was most likely assumed to be the case for the developers of the natural language instrument, titled the Force Concept Inventory (FCI) (Hestenes, et. al., 1992). Items and responses on the FCI are framed so that those completing the instrument are not required to know rote definitions of terms; it is strictly a conceptual inventory that does not depend on distinguishing technical terminology.

In formal learning environments, the usage of terminology is technical and precise. In the case of physics instruction, overlapping colloquial and technical usage of terminology is common and problematic. Ascribing "force" as synonymous with "energy" and "momentum," for example, reflects a lack of differentiation among fundamentally different concepts. As a solution to typical outcomes of this kind of error, some physics curricula were developed to improve problem solving abilities. Other physics curricula concentrated on deepening conceptual understandings. Only recently has newly published physics curricula and teaching strategies reflected the importance of correct technical language usage and the adoption of technical terminology *after* students develop concepts. Such courses and strategies range in level from those completed by elementary school education students (Marek & Cavallo, 1997 and Goldberg et. al. 2006) to those enrolled as physics majors at the undergraduate level (McDermott & Shaffer, 2002 and Maier & Marek, 2005). Although it is left as a generally understood presupposition that the use of language for communicating ideas accurately is critical, little research exists reporting potential long-lasting effects of students' original colloquial usage of technical terminology (Lemke, 1990).

Evident from adopted introductory textbooks is that usage of technical terminology isn't consistent among qualified textbook authors. In a review and analysis of usage of language in physics texts, Brookes (2006) demonstrated varying contextual and grammatical uses of the terms "heat" and "force." Using a term, as Brookes reports, as a noun and as a verb within the same chapter of a text implies varying roles and properties of the concept the term identifies. Brookes and Etkina (2007) present similar findings regarding quantum mechanics paying special attention to the impact of varied language usage within a text's narrative on students' understandings.

Germane to this study are considerations of how these works and developments impact students' usage of technical terminology following instruction. Specifically, this study will involve the collection of empirical data to determine if final usage of the term "force" is indicative of conceptual change, reasoning ability, neither or both. Implementing a full scale study that compares the degree of change of language usage among students across multiple instructional settings is problematic. This is also the case of a comprehensive tracking of vocabulary usage in the classroom. Answering questions such as what was meant by what was said, did students mean force when they said energy, and did students correctly use the terminology of energy while really thinking force, for an entire class over the span of an academic year would present difficult research challenges (Lemke, 1990). The intent of this study is to investigate the degree to which colloquial language usage persists; namely usage of "force" synonymously with other terminology. Further, this study will investigate the nature of the relationship between students' changes in language usage and their reasoning abilities. It may be that students' vocabularies prior to taking physics and their reasoning abilities do not exhibit a correlation. However, as shown by Coletta and Phillips (2005), students' reasoning abilities should be correlated to their conceptual gain in Newtonian physics. As will be discussed in greater detail in the literature review, the ability to accommodate new concepts and correctly associate technical terminology long-term should also be correlated to reasoning ability. Such relationships should be reflected in correlations between gains on a conceptual natural language instrument and accurate long-term adoptions of new technical terminology in language usage measured by another instrument.

Purpose of Study

The purpose of this study is to explore one facet of students' colloquial language usage and its effect on their learning physics concepts related to force. This study addresses the following questions:

- How do trends in students' colloquial usages of terminology related to "force" prior to and following course instruction compare to conceptual gains of the same concept?
- What relationships, if any, exist between colloquial usages of "force" and students' scientific reasoning abilities?
- Are students' reasoning abilities and conceptual gain, as measured by gains on a test of conceptual understanding of force, mutually exclusive predictors of measured changes in colloquial usages of "force"?
- Does lack of conceptual change, as measured by gains on a test of conceptual understanding of force, necessarily preclude a positive change in colloquial language usage?
- And in contrast, is significant conceptual change, as measured by gains on a test of conceptual understanding of force, necessarily mirrored by positive changes in colloquial language usage?

A tool for the systematic and quantitative measurement of students' conceptual gains in Newtonian mechanics for introductory physics is well-established. The Force Concept Inventory (FCI) is a commonly used instrument in physics education research, serving as the source of extensive research data (Hestenes & Wells, 1992). As a result, significant pools of data exist, as well as insightful published works of physics education research groups. Also well established in the science education research community is Lawson's Classroom Test of Scientific Reasoning Ability (TSR). This instrument evaluates learners' levels of mental functioning in an array of science classrooms (Lawson, 1978 and 2008). Performance on the TSR is correlated to and therefore serves as a measure of participants' Piagetian developmental stages. Coletta & Phillips reported that gains on the FCI are positively correlated to performance on the TSR (2005).

Students' conceptual gains measured by pre- and post-test results of the FCI will be collected along with student performance on the TSR. A third instrument, the Mechanics Language Usage instrument (MLU), was designed for this study to fulfill the need for an instrument tracking terminology usage. The purpose of employing all three of these instruments is to determine if changes in language usage is predicted by conceptual gains as measured by FCI pre- and post-testing. An additional component of the study is to determine whether final language usage as measured by MLU post-test, can be predicted by performance on the TSR.

Significance of Study

This study's findings should produce direct evidence either in support of or in opposition to the argument that correct usage of terminology on the part of the student is requisite for accurate accommodations of concepts. For example, one could argue that high gains on the FCI should not be possible for students reflecting little to no change in their language usage. Results to the contrary would indicate that conceptual development occurs independently of changes in language usage. Something could also be said of low FCI gains and high positive MLU changes. Just because terminology is used correctly does not necessarily imply that students have an accurate understandings of the concepts.

For this investigation, the Mechanics Language Usage instrument was developed to measure changes in students' usage of "force." Prior to this study, no instrument was available in the research literature to obtain these measures. Subsequent versions of the MLU could be created to measure the change in usage of other terminology. The format and structure of the MLU represents a new type of instrument that will be applicable in other content areas within and beyond physics.

Problem Statement

The persistence of colloquial usage of the term "force" following course instruction may take on different forms. Despite completing coursework, students may still equate "force" with any number of other terms without distinction, or change their usage more or less accurately. And, while practitioners and researchers agree correct usage of technical terminology is important, there does not currently exist a paper and pencil instrument that gauges whether or not students' usage of the term "force" changes over time. Consequently, there is little quantitative evidence available to test the relationships between initial and final language usage with conceptual gains and reasoning ability. This study addresses each of these vacancies in the science education research community.

Research Questions

This study is designed to investigate the following research questions:

1. Are there significant changes in students' colloquial usage of the term "force" following instruction?
2. What are the relationships among colloquial language usage of "force," scientific reasoning ability and conceptual change for students in a traditionally taught introductory college physics course?

CHAPTER 2

Literature Review

Context of Study's Foundations

This study is an investigation of introductory physics students' language usage as they learn concepts in Newtonian mechanics. Because language usage permeates thought and communication prior to, during and following the development of concepts, boundaries between learners' language usage of terminology in the "technical sense" in the context of formal physics instruction from the "colloquial sense" (everyday usage) are inherently blurred. However, the effects of merging these usages can be investigated. Learners of varied experiences and reasoning abilities delineate language usage differently and arrive at different levels of understanding of physics concepts. Therefore, the variables for this study span multiple disciplines. Consequently, this literature review presents work pertinent to this study from educational theory, language and semantics, ontology, and physics education research.

Language Usage in Context of Piagetian Theory

In the context of Piaget's model of mental functioning, there are specific processes a learner engages in while developing a new concept (Piaget, 1975; Renner & Marek, 1990; Phillips, 1975). Initially, learners assimilate a new experience or observation with what they already know. Confusion on the part of the learner arises during assimilation if he or she is unable to account for what is observed. Piaget termed this state disequilibrium; other equitable terms are cognitive conflict and cognitive dissonance. In order to re-equilibrate, the learner must somehow account for the new experience (Maier & Marek, 2005). For the learner, this is only possible through the development of a new concept, meaning accommodation has taken place. Once concept accommodation occurs, it becomes a piece of the greater whole of the learner's understanding and is organized as part of his or her mental content and new mental structures are formed. As a result of these processes, assimilation and the accuracy of accommodated concepts directly affect the learner's future processes in mental functioning.

Evident from earlier work is the relevance of correct language usage during instruction and in textbooks (Crouch, et. al., 2001). In the physics community the question of language usage in the past has been categorized as a rhetorical case of "semantics" juxtaposed with other subjects of merit that educators should be more especially attuned, but not necessarily presented as material for researchers to pursue (Williams, 1999; Touger, 1991; Touger 2000; Styer, 2001). A familiar example of this among physics instructors is the student usage of the term "mass" versus "weight" when reporting the mass of an object. While physicists agree mass and weight are not the same (weight is a force vector and mass is a scalar physical property of an object), there is no agreement regarding the level of technical accuracy to require of students prior to developing the concept named force. Subsequently, there is often confusion when students are later presented problems requiring distinctions to be made to adequately describe a system.

As a hypothetical example from the sciences, consider a teacher guiding students toward a model of the scientific method and the implications of what a “scientific theory” means within the scientific community. As examples throughout the duration of the course, the class studies many scientific theories and hypotheses including continental drift hypothesis, plate tectonic theory, big bang theory and the nebular hypothesis. Now imagine the teacher attempting the same feat with students who regularly use the terms “theory”, “hypothesis,” “inference” and “guess” interchangeably in everyday language. Following instruction, how well will these students be able to distinguish among the consequences and meanings of the above scientific developments after using such terms as equivalences for several years? Even if students distinguish these terms from one another successfully on unit tests, how accurately will students associate appropriate terminology to concepts on a long-term basis? To what extent will the old usages of the terms persist? Of what consequence will this hold for these students’ future studies and perceptions of science? Despite best teaching practices, colloquial language usage could still present a significant and potentially lasting obstacle.

Such is the case for introductory physics, especially so with Newtonian mechanics. Because of the cumulative nature of coursework, the challenge of the teacher often becomes balancing students' developments of conceptual understandings with acceptable colloquial terminology until further delineation and specificity in usage of terminology is required. Using a previous example, a teacher concedes to student usage of weight and mass interchangeably until the concept named force is developed.

Communication and Development of Concepts as More than Just "Words"

In Hayakawa's *Language and Thought in Action* (1940), he offers several viable points regarding the role of language, in developing conceptual understandings, and in communicating ideas. Although Hayakawa's arguments in this text are primarily to shed light on human semantic responses to politics, governmental policy, and interactions among each other, there are several underlying premises that warrant exploration for this investigation. Hayakawa's attention to how individuals "choose" to interpret or assign terminology is of particular interest. The pooling of knowledge, the relationship between classification and abstracting, and the "one word, one meaning" fallacy are also of interest and will be addressed.

Hayakawa contends that all of what we know, we learn through the use of language, spoken and written. Early in development, we learn to interpret noises as conveyers of meaning that represent "things" experienced. At the outset, this is similar to some of what we could learn from Piagetian theory. Namely, that learning is experience-based. However, Hayakawa's emphasis is on language usage and development, not on conceptual development. His argument is that only after significant development do learners differentiate language as *representations* of things from language as *the meaning* of things. Regardless of age, what individuals ultimately know consists of a pooling of knowledge from the incorporation of what is read, heard and experienced. This pooling of knowledge is dependent on one's own working definitions of words, since language is the fundamental common theme that permeates all of what is "known." This is analogous to what a physics instructor would contend: a holistic understanding of physics is cumulative in nature; each succeeding conceptual understanding dependent on previously developed concepts.

Based on varied experiences and contexts of the usage of the term "force," each person arrives at his or her own set of rules governing proper use of the term, resulting variants of others' sets of rules. Up to this point, one could argue that the solution to the problem of people understanding force to mean different things is to simply and collectively proclaim an absolute definition of force to replace any alternatives. Such an effort would ignore any long held conceptions and associations students ascribe to "force," however accurate or inaccurate they may be. In the context of Piagetian theory, "force" is already woven in to one's mental content and mental structures; and in turn used to further assimilate other experiences.

Moreover, Hayakawa contends that in the context of written and spoken language, the belief that one word carries only one true meaning is a fallacy. The meaning of a word, even if singular upon its origin, changes over time. Science historians could easily provide evidence for this within the discipline of physics itself. As a brief example, the meanings of work and energy, as understood by physicists today, involves a rich history that is heavily rooted in cultural influences in experimentation and usages of the term as well as religious interpretations of its meaning and origin (Smith, 1998). Even today, there is significant evidence that the usage of the particular terms among physicists, and in formal contexts, is varied and problematic at best (Hilborn, 2000, 2003; Mendelsen, 2003a, 2003b; Bauman, 1992a, 1992b; Mclldowie, 1995).

Communication and Development of Concepts as More than Just "Semantics"

Hayakawa acknowledges that because of different backgrounds, people classify their experiences differently through the selection and use of words. Similarly, Piaget's model of mental functioning describes the "end" process of concept development as the organization of the concept within one's own mental content and mental structures, which is a function of life experiences and prior knowledge. However, these processes are not truly end processes, as one's mental content and mental structures are the basis for future assimilation (Piagetian) and classification (Hayakawan). The mechanism Hayakawa proposes for how individuals gain an understanding of their environment is through a process called "abstracting." This process is an adaptation of Korzybski's structural differential from general semantics (1933). Although a thorough examination of semantics falls out of the scope of this study, the value in exploring some poignant examples provides insight regarding the merit of studying colloquial usage of technical terminology by students.

For example, an image of Newton's Cradle is presented to a physicist (also known as a "momentum demonstrator"). One of the end steel spheres is raised and released. Upon impact with the remaining four motionless spheres of equal mass, the farthest sphere is ejected with the same speed away from the others. The breakdown of the physicist's abstraction is as follows:

1) collision of objects → momentum of objects will change
2) changing momentum in the absence of external forces → conservation of momentum
3) conservation of momentum for rigid objects → elastic collision
4) elastic collision → conservation of kinetic energy

The understanding of the concepts for an experienced physicist may result in the above abstractions occurring in very short order—without a conscious recognition of each individual abstraction. This is because through study and experiences, each level of abstraction leads to categorizations that fit within his or her body of knowledge.

In contrast, for two physics students who complete studies of mechanics, the abstraction breakdowns might consist of:

Student 1

1) collision of objects → model system
2) model system → no loss of energy
3) no loss of energy → energy is conserved

Student 2

1) collision of objects → closed system
2) closed system → no external forces
3) no external forces → internal forces only
4) only internal forces → Newton's third law of motion
5) Newton's third law of motion → force of impact ball = force of ejected ball

There are several reasons this example is appropriate. For Student 1, the final outcome will probably lead this student toward correctly solving certain problems but it does not reflect the true nature of what is involved in other aspects of conservation of momentum. Applying the same abstractions to an inelastic collision would expose this. For Student 2 the system holds a different meaning, leading toward an over simplified statement of one of Newton's laws of motion. To be clear, all three abstractions are different, yet they all were initiated by the same image. And, like the physicist, the students' abstractions may occur without the students being consciously aware of the steps of their abstractions.

Now consider the same scenario applied to college students yet to take physics courses. These students, like any anyone else with an established vocabulary, already categorize "force," "momentum," "energy," and "collision" among other terms in their body of knowledge because this vocabulary used colloquially. The question that arises is what kinds of abstractions do these students make upon reading and/or hearing these terms? Also, how lasting are these abstractions following classroom instruction and experiences in physics class? Thus, semantics plays a role much more important than "semantics" in the punitive sense.

The irony in the titles "Communication and Development of Concepts as More than Just 'Words'" and "Communication and Development of Concepts as More than Just 'Sematics'" are in the downplay of the terms in quotes, if read with colloquial interpretations. Often, less attention is paid to the importance of the "words" students use in their explanations while differences among terminologies defining concepts are brushed away as just "semantics." This is akin to brushing aside a scientific argument because it is based on just a "theory." As it turns out, words are more than "words" and semantics are more than "semantics" when interpreting new stimuli with already existing mental structures.

Conceptual Change

Up to this point in the narrative, emphasis has been placed on the usage of colloquial language and technical terminology while "conceptual change" has been referred to only occasionally (most notably in the Purpose of Study and in one of the Research Questions). This is not to imply conceptual change should be considered secondary. To the contrary, conceptual change is ultimately the goal of instruction and the mechanism for developing new understandings. Described below are a few of the more relevant works linking conceptual change and language usage.

In 1982, Posner et. al. elucidated a theory of conceptual change coupling disequilibrium and accommodation from Piagetian theory with historical accounts of scientists' changing ideas of nature. Conceptual change, they argued, is initiated when one's current mental model of a system is in direct contrast with observations or new experiences. Conceptual change occurs when one's mental model undergoes a change so that what remains is an understanding that accounts for what was already known in addition to new observations or experiences. In other words, rather than working from a lack of a conception to developing a new concept, Posner et. al.'s theory addressed the mental processes involved when already existing understandings are challenged and undergo change. It is this description that distinguishes conceptual change from conceptual development.

Other factors such as learning environment, communication (instructor and narrative in texts), social background and affective components were also recognized as contributors to conceptual change (Strike & Posner, 1992; Pintrich, et. al., 1993; Hammer, 1996; and Greca & Moreira, 2002). A full description of these factors and the results of research stemming from the investigation of them falls out of the scope of this study. However, it is important to note that much of the research literature citing FCI data and results has as its purpose, comparisons of modes of instruction (comparing two or more different learning environments). Also of importance is recognizing that varying degrees of conceptual change occur among different individuals. Vosniadou (1994) presents a strong case for the degree of conceptual change being a function of prior knowledge and experiences. The difference between Vosniadou's work from the earlier work of Posner et. al., Strike & Posner and Pintrich et. al. is that Vosniadou categorized conceptual change into levels ranging from trivial to fundamental. Vosniadou also described how the degree of conceptual change may act as the source of misconceptions. For instance, by simply adding a simple modification to a pre-existing naïve model of Newtonian mechanics, a learner may set the stage for interpretive errors in future thought by not fully changing their original model.

An alternative perspective is known as the "knowledge as pieces" model. In this theory of knowledge, phenomenological primitives (p-prims) make up pieces of knowledge that are weakly bound together (diSessa, 1993). Each p-prim in physics, for example, is based upon one's intuitive knowledge of physics. Collectively, p-prims establish one's knowledge base and serve as the network from which one draws understanding. According to diSessa, the conventional view of misconceptions inaccurately assumes that learners have stable, incorrect cognitive structures. Instead, diSessa argues that "misconceptions" are simply instantaneous knowledge states of one's emergent knowledge. This theory of knowledge suggests that one's knowledge is based entirely upon experience and is in a continuous state of flux.

Regardless of the mechanism for change in one's understanding of physics concepts, our current models for conceptual change emphasize experience and communication (either written, through social interaction or both). The emphasis is placed on the role of experience and communication toward making changes or furthering one's knowledge state. It may be that researchers never arrive at a finite, robust and universal definition of conceptual change. This study has embedded the assumption that although attempts at precise and direct measures of conceptual change are idealistic, changes in responses to items on a conceptual assessment instrument can serve as an indicator that conceptual change has occurred.

Furthermore, it is assumed that conceptual change that occurs via traditional lecture requires communication using technical terminology and that this terminology is composed of vocabulary that is a subset of language already known to the learners. What follows next is evidence that language usage is a facilitator of conceptual change, playing an integral role in how learners build mental models.

Ontology and Blending

One's ontology is, in its truest sense, how an individual specifies a concept to him/herself in thought. This specification cannot fully be described in words to another individual that would lead to an exact replicated understanding. A tenet of this study is the idea that adopted usages of terminology directly affects future learning. While learners' prior experiences are factors effecting concept development, usage of language most often is an unavoidable conduit through which ideas and understandings are expressed. Therefore, language is an integral component during the learning of new concepts. Philosopher W. V. Quine addressed the intrinsic interplay between one's usage of language and ontology concluding that language usage establishes one's ontological understandings (1984). Despite this straightforward link between language usage and concept development, little quantitative work exists to date documenting its significance (Brookes, 2006).

Collectively, the organization of one's mental structures constitutes his/her mental content. And, in turn, one's mental content then acts as the primary source individuals draw from as they make sense of new experiences. Although concepts developed by individuals aren't necessarily stored as strings of words, specific terminologies are associated with understandings. Therefore, mental content is arguably a function of one's own language usage over time, whether spoken or unspoken.

Consider the following as an example to the contrary: Obviously, a textbook for a course that a student has completed does not constitute the mental content of that student. So what does a textbook constitute? Can a textbook or any written document truly and completely represent one's mental content? Can one's understanding of concepts be fully understood by another individual? These are questions of ontology. Ordinarily of primary concern to philosophy, ontology will be broached briefly in this study as a means of providing a framework for demonstrating the development of conceptions' dependencies on language usage. For the purposes of this study, a discussion of ontology is also valuable for considering conceptual change due to blending understandings.

In his text, *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*, Mark Johnson (1987) implicitly ties much of the content presented in the sections above to ontology. He argues that conceptual imageries and phraseologies are tied, but not necessarily bound indefinitely, to those of other concepts (Johnson, 1987). Such phraseologies are composed of associated words and statements that individuals come to use to frame their understandings of concepts.

Similar to disequilibrium caused by a discrepant event in Piagetian theory, ontologies are challenged regularly via new data and new experiences. There is evidence that a single reading impacts individuals' understandings of concepts. In an investigation of over 80 high school ninth graders, students' conceptual change was measured following a reading assignment (Palmer, 2003). The intent of this study was to compare conceptual change among students reading didactic passages to students reading passages that were refutational in nature. The subject material of the passages were the same and consisted of content related to commonly held misconceptions in biology. As might be expected, the degree of impact varied across the sample (Palmer, 2003). deLeeuw & Chi provide evidence that the thought processes of students while reading challenging text is facilitated by a process they term self-explanation (2003). In this process, students use language that is familiar to them in order to rephrase and explain the narrative of the text. Of importance to this study is their conclusion that learners engage in self-explanation to refine their conception of the content presented in the narrative, not as a process to discern the text. The process deLeeuw & Chi outline is therefore an interpretive one, heavily dependent on the meaning of the terminology learners use to help them formulate conceptions. This use of one's already existing mental content to interpret and verbalize internally new data for further understanding resides within one's ontology.

It is at this point where the distinction between Piaget's mental content and ontology becomes better defined. According to Piaget, changes to one's mental content take place as a function of experiences *and level of thought* which is a function of the learner's stage of development. Changes to ontology occur via blending of what is already "known" to be true by the learner and interpretations of the new experiences. This requires a working vocabulary and the use of imagery, creating conceptions that the learner cannot fully express to another individual (Lakoff & Johnson, 2003).

Recent work completed by Podolefsky and Finkelstein (2007) reference blending as an explanation for students' development and modification of previously held conceptions (2007). By presenting alternative analogies or models of physical systems to students, their understandings changed to account for what was presented within what they already knew. In many cases, previous understandings underwent significant change. In their work, blending is presented as a strategy for helping students challenge what they know with new observations to promote improved conceptual understandings. This is similar in purpose to the learning cycle curricula for secondary school science programs developed originally in the 1980's (Renner et. al., 1985). While Podolefsky and Finkelstein's blending emphasizes characterizing and challenging mental models of understanding, learning cycle curricula has grown from attention to students' stages of intellectual development. These are just two examples of how science educators and science education programs are recognizing the importance for teachers to be cognizant of students' prior knowledge and intellectual ability.

Common Ground

As different as the premises among the works of Piaget, Hayakawa and Korzybski, and Lakoff and Johnson may seem, research in physics education and science education reveal underlying commonalities. At the very least, the implications for student learning that stems from related research demonstrate that learners' conceptions are influenced by their experiences. Further comparison traces a path that ultimately leads back to students developing concepts by challenging what they already know with new experiences. Whether one contends this occurs via learners developing new semantics in language usage, new representations of imagery and phraseology, or newly-developed blended models of concepts, a likeness exists in the outcome.

The most practical manner to bring this section to a close in a way that also brings it full circle with the intent of this study is to consider the developmental stages of students taking introductory college physics. By the time students are typically enrolled in introductory physics, they are most notably in transition between concrete operations and formal operations (McKinnon & Renner, 1971). An interesting dynamic results. While students are beginning to apply their reasoning abilities to more complicated systems, there still remains a dependence on being able to see or physically manipulate a system to accommodate a concept (Renner & Lawson, 1973).

Within the framework of Piaget's theory of intellectual development and the nature of this study, the concrete operational stage and the period of transition to formal operational stage lend to the susceptibility of misaligning technical terminology with its colloquial usage. The argument is as follows: if students are concrete or transitional learners, they will be less able to delineate their existing colloquial language usage with the terminology as specified in the classroom. In other words, concrete learners will more likely hold colloquial usage of technical terminology synonymous with technical usage of terminology in the classroom—having a direct impact on distinguishing the concepts the terminologies define. It is the intent of this study to investigate the nature of this interval of development and attempt to determine the significance of any dependencies.

CHAPTER 3

Research Methodology

Research Design

The research design for this study was a non-randomized control group pre-test—post-test quasi-experimental design. This research design, also referred to as ex post facto research, is appropriate for circumstances where the investigator is not in direct control over the independent variable. For this study, the independent variable was time of course instruction. Three separate appraisal instruments, two of which were administered as pre- and post-tests, were used to assess conceptual change, change in language usage and reasoning ability. The reasoning ability appraisal instrument was only administered once. In the narrative that follows, greater detail is provided regarding the nature of the sample, the instruments used and the analyses performed.

Research Sample

The population for this research being sampled from is students taking introductory college level physics for non-majors. Students in this population are typically pre-professional students completing requisite coursework for programs in the health fields consisting primarily of biology and chemistry majors.

To represent this population, cooperation from two major mid-western universities was solicited. The rationale for soliciting participation from two different institutions was primarily a precautionary measure for data collection. In the event of data collection errors at one of the participation sites, a second source of data would increase the likelihood of collecting at least one complete set of data. This was especially advisable because the preparation of necessary paperwork and consent forms for Institutional Review Board (IRB) approval occurred prior to the finalization of course schedules for universities. Therefore, due to possible unexpected changes of teaching schedules, collecting data from more than one research site increased the likelihood of securing at least one full set of data.

A second precautionary reason for maintaining two collection sites was a product of the research design. Because three separate instruments were to be administered—two of which required pre- and post-test data—five separate sets of paired data per research site were required. Therefore, if an attrition rate of 50% were assumed for all of the volunteering participants over the span of an entire academic semester, the total research sample could have been reduced from 250 to fewer than 80 participants, had data only been successfully collected from one research site (250 corresponding to successful data collection from both sites with an attrition rate of 50%).

Additional research considerations supporting the solicitation of two research sites had more to do with possible exploratory analyses. If the course deliveries were similar, then cross-validation analyses could be performed to test for consistency between groups. And, if the course structures varied in notable ways (i.e. deviating instructional strategies) then between groups comparisons could be made.

Course and standardized exam performance such as college entrance exam scores, GPA, course average were not collected for this study. Demographic and attitudinal data including gender, prior coursework in physical science, confidence levels in physics and English proficiency were also not used in the analyses.

Treatment

The treatment for this study is course instruction in introductory college physics for non-majors. Because this treatment was applied via two separate instructors at two different institutions (yielding two separate samples), information summarized in this section is provided for reference. Data from both samples were combined for analyses, but distinguishable participant numbers per sample were assigned so that the data could be discerned from one another at any point in the analyses. The samples were labeled Sample 1 and Sample 2. Each sample's demographics were primarily the same and the modes of instruction were comparable. The syllabi for the courses have not been included in this study for anonymity purposes. However, general descriptions of course designs for each sample are described below.

Sample 1 students met three times per week in a large lecture hall. The initial enrollment in the course was nominally 260 students. Attendance at an additional weekly discussion hour was required. These discussion sections were led by physics graduate students. During lecture, notes were presented digitally using PowerPoint and a digital projector. Worked examples were done by hand during class, projected real time. An electronic response system was used regularly during class time and responses were recorded for a grade. Class notes, including worked examples and in-class quizzes, were provided to students electronically via the Internet. Regular chapter assignments were online using WebAssign™ (automatic grading). In addition, group problems were assigned regularly (typically worked on during discussion hour). The breakdown for the final grade for the course consisted of weekly group problems (10%), daily in-class questions (10%), weekly assignments (20%), three exams (40%) and one final cumulative exam (20%). Laboratory was not required; if completed, grading and credit hours earned were kept separate from lecture. Unfortunately, it is unknown which participants were enrolled in laboratory for this sample. For completing any part of the surveys, participants were awarded credit for one full group problem—participation was not required and could end at any time without penalty.

Sample 2 also met in a large lecture hall for lecture three times per week. The initial enrollment in the course was nominally 280 students. Daily lecture notes were presented digitally using PowerPoint and were provided to students electronically via the Internet. Students were required to attend laboratory each week and submit worksheet laboratory reports. Required laboratory was taught and graded by graduate students, consisting of prescribed exercises verifying content from class lecture. An electronic response system was also used in Sample 2 regularly during class time and responses were recorded for a grade. Weekly assignments consisted of a combination of written assignments using the textbook and online assignments using MasteringPhysics™ (automatic grading). In addition, shorter problem sets were due at the beginning of each class. The breakdown for the final grade for the course consisted of laboratory (20%), daily in-class questions (10%), weekly assignments (30%), three exams (30%), and one cumulative exam (10%). For completing any part of the surveys, participants were awarded credit for one full written assignment—participation was not required and could end at any time without penalty.

Instruments

To investigate the nature of the relationships among change in language usage, conceptual gain and reasoning ability, three separate measures were required. Of these measures, change in language usage and conceptual gain required pre- and post-testing while reasoning ability required only one set of data. The instrument chosen as a measure of conceptual gain was the Force Concept Inventory (FCI). The instrument used to measure reasoning ability was Lawson's Classroom Test of Scientific Reasoning (TSR). The instrument used to measure change in language usage, the Mechanics Language Usage Instrument (MLU) was developed specifically for this study. Reported below is a synopsis of each instrument.

The FCI is a quantitative instrument developed to measure students' understandings of Newtonian mechanics (Hestenes, et. al., 1992a). This 30-item multiple choice instrument was designed to reveal common ideas students hold of force and motion and is well established as an indicator of students' understandings of basic Newtonian mechanics. Comparisons of pre- and post-test performances on the FCI serve as a measure of change in students' understandings of Newtonian mechanics concepts. Because the FCI is designed to isolate conceptual understanding from rote memorization of definitions, the narrative of the test questions and responses of the FCI are colloquial or conversational in nature. Therefore, students are not prevented from answering a question due to a lack of knowledge of definitions of technical terminology.

Other assessment instruments that assess conceptual knowledge of the same nature include the Mechanics Baseline Test (MBT) (Hestenes, et. al., 1992b) and the Force and Motion Conceptual Evaluation (FMCE) (Thornton & Sokoloff, 1999). However, since performance on these instruments is based upon accurate knowledge of technical terminology, the MBT and FMCE were deemed inappropriate for this study. The method of analysis of pre- and post-test FCI data commonly reported is the determination of the measure of student "gain" on the instrument. To determine gain on the FCI, the following relationship is used:

$$G = \frac{postscore\% - prescore\%}{100 - prescore\%}.$$

The argument made by users of the FCI and the reported gain is that differences among gains on the FCI indicate conceptual change on the part of the students. This gain, commonly known as normalized gain, is directly linked to extensive study in the physics education community and is well documented as a function of mode of instruction, gender, grade level (high school to college level) (Hake, 1998). This algorithm is typically used to compare classes employing different modes of instruction and has become a staple for gauging a rule of thumb index for instructors wishing to monitor their teaching effectiveness.

Unfortunately, other than arguments made from a preponderance of empirical evidence that gains on the FCI are correlated to effective interactive modes of instruction, no statistical validation data on the FCI has been made publicly available (Huffman & Heller, 1995). In response to Huffman and Heller’s publication, the authors of the FCI published a follow up article that attested to face and content validity, but did not report statistical measures of validity or reliability (Hestenes & Halloun, 1995).

However disconcerting the lack of validation data on the FCI may be, studies over the years using the instrument have shown many interesting trends. Most relevant for this study, for example, is the recent work done investigating whether

FCI gains are a function of level of cognitive development (Coletta & Phillips, 2005). Using the TSR, Coletta and Phillips demonstrated that a significant positive correlation exists between gains on the FCI post-test and reasoning ability. In fact, they report that performance on the TSR is a greater predictor of FCI post-test scores than FCI pre-test scores. Because greater gains on the FCI indicate progress toward an understanding Newtonian mechanics (which requires formal reasoning abilities), this result is consistent with learning theory—Piaget's model of mental functioning in particular. Appendix E provides information on accessing the most current version of the FCI.

The TSR is a 12-item instrument consisting of 12 leading questions, each paired with a follow-up question regarding the reasoning for the choice of the preceding leading question. To be scored as correct, the leading and corresponding follow-up question of a coupled pair must both be answered correctly. The reliability measure for the TSR is 0.78 while the correlation between test results and personal interviews is $r = 0.76$ (Lawson, 1978, 2007).

For this study scores obtained using this instrument were recorded on a continuous scale from 0 - 12 for analyses. In addition, TSR binned scores in categories reflecting concrete, transitional and formal operational learners from this instrument were used for analyses. Participants were categorized by the following means: 0-4 correct responses were categorized as concrete operations; 5-8 correct responses as transitional; and 9-12 correct responses as formal operations. The TSR is included in Appendix F and Appendix G is the TSR key.

Collectively, the FCI and TSR do not address the effects of colloquial usage of language on instrument performance. Their design was to exist as independent of technical terminology as possible. To acquire a measure of change in usage of technical terminology, the Mechanics Language Usage (MLU) instrument was developed for this study. The MLU was scored by tallying the number of instances students selected "force" *and* another term in response to a leading question. The MLU consists of five multiple choice and two free response items designed to identify alternate word associations students retain with the term "force." All of the MLU items and selectable responses of the instrument consist of material from Newtonian mechanics. The multiple choice items of the MLU were carefully created to closely correspond to selected FCI items. Multiple responses for each item are possible on the MLU and is so indicated at the end of each test item. However, no fully correct response to a single question includes the selection of "force" in addition to some another choice. Because the purpose of the MLU is to track student responses that associate superfluous terminology with force, "correctness" in responses is secondary. The MLU pre- and post-tests are Appendices H and I, respectively.

Deliberate steps were taken to make improvements upon and test for validity and reliability of the MLU. Face validity was completed using a separate sample of experienced physics students (n = 12) who completed the instrument, providing written and oral feedback on the MLU. Content validity was completed by a panel of experts consisting of a science education researcher, two physicists (one retired), two high school physics master teachers, and a retired high school science supervisor. Given details of the purpose of the instrument, each of these individuals critiqued the MLU.

Test-retest reliability was performed with a sample of psychology students (n = 39). None of these students were enrolled in any physical science course during the time interval the pre- and post-tests were administered. Therefore, the test-retest sample did not receive a treatment (physics instruction) and a time span of 13 weeks lapsed between the pre- and post-test. The correlation between the MLU pre- and post-test for this sample was $r = 0.507$, $r^2 = 0.257$. This is an indication that nearly 26% of the variance in the post-test is accounted for by pre-test performance. As a final measure of reliability, a comparison of pre- and post-test data of the test-retest sample yielded a Cronbach's alpha value of 0.673 ($\alpha = 0.673$). This is further evidence that responses participants choose on the MLU post-test will be consistent with pre-test performances without a treatment.

Procedure

MLU and FCI pre-tests were administered early in the spring 2008 semester prior to instruction on forces. Near the end of the semester (12 – 14 weeks later), these same two instruments were administered to obtain post-test measures. The TSR was administered about halfway through the semester so that testing fatigue could be reduced. Changes in performance from pre- to post-test were then determined on the MLU (using counts of instances “force” was associated with other terminology) and FCI (using normalized gain).

To reduce the chance of a bias toward selecting “force” on the MLU (literal, correct, or incorrect), the MLU pre-test was administered prior to the FCI pre-test and likewise for post-testing. The rationale for this order is that while the FCI is a natural language instrument, “force” does appear in many of the instrument items. Other terminology on the MLU (“strength,” for example) does not appear at all or in the same frequency on the FCI as “force.”

Due to time constraints and at the request of Sample 1’s instructor, the MLU was administered to students as a take-home activity. Students had effectively up to four days to complete the MLU pre-test. The same protocol was followed for the MLU post-test for this sample. The FCI and TSR were completed during lecture and discussion hours. For Sample 2 participants, all of the instruments were completed during class time. To assist with administering the surveys, detailed instructions were left for teaching assistants when it was not possible to administer them personally. See Appendices C and D for the pre- and post-test administration instructions.

The primary analysis consisted of determining if performance on the TSR and gains on the FCI are significant predictors of post-test MLU scores. To make these comparisons, linear regressions were performed. As discussed earlier in the literature review, the development and usage of language is a function of one's developmental stage. As a consequence, one's colloquial language usage serves as a mechanism for constructing and articulating mental models. Terminology that is re-defined in the physics classroom coexists with preexisting colloquial terminology. In other words, in the classroom students accommodate concepts using technical terminology that already has varying meanings for them. Therefore, it was anticipated that change on the MLU would be be directly correlated to performance on reasoning ability. It was also anticipated that despite high gains on the FCI, students would still retain some colloquial usage of terminology—effectively placing a cap on change in language usage as measured by the MLU.

CHAPTER 4

Results

Outline of Presentation of Results

The order of this chapter begins with providing instrument data from each of the collection sites. These data are presented in various forms to support combining the samples for further analysis. The analyses of the combined samples that follow can then be considered in two ways: generative and model specific (specific to the research design). Efforts were made to transition between these modes to help facilitate the greater context of the data leading to regression analyses, which are then further discussed in the conclusions and discussion chapter.

For completeness, analyses performed with data from the combined samples are presented with attention to statistical differences. Differences among reported significances that may exist if one sample or the other is excluded are reported in footnotes where appropriate. This is warranted, in light of implications stemming from between-group comparisons that are in the next section. All of the computations and results that follow were obtained by using SPSS Graduate Pack, version 16.0.1.

Between-Samples Comparisons

The necessary first step to determine the viability of combining data from each data collection site into a single sample involved performing between-group comparisons. To begin, independent samples t-tests were performed using data collected from each instrument. The results of these comparisons are in Table 4.1 below.

Table 4.1
Independent Samples T-Test for Instruments

Measure	*p-value (2-tailed)*	η^2
MLU pre-test	0.011[a]	0.025
MLU post-test	0.157	0.010
ΔMLU	0.271	0.005
FCI pre-test	0.436	0.003
FCI post-test	0.008[b]	0.030
FCI gains	0.000[b]	0.072
TSR	0.092	0.013

[a] Although the MLU pre-test is statistically different between samples 1 and 2, the MLU post-test and ΔMLU are not, indicating overall similarities among the samples regarding the usage of "force."

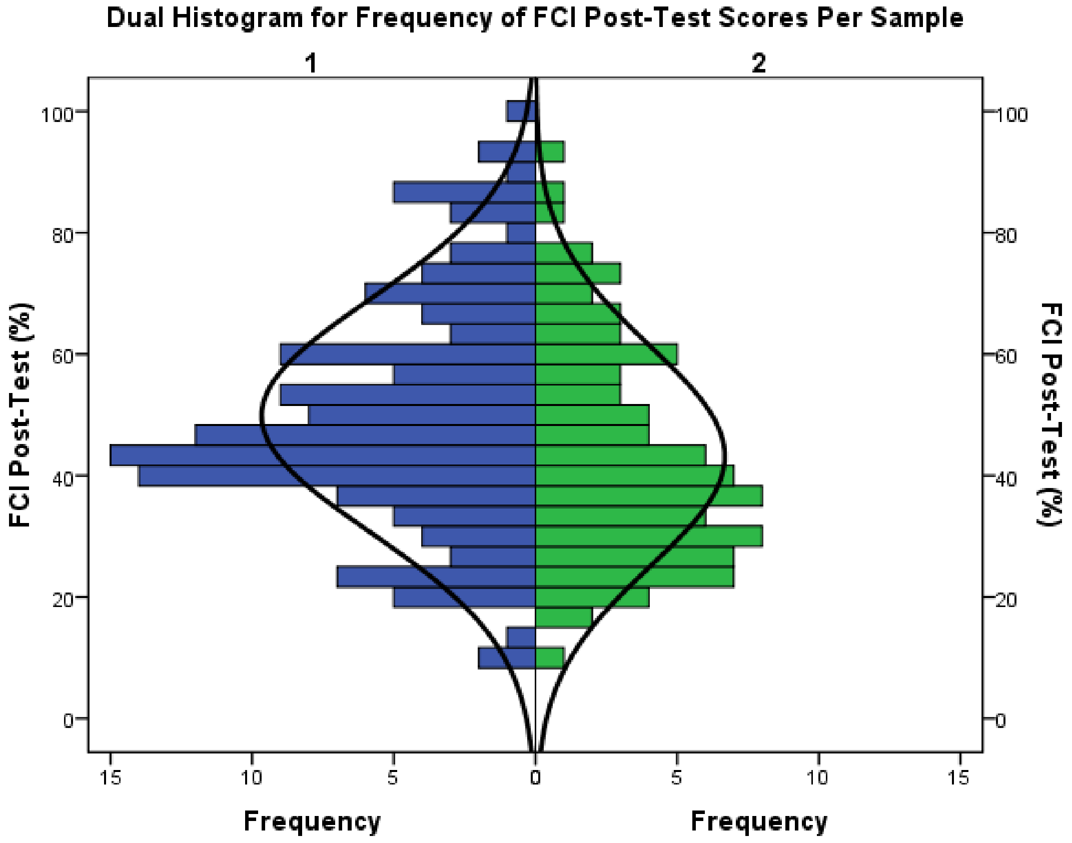
[b] These *p*-values are cause for concern as they indicate that there is a statistical difference between samples 1 and 2 regarding performance on the FCI post-test and overall gains on the FCI.

As can be seen by the range in *p*-values, these results do not provide overwhelming evidence that samples 1 and 2 are equivalent. To the contrary, the differences among *p*-values challenge the legitimacy of combining the samples and needs to be addressed. The rightmost column, consisting of values of eta squared, contains correlation ratios yielding a measure of the strength of the relationships between the grouping variable (research sites) and scores on the measures. To justify combining samples on the basis of similarity, small values of eta squared are desired. Values of 0.01 are considered small effects, 0.06 medium and 0.14 percent large

(Cohen, 1977). Based on this convention, the greatest issue lies with the apparent moderate difference found in the FCI gains between the samples.

Another possible explanation for the statistically significant differences could be the result of a violation of the assumption of normal distributions among samples or due to unequal sample sizes. To investigate the possible violation of a normal distribution, a vertical dual histogram plot was generated for the FCI post-test and FCI gains. Superimposed on the histograms are solid lines representing the normal curves for each sample.

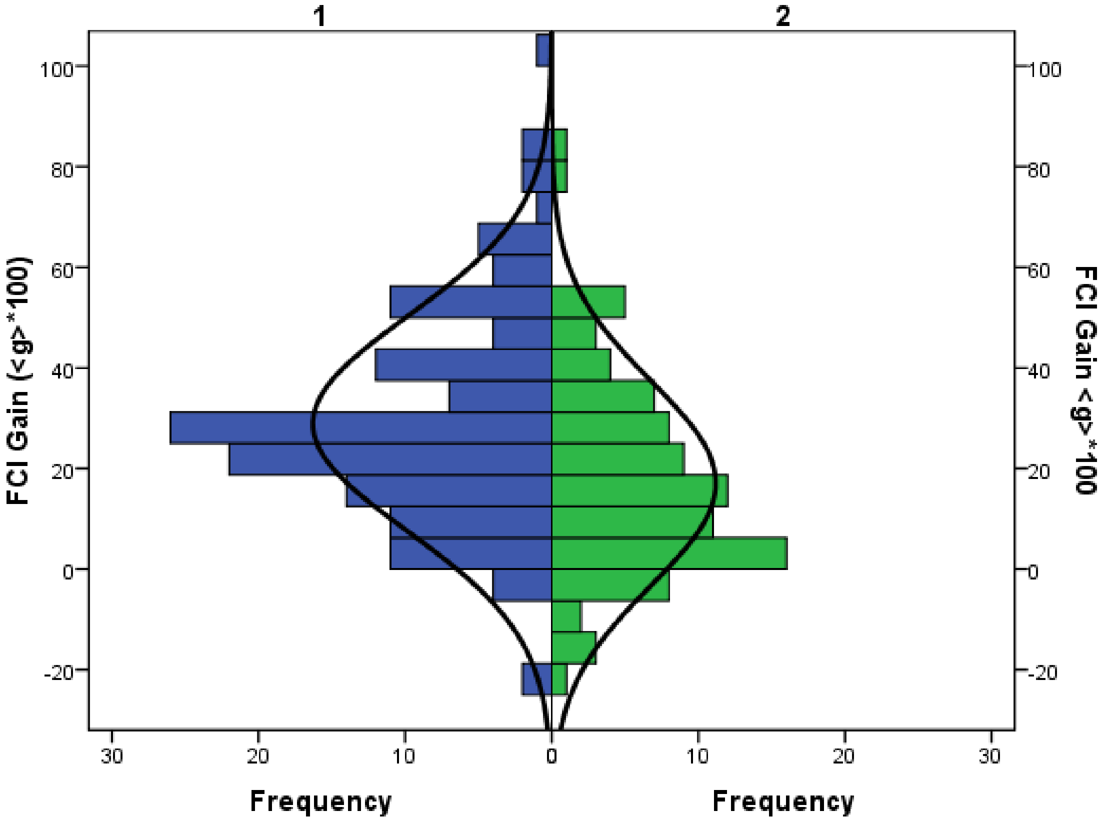
Figure 4.1



In the dual histograms plot of Figure 4.1, quick inspection shows that although the range of FCI post-test scores for the samples do not violate the assumption of a normal distribution, sample 2 is skewed toward a lower FCI post-test score average. As can be seen in Figure 4.2 below, the distributions for the FCI gains are normally distributed for each sample but skewed toward a lesser mean for sample 2.

Figure 4.2

Dual Histogram for Frequency of FCI Gains Per Sample



Combining the samples yields distributions that do not violate the assumption of a normal distribution for FCI post-test performance or for computed FCI gains:

Figure 4.3

Frequency of FCI Post-Test Scores

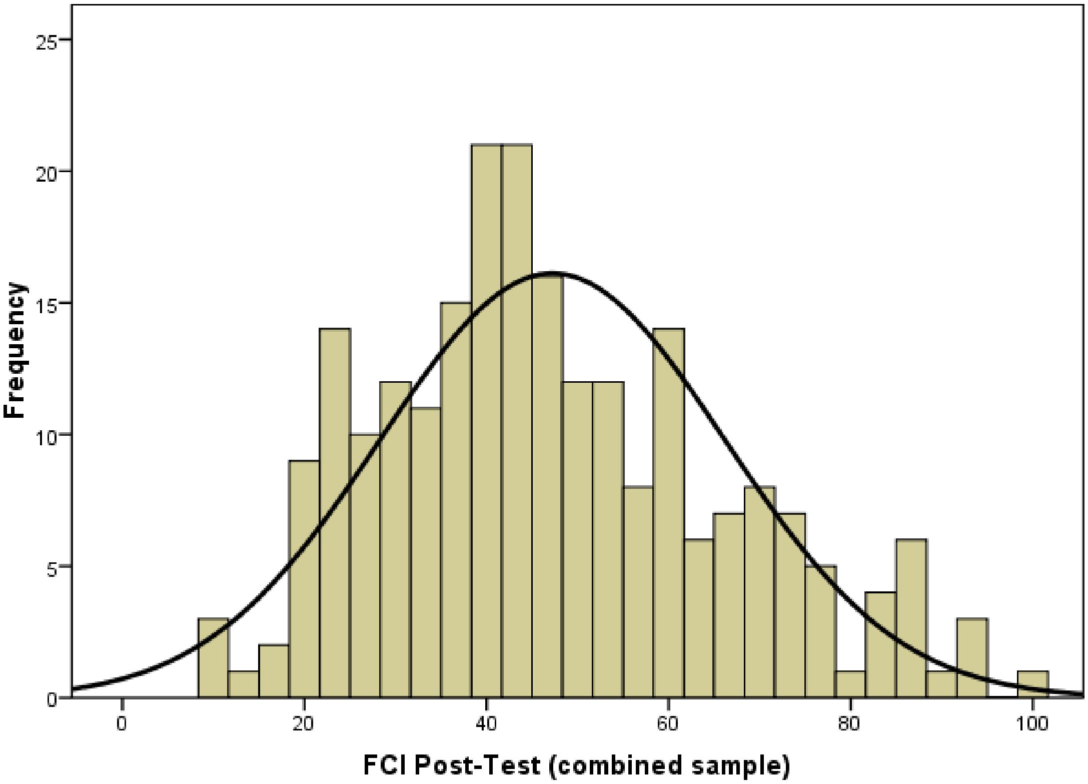
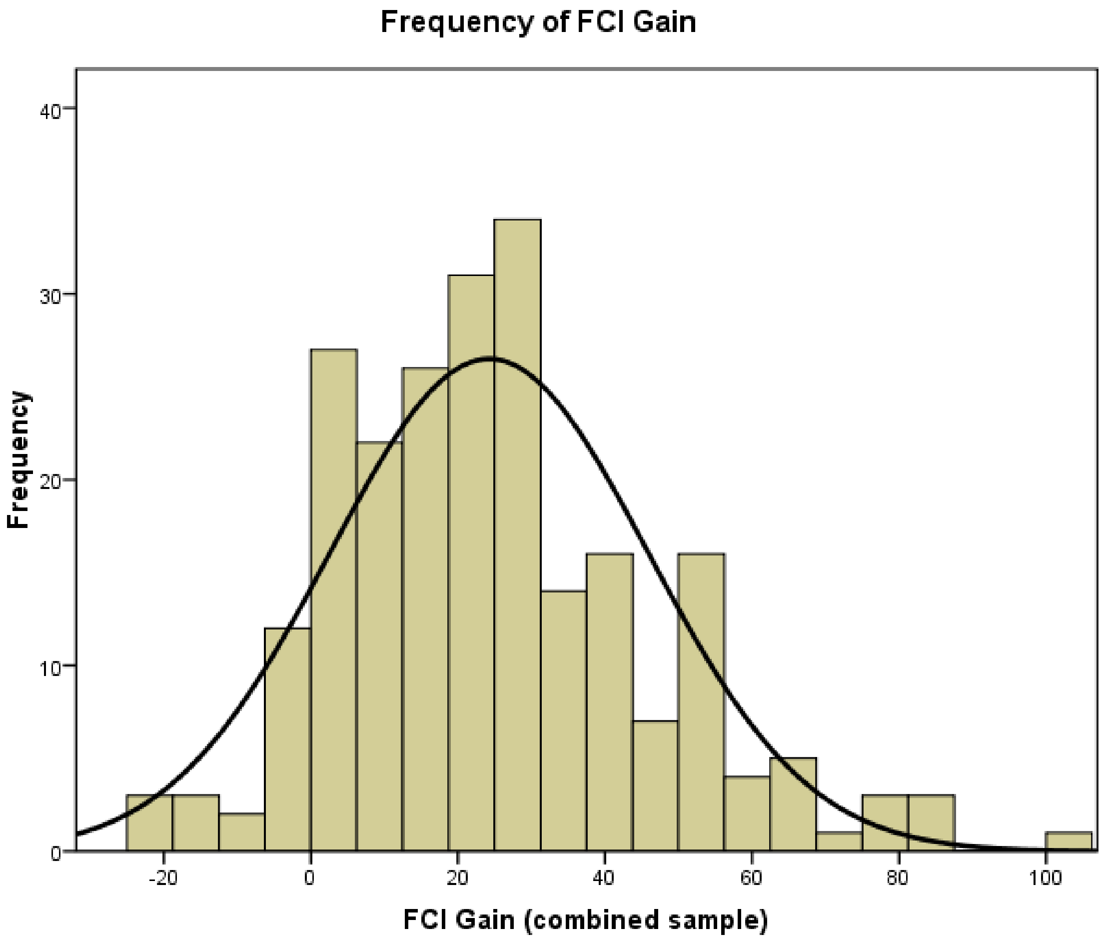


Figure 4.4

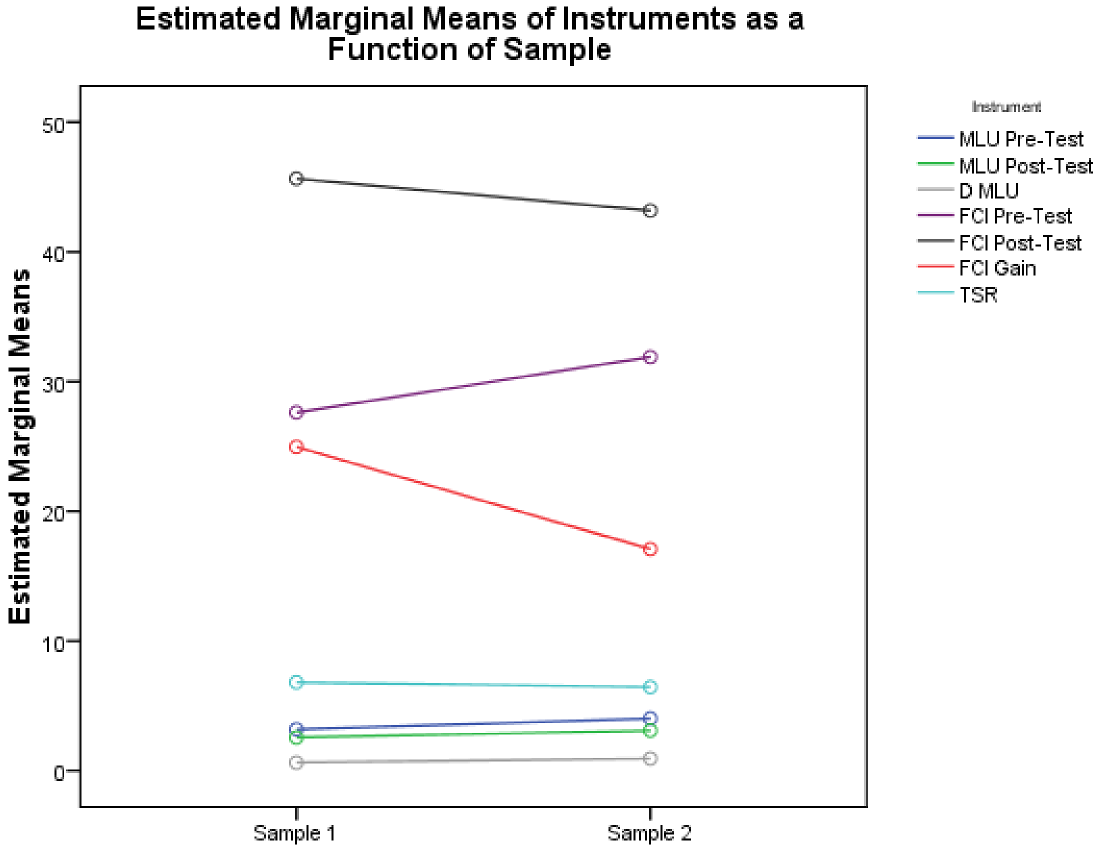


Neither sample individually violates the assumption of a normal distribution for FCI gains or FCI post-test scores. When combined, the samples collectively also do not violate this assumption for the same measures. However, more detail in discerning the source and detriment of the statistical difference between the samples on FCI performance is in order. The second most logical place to look is in the differences that exist among the sample sizes.

A strategy for inspecting the data in a way to account for varying sample sizes is to compare the means of all instruments simultaneously using a split-plot. Because the number of participants at each research site was different, plotting unweighted (marginalized) means in this manner is most appropriate.

In Figure 4.5 below, the greatest differences across instruments for the means between the research sites occur with the FCI post-test (the upper most line), the FCI

Figure 4.5

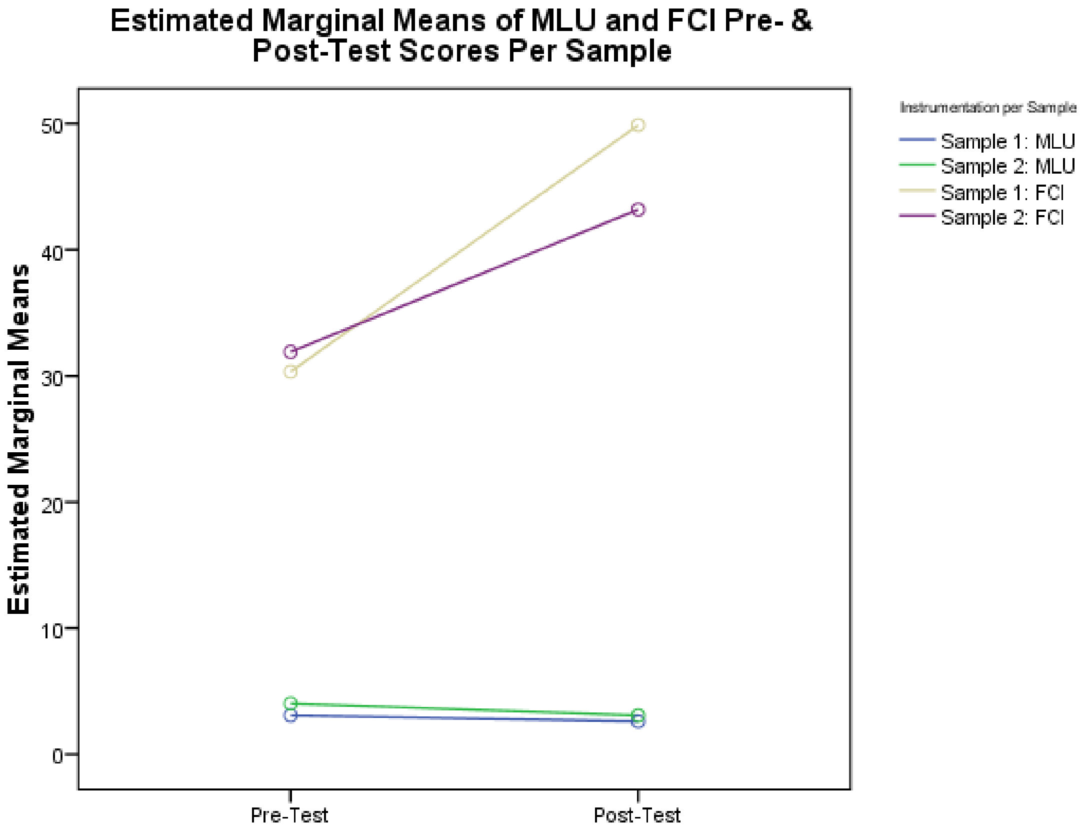


pre-test (the second line from the top) and the FCI gain (the third line from the top). Because FCI gain is a function of the FCI pre- and post-tests, it follows that if a difference exists among any one of these means for a sample then differences would exist among the other two.

It is clear from this figure that means on the TSR, MLU pre-test, MLU post-test and ΔMLU (the four lowermost lines from top to bottom, respectively) do not deviate much among the samples. The means of the remaining three instruments warrant further investigation as the steeper slopes represent greater differences among the samples. To make a general comparison of the trends of each sample for which there was pre- and post-test data, another plot of marginal means was generated.

In the second marginal means plot below (Figure 4.6), FCI and MLU pre- and post-test performances indicate that while differences between the samples exist, the relative trends are the same. Namely, FCI scores increase and MLU performance indicates a decrease in the occurrence students use “force” synonymously with other terminology.

Figure 4.6



Based on the supporting information presented in this section, the samples were merged for the analyses of this study. Justifications for merging the sample include 1) although statistically significant differences existed among some of the instrument performances, the effect size was medium at most, 2) the marginalized means per sample exhibited the same general trends, 3) the combined sample does not violate the assumption of a normal distribution, 4) trends within the two samples separately are in the same direction, and 5) the sample whose distribution was the most skewed was the smaller of the two samples. Motivation for merging the samples is primarily to increase the capability to generalize to the population of physics students taking traditional lecture introductory college physics. The statistical significance noted at the beginning of this section will not go unheeded, however. For each analysis that follows, special note is made whether or not statistical outcomes change significance when either sample is considered separately.

Combined Samples Data Analyses

As a means for identifying possible unanticipated relationships and confirmation that the hypotheses put forth earlier are plausible, correlations were computed with the data across all instruments. Instruments with a subscript "pr" correspond to pre-test, "po" corresponds to post-test, "Δ" corresponds to change (only for the MLU instrument), "<g>" corresponds to gain (only for the FCI instrument) and "∑" corresponds to sum (only for the MLU instrument). Values in the $\mathrm{MLU}_{\sum}$ column represent the sum of occurrences participants used "force" synonymously with other terminology on the pre- and post-tests, as opposed to the difference between pre- and post-test MLU performance (ΔMLU). The purpose for including this column addresses possible misinterpretations of trends involving ΔMLU data. This will be more clearly explained toward the end of this section.

Table 4.2

Correlations Among Instrument Performance (Combined Samples)

	FCI_{pr}	MLU_{pr}	FCI_{po}	MLU_{po}	MLU_{Δ}	$\text{FCI}_{<g>}$	TSR	$\text{MLU}_{\sum}$
FCI_{pr}	-	-0.185**	0.674**	-0.129	-0.071	0.140*	0.381**	-0.198**[b]
MLU_{pr}		-	-0.163*	0.295**	0.688**	-0.067	-0.170**	0.843**
FCI_{po}			-	-0.145*	-0.038	0.805**	0.481**	-0.192**[b]
MLU_{po}				-	-0.491**	0.086	0.022	0.763**
MLU_{Δ}					-	0.004	-0.138*[ab]	0.189**
$\text{FCI}_{<g>}$						-	0.354**	-0.094
TSR							-	0.053[c]
$\text{MLU}_{\sum}$								-

**Correlation is significant at the 0.01 level (2-tailed)

*Correlation is significant at the 0.05 level (2-tailed)

[a] This correlation is not significant for only Sample 1 data

[b] These correlations are not significant for only Sample 2 data (for 0.01 or 0.05 level, 2-tailed)

[c] This correlation is significant for only Sample 1 data (at the 0.05 level, 2-tailed)

The number of instances statistically significant correlations exist across all of the research instruments is a favorable sign that interdependencies exist. It was expected that statistically significant correlations would exist among MLU_{pre}, MLU_{po} for measures ΔMLU and $\text{MLU}_{\sum}$, since the latter are computed from the former. The same expectations are appropriate for FCI_{pre}, FCI_{po} and $\text{FCI}_{<g>}$. Perhaps the most telling correlations are those existing among performance on the TSR and other measures. All correlations between FCI measures and TSR performance are statistically significant. While this is anticipated based upon prior research, we see from Table 4.2 that although MLU_{pr} has a statistically significant correlation with TSR performance, MLU_{po} does not. This implies that how students change their usage of “force” may be independent of their reasoning ability. This implication will be further investigated later in this chapter.

Also of interest are some of the correlations that are *not* statistically significant. For example, gains on the FCI has a near zero correlation to change in performance on the MLU (pre- minus post-test). This is supportive of the notion that despite high gains on the FCI, students will continue using "force" as they had prior to instruction. However, before jumping to conclusions, further analyses are in order to discern more of the nature of these relationships. This is especially warranted since greater sample sizes may yield statistically significant relationships even though the relationships are weak. For example, Table 4.2 indicates TSR and MLU_{pr} scores share a statistically significant correlation at the 0.01 level. However r^2 is only 0.029, which means roughly only 3% of the variance in one measure is accounted for in the other.

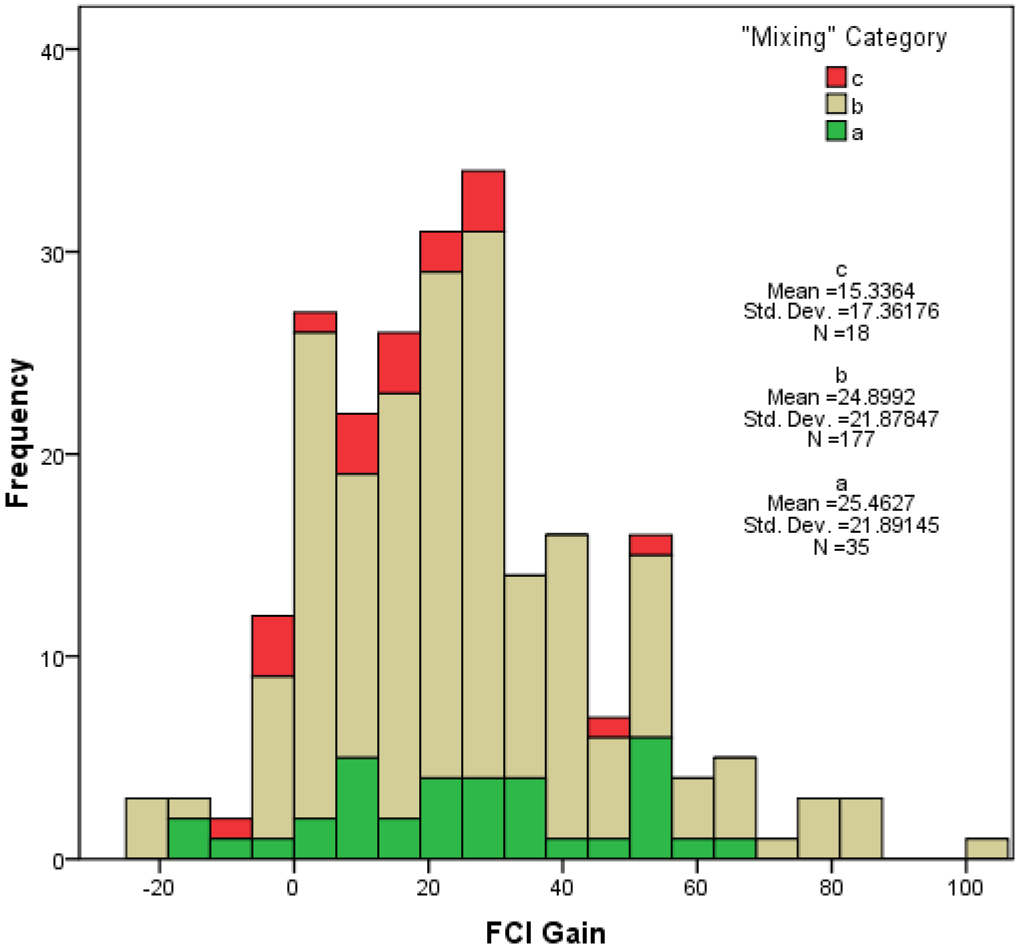
General Trends

Presented in this section are general trends that exist among the data. Most of these are presented in the form of plots either preceded or followed by a discussion of relevance. Individually, these trends are exploratory in nature. Collectively, these trends are systemic, predictable outcomes branching from the original hypothesis and research questions of this study. These scatter plots and histograms are useful for thinking of the data by offering supportive trends or trends in opposition to initial ideas. They are also useful in framing possible further questions. To draw further conclusions would be unwarranted as scatter plots and correlations are not tests of research models.

In the histogram that follows, the frequency of participants who changed how they used “force” synonymously with other terminology is plotted as a function of their FCI gains. Values for “mixing” force with other terms were obtained for each participant by subtracting the MLU post-test results from the MLU pre-test results. “Mixing” category “b” represents the participants whose change of the usage of “force” synonymously with other terminology remained within one standard deviation of the mean of the total sample. Category “a” represents participants whose usage of “force” changed more than one standard deviation from the mean toward less usage of the term synonymously with other terminology (a more favorable result of instruction). Category “c” represents participants whose synonymous usage of other terminology of “force” increased following instruction (a less favorable result of instruction). The standard deviation for change in the number of occurrences participants used “force” synonymously with other terminology was 3.145. The greatest possible number of occurrences a participant can use “force” interchangeably with other terminology on the MLU is 21 times. Therefore, the greatest possible change from pre- to post-test is ± 42.

Figure 4.7

Frequency of FCI Gain Per Level of "Mixing" Force With Other Terminology



One trend in this histogram important to this study is immediately apparent. Superimposing the three distributions of mixing categories within the same range of FCI gains reveals that each category has comparable means. Negative change on MLU performance is skewed toward lower FCI gains while no positive change greater than one standard deviation in MLU performance was recorded for the highest

FCI gains measured. Most participants fall in category b, indicating that very little change in MLU performance across FCI gains occurred.

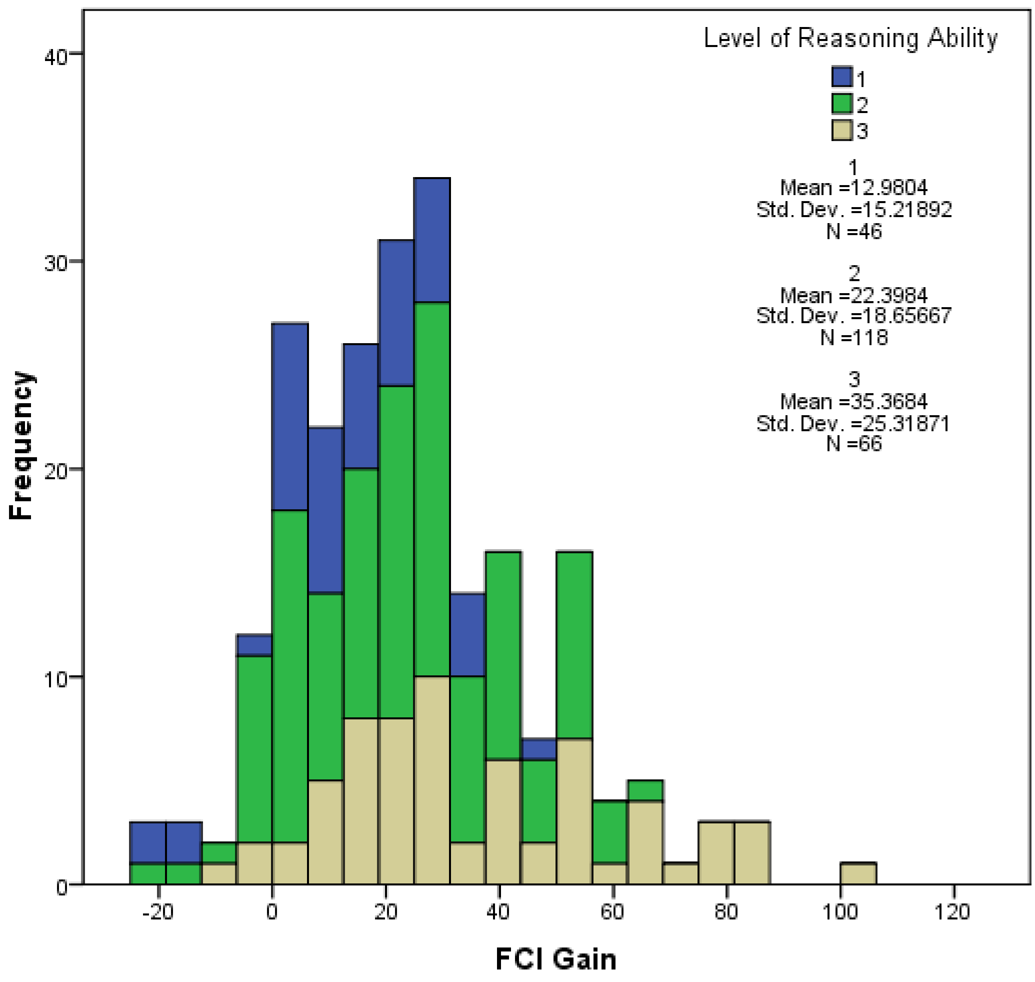
Incidentally, representing TSR performance frequency as a function of MLU "mixing" categories yields a very similar result. For ordinary normal distributions of data this might be expected: means of each category within a given sample may share common means for a different measure. In contrast to this result, as might generally be the expectation of physics instructors, one would expect that participants in category "a" should have FCI gains well above those of categories "b" and "c." In other words, the normal distribution of category "a" should be shifted toward the right if expectations are that students who demonstrate a greater understanding of Newtonian mechanics would also use "force" synonymously with other terms *less*. According to the above histogram, this is not the case.

A trend similar to the one just eluded to does exist when participants are categorized by level of scientific reasoning ability, however. To generate the plot above, participants were categorized as "concrete," "transitional" or "formal" operational learners according to their performance on the TSR (scoring as described in an earlier section).

In Figure 4.8, the normal distribution of category "3" has a mean FCI gain that is greater than those in the other categories. The reverse is true for category "1." These trends are consistent with recent work (Coletta & Phillips, 2005). Coletta and Phillips report that only students of greater reasoning abilities are capable of high FCI gains in traditional lecture-based physics courses. This is supported by the statistically significant correlation of 0.354 discussed earlier and as indicated in Table 4.2. In addition, a statistically significant correlation of 0.364 exists (at the $p = 0.000$ level, 2-tailed) between developmental stages as categorized above and FCI gains.

Figure 4.8

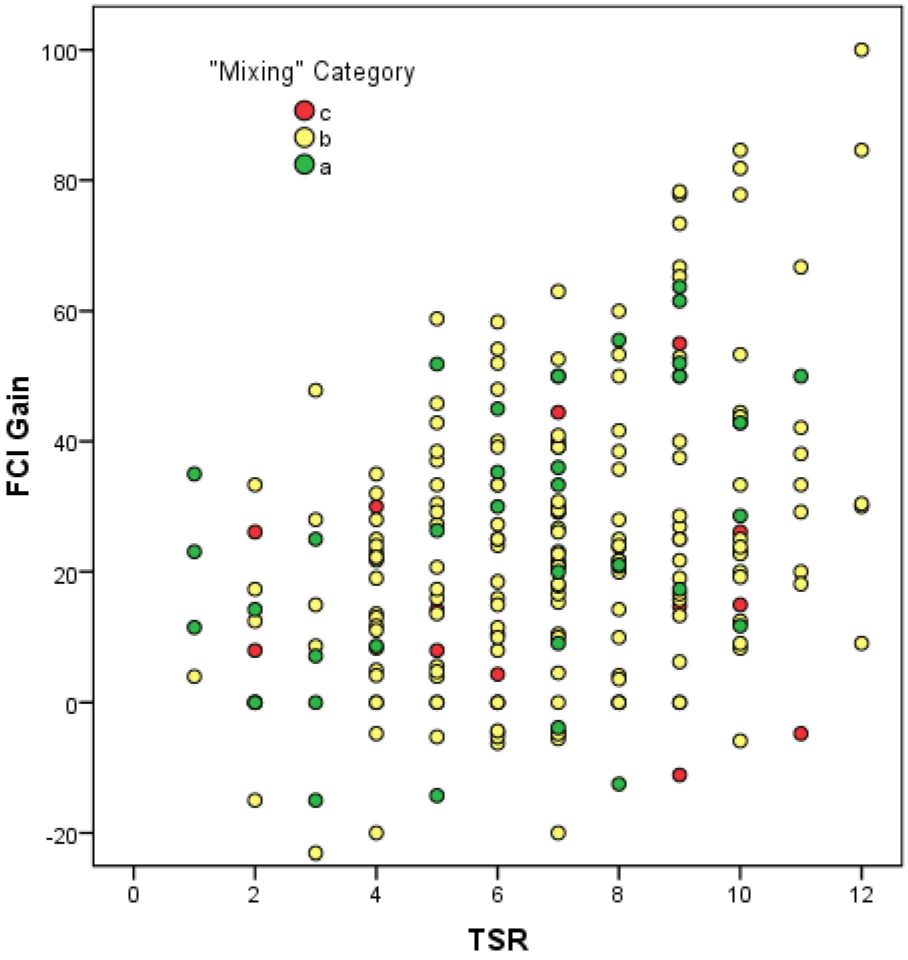
Frequency of FCI Gain Per Level of Reasoning Ability



When scatter plots of FCI gains, TSR performance and change in MLU performance are created, interrelationships related to those already described come to light. In the first scatter plot presented below, FCI gain as a function of TSR performance is plotted. The data have been categorized by the same "mixing" categories as before. Quick inspection of the plot reveals that distributions of data

Figure 4.9

FCI Gain as a Function of TSR Performance

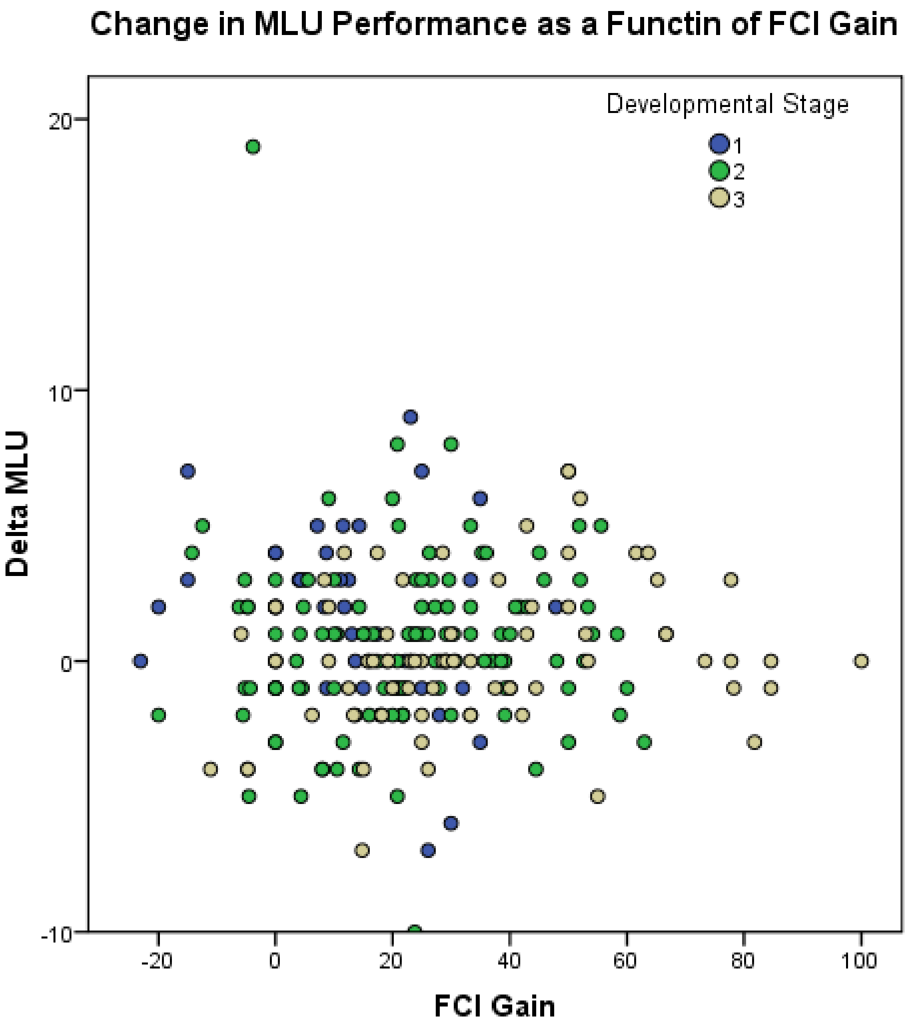


points per category are in general, random. The only obvious deviation from the random distribution of data points can be seen in the upper right hand corner of the plot. Note that a cluster of participants with an average change in MLU performance exists in this region.

Why the attention to this particular cluster of points? This region of data points corresponds to students with high FCI gains and high TSR scores, yet little change in their MLU performance. If actually a significant effect, this clustering of points implies one of two things: 1) that despite high FCI gains and greater reasoning ability, students will have only average changes in MLU performance (retaining their original colloquial usages of “force,” or 2) students with high FCI gains and greater reasoning abilities use “force” correctly already. Ironically, neither of these implications are reported anecdotally by physics instructors. Caution must be exercised before drawing conclusions: the cluster of points results from a simple scatter plot consisting of fewer than 10% of the entire sample size.

In the final two scatter plots presented in this section, performance on the MLU is plotted as a function of FCI gains. In addition, students are categorized by developmental stage as determined by performance on the TSR. The first plot (ΔMLU vs. FCI Gain) takes on the shape of an arrowhead. Of particular interest here is the stark boundary that exists just beyond the FCI gain value of 60. Only students of the most developed reasoning abilities exist in this range. This suggests that a reasoning ability threshold exists for FCI performance, and hence conceptual understanding in introductory college physics class (lecture based). There also exists a lower limit for FCI gain values. No students thinking primarily at the concrete level achieved an FCI gain greater than 50.

Figure 4.10

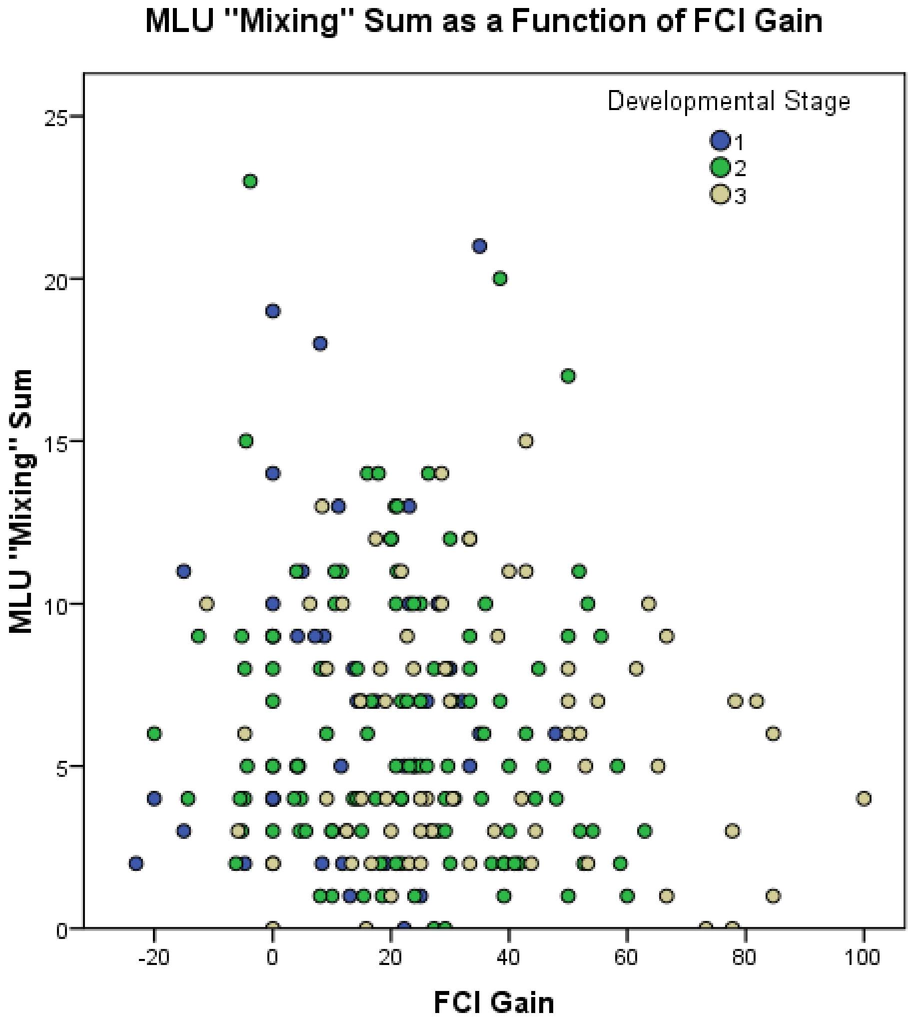


Although this trend is not the focus of this study, it is related. Based on Figure 4.10, for instance, one could argue that attention to language usage is not important since students who demonstrate an understanding of the concepts do not necessarily use “force” independently from other terms. Such an argument is in direct contrast with the primary motivations for conducting this study.

There exist other explanations to explain the right-most cluster of points in the above scatter plot. Perhaps those students with FCI gains 60 or greater simply use “force” technically correct prior to instruction and exhibit no change in their usage following instruction. Or, perhaps these students grossly misused “force” on the MLU pre-test and post-test; a difference between the two would yield values close to zero. To address these alternative accounts, a scatter plot of MLU “Mixing” Sum as a function of FCI Gain was created. “MLU Mixing Sum” is the total sum of occurrences students used “force” interchangeably on the MLU pre- and post-tests.

The scatter plot in Figure 4.11 below indicates a general negative correlation between students using “force” interchangeably with other terminology and gains on the FCI, which is what one would expect following effective instruction. What this plot does not indicate is the earlier suggestion that students with the highest gains use “force” interchangeably in excess on the MLU pre- and post-tests, netting near zero values on the ΔMLU scatter plot.

Figure 4.11



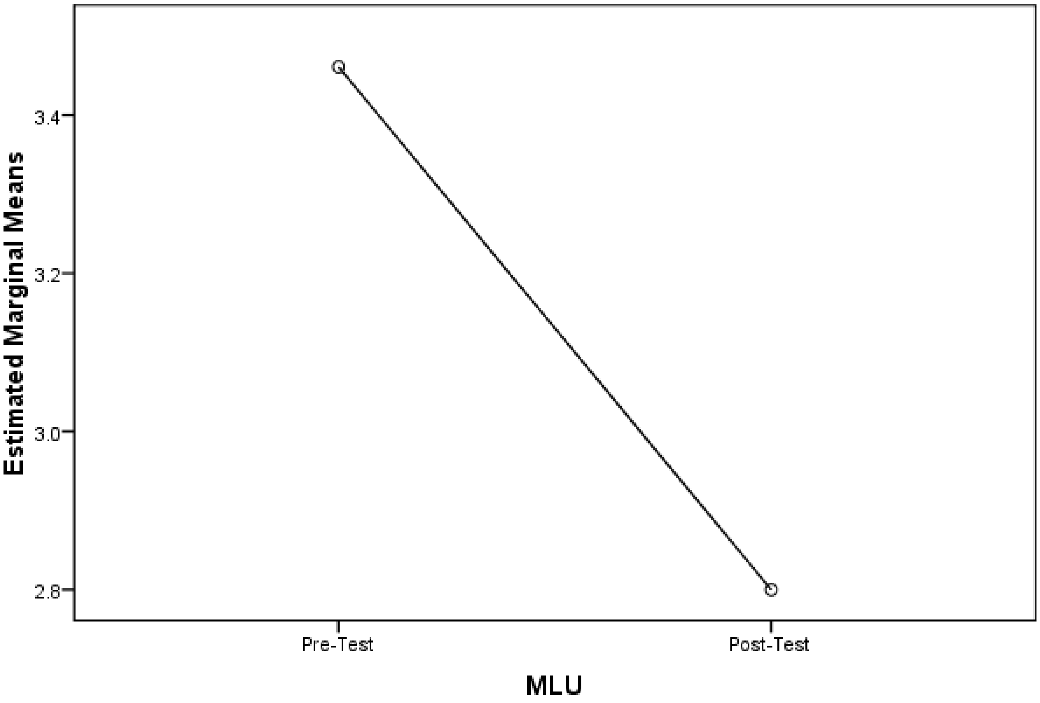
The general idea thus far then, is that the combined sample data follows a normal distribution while containing embedded sub-trends. Although many of these trends can be accounted for in the correlations that exist among the measures, not all of the sub-trends were expected. One of the most notable outcomes is the reasoning ability threshold on FCI gains which is consistent with what has already been reported in science education research. In the next section, greater statistical substance is provided that places the material from this section in context with the background of the study, literature review and the research methodology.

Quantifying the Significance of ΔMLU

Before performing regression analyses involving MLU pre- and post-test results, it is necessary to investigate the magnitude of average changes in MLU over the course of data collection. To do this, a general linear model utilizing split-plot design was used for the combined sample. From pre- to post-test, there is a significant change on MLU performance (*p* = 0.002) albeit the relationship is weak.

Figure 4.12

Estimated Marginal Means of MLU Pre- and Post-Testing



This is reflected by a low partial eta squared ($\eta^2 = 0.042$) and in the nearness of values of the marginal means of the plot above.

In the analyses that follow, scores from other instruments are used as predictors of MLU post-test scores and change in MLU performance.

Regression Analyses

The primary research objective for this study is essentially to determine whether or not students retain colloquial usage of "force" following instruction. For quantitative analyses, the strategy for testing this was to measure the extent post-test performance on the MLU could be predicted by performances on other instruments. The statistical model that best fits this research design is regression analysis. FCI post-test, TSR and MLU pre-test performances were selected as the independent variables and subsequent MLU performance the dependent variable. Two sets of regression analyses were performed: one with MLU post-test performance as the dependent variable, the other with ΔMLU as the dependent variable. Originally, only ΔMLU was going to be used as the dependent variable. However, ΔMLU is a function of both MLU pre- and MLU post-test performances. To truly isolate usage of "force" at the end of instruction required using MLU post-test scores in regression analyses. Reported below are the statistical results for the regression analyses performed.

TABLE 4.3
Multiple Regression (MLU post-test as the Dependent Variable)

Independent variables	β *(standardized coefficients)*	*p-value (2-tailed)*	R^2
MLU pre-test	0.288	0.000	
FCI post-test[a, b]	-0.144	0.047	0.103
TSR	0.096	0.185	

[a] Performing multiple regression using FCI gain yields $\beta = -0.087$, $p = 0.200$, $R^2 = 0.094$.
[b] Performing multiple regression using FCI pre-test yields $\beta = -0.101$, $p = 0.145$, $R^2 = 0.096$.

Outside of MLU pre-test scores, performance on the FCI post-test is the greatest predictor (albeit inversely) of the frequency of occurrences students mix "force" with other terminology following instruction. This is indicated by the magnitude of the β terms. In short, those with greater FCI gains mix "force" less with other terminology. There is not a significant statistical relationship between TSR performance and the frequency students use "force" synonymously with other terminology following instruction.

Data from Table 4.3 suggests that reasoning ability is not a strong predictor of how frequently students use "force" synonymously with other terminology. Instead, the greatest predictor—as expected—is initial language usage followed by final conceptual understandings. This is consistent with the notion that language usage is resilient and that organizing conceptual understandings has a lasting effect on one's mental models.

Presented next are data revealing similar relationships, but with the change in MLU scores as the dependent variable (pre-test minus post-test; ΔMLU for short).

These results are consistent with the order and level of statistical significance FCI and TSR performances held in the results presented in Table 4.3.

TABLE 4.4
Multiple Regression (ΔMLU as the Dependent Variable)

Independent variables	*β (standardized coefficients)*	*p-value (2-tailed)*	R^2
MLU pre-test	0.693	0.000	
FCI post-test[a, b]	0.109	0.047	0.483
TSR	-0.073	0.185	

[a] Performing multiple regression using FCI gain yields $\beta = 0.066$, $R^2 = 0.692$, $p = 0.200$.
[b] Performing multiple regression using FCI pre-test yields $\beta = 0.077$, $R^2 = 0.692$, $p = 0.145$.

The primary difference between the results of Tables 4.3 and 4.4 is in the variance of the dependent variable accounted for by the independent variables (R^2). Nearly 50% of the variance of ΔMLU is accounted for in multiple regression analysis as opposed to only 10% for MLU post-test. This increases to nearly 70% when FCI gain or FCI pre-test is used instead of FCI post-test results. The reason for this statistical behavior is a result of the properties of ΔMLU scores: they are a function of both MLU pre- and post-test scores. Likewise, FCI gain is a function of both FCI pre- and post-test scores. These tables illustrate that using dependent variables which are a function of the independent variables may yield conspicuously high correlations in regression analysis. However, for analyses presented in this study, it is noteworthy that in spite of these heavy dependencies on the MLU pre-test, the FCI post-test was still a significant predictor of MLU post-test performance and ΔMLU.

For thoroughness and to serve as cross validation between samples, regression analyses were also performed using data from each sample separately. Although the p-values varied, statistical significances remained with FCI post-test scores.

CHAPTER 5

Conclusions and Discussion

Structure of Conclusions and Discussion Chapter

In the Purpose of Study, several motivating questions were posed. In this section, answers to these questions will be presented in the order the questions were originally offered. Immediately following, the more specific research questions will be addressed based on the data collected and the subsequent analyses reported in Chapter 4. This will be followed by a recap of the statement of the problem with concluding remarks. And finally, the chapter will close with a discussion of research limitations and considerations for further research.

Answers to Motivating Questions

The first motivating question posed was “How do trends in students’ colloquial usages of terminology related to “force” prior to and following course instruction compare to conceptual gains of the same concept?” The correlation between ΔMLU and FCI gain is strikingly low ($r = 0.004$). In fact, there exist only a few significant correlations between any MLU measure and any FCI measure. Those correlations that are statistically significant have low values of r^2 (the maximum value of which is 0.039), indicating weak relationships. Furthermore, the scatter plot of MLU_{Σ} as a function of FCI gain also has a low correlation ($r = -0.094$). Perhaps most telling is Figure 4.7 illustrating that students categorized by degree of "mixing" "force" have very similar means for FCI gains. However, the scatter plot of ΔMLU as a function of FCI gain shows that the variance in the change of the usage of "force" decreases with increasing conceptual gain. This result is in direct contrast with the assumption that language usage improves the most for students with greater conceptual gains.

The next question was "What relationships, if any, exist between colloquial usages of "force" and students' scientific reasoning abilities?" There were similar findings in the answers to this question as the first: scant statistical significances and weak relationships (low r^2 values) for those correlations that were significant. Notable trends from scatter plots indicate that per category ("mixing" or developmental stage), more subtle interactions may be taking place. For example, in Figure 4.10 students of fully developed reasoning abilities and achieving the greatest FCI gains do not change how they use "force" beyond one standard deviation from the mean.

The answer to the third motivating question is no: students' reasoning abilities and conceptual gain, as measured by gains on a test of conceptual understanding of force, are not mutually exclusive predictors of measured changes in colloquial usages of "force." Table 4.4, footnote "a" indicates that although collectively TSR and FCI gain account for nearly 70% of the variance in ΔMLU, each variable is not a significant predictor individually.

The final two questions are closely related. The answer to the question “Does lack of conceptual change, as measured by gains on a test of conceptual understanding of force, necessarily preclude a positive change in colloquial language usage” is no. Figure 4.10 clearly shows that participants with low FCI gains are capable of making positive changes in how they use “force.” Ironically, the same scatter plot indicates the reverse may be the case for students who demonstrate greater conceptual change. Therefore the answer to the final motivating question “Is significant conceptual change, as measured by gains on a test of conceptual understanding of force, necessarily mirrored by positive changes in colloquial language usage” is also no. Students with greater conceptual gains change how they use the term “force” less than those of lesser gains.

Answers to Research Questions

Each research question posed at the beginning of this study are intimately tied to measures obtained using the MLU. Therefore, deliberate acts were carried out to ensure the instrument’s credibility for this research. Within the analyses section, it was demonstrated that pre- and post-test performances on the MLU were consistent among the separate samples. Prior to those results, details of the instrument’s validity and reliability were provided. The MLU, therefore, is a fairly robust instrument regarding consistency across samples. With this knowledge in the foreground, each research question is addressed below.

The first research question was: “Are there significant changes in students’ colloquial usage of the term “force” following instruction?” Using a general linear model with repeated measures of variables, the effect (pre- and post-testing amidst instruction) was found to be statistically significant ($p = 0.002$).

Much of the substance in the answers to the second research question exists within the discussion of the motivating questions. “What are the relationships among colloquial language usage of “force,” scientific reasoning ability and conceptual change for student in a traditionally taught introductory college physics course?” Based on the results of this study, some general conjectures can be made from the results addressing the second research question.

Initial usage of the term “force” synonymously with other terminology does not appear to be related to scientific reasoning ability or conceptual gain in Newtonian mechanics. Although it was observed that gains on the FCI do predict changes in performance on the MLU, it was a weak relationship most likely statistically significant due to a small effect detected in a large sample.

Interrelationships between these variables may not be fully describable via linear regression models. Referring back to Figure 4.10, it is very interesting that the variance of change in MLU performance appears to be a function of FCI gains. In other words, students achieving lesser gains on the FCI change how they use “force” more than students of greater FCI gains; this trend is non-directional. The expectations were that students of lesser FCI gains would have more negative values for ΔMLU and students of greater FCI gains would have more positive ΔMLU scores.

Revisiting the Statement of the Problem

The findings of this study are consistent with the anecdotal claims referenced in the problem statement of Chapter 1. Specifically, this study provides quantitative evidence that residual colloquial language usage exists among students following instruction. And, while students who achieve the greatest gains on the FCI change their usage of “force” the least, students achieving the least gains on the FCI change their synonymous usage of “force” with other terms the most; either for the worse or for the better. Because of this kind of relationship, performance on the FCI is only a weak predictor of change on MLU post-test performance. The irony is that students whom instructors assume stand the greatest chance at improving their language usage (by demonstrating sound conceptual understandings) actually show the greatest resistance toward change. The additional nuance is that other students with lower gains change their usage of “force” the most, but in either direction; toward the better or toward the worse.

Colloquial usage of “force” appears to be independent of reasoning ability since reasoning ability is not a significant predictor of MLU performance. However the relationship between FCI performance and reasoning ability has been shown to be significant in the correlations presented in this study, verifying the findings of other work in physics education research.

Left unanswered at this point is what actions should be taken to address colloquial usage of terminology in the classroom? More research needs to be completed before this question can be answered because the full nature of the issue has not been ascertained.

Limitations of the MLU

As used in this study, scoring the MLU presented challenges that were in part overcome, but in the end resulted in limitations of conclusions that can be drawn from its use.

Although ΔMLU and MLU_{Σ} values were tabulated and used in the analyses of this study, one limitation of the MLU is that students who choose only “force” responses to items on the MLU pre- and post-tests will exhibit no change in MLU performance. Similarly, students mixing combinations of terms other than “force” or students who answer items on the MLU technically correct will appear the same statistically. The issue is that identifying these students as “resistant to changing their language usage” places all of them in the same category despite the very different reasons governing their choices. This could be remedied by a detailed tracking of responses to the MLU, including additional categorization schemes. As the MLU is structured to allow multiple responses, this approach may be required if used in future studies.

Another limitation of the MLU lies in its subtle sentence structure. For example, in item three of the MLU, if a student interprets the “force” response to mean the force *acting* on the boy has increased, he/she would be correct (since the tension in the cord increases as a pendulum passes through the equilibrium position, and technically, the distance away from Earth has decreased—which increases the gravitational force acting on the boy). As written, and originally intended, the “force” response is incorrect as it implies that the boy possesses and increased force (forces cannot be possessed or held by an object, they are only the result of interactions between objects). Therefore, the phraseology and tense of the leading questions and answers may be affecting MLU performance more than anticipated.

Considerations for Further Research

Based on the outcome of this study, there are numerous directions one can go to contribute more work in this area.

First, the MLU presented and used for data collection of this study is in its first iteration. Despite efforts to establish validity and reliability, more work could be done to improve the instrument. Increasing the discrimination of scores on the MLU is certainly one area for improvement. Originally, the MLU was kept short to facilitate collection of data from large research samples. Researchers contemplating making revisions to the instrument in the future will need to balance sacrificing length for increased discrimination if greater differences among marginal means are desired.

Validating the quantitative results with a qualitative measure is another way the use of the MLU could be greatly enhanced. Originally considered for use in this study was a short video clip of world cup soccer players making a goal. The "announcers" calling the plays used "momentum" and "force" colloquially (the announcers were actually reading a choreographed script very similar to the leading questions of the current version of the MLU). The intent was to use this footage as a primer for discussion among students in focus groups. The discussion would then be transcribed and analyzed via qualitative methods. Due to lack of human resources this was not undertaken, but would serve as excellent follow up research.

Independently, the MLU pre- and post-test scores plotted against FCI gains have a normal distribution. However, ΔMLU plotted against FCI gains displays a convergence in ΔMLU toward greater FCI gains. Investigating the nature of this variance could prove very challenging and fruitful. For example, what is the cause for some individuals to exhibit significant positive changes in language usage while others of the same reasoning ability and FCI performance exhibit significant negative changes in language usage? Arguably, the ability to answer this question would have a dramatic and direct impact on improving instruction.

There are of course many other possibilities for variants of this research study. For example, an experimental design using the MLU, a control group and/or multiple instruction modes across at least two different samples would be invaluable. It could be that had multiple instructional modes been a part of the design of this study, more assertive conclusions could have been made about language usage and the utility of the MLU. Alternatively, research could be designed to investigate the level of colloquial language usage *instructors* use. The usage could be categorized as personal colloquial usage (instructors' own colloquialisms remnant of their early language development) or as generalizing (instructors choosing to use colloquialisms as an attempt to relate to students' colloquialisms). The effects of each of these usages of colloquialisms would shed light on pertinence of how instructors choose the words they use during instruction.

A multitude of opportunities exist in investigating whether or not the findings presented in this study change or remain the same when including demographic and attitudinal data in the analyses. For example, do differences among males and females exist in the usage of "force" colloquially? Do students with greater amounts of prior physical science coursework exhibit the same levels of change of language usage across reasoning abilities observed in this study? Also, would the same results be found for a population whose native language is not English?

In a similar vein, a wealth of opportunity for further research rests with investigating the interplay of usage of terminology with identified misconceptions versus lack of a conception. For example, how does change in usage of “force” for students with particular force and motion misconceptions compare to individuals broaching the material for the first time? Are there discernable differences in the change of usage of “force” for students of distinctly different misconceptions? These are valuable research questions to pursue, as they could potentially reveal pathways from initial knowledge states (ranging from no conception to stable misconceptions) to accurate and complete understandings via diagnosing language usage during instruction.

To close, one last suggestion for research is offered since it is truly at the heart of what motivated this study from its infancy. Administering the MLU (or a derivative) and the FCI (or other natural language instrument) to multiple samples of varying reasoning abilities stands to offer the greatest benefit of an instrument such as the MLU. The explanation for this claim is as follows. Choosing samples of different reasoning abilities and detecting significant differences in language usage would enable researchers to offer insight as to what students “do” per reasoning ability in thought with the terminology they use colloquially. That is, this kind of research design would help determine if colloquial language usage is a function of conceptual understanding or if conceptual understanding is a function of colloquial language usage for learners of varying reasoning abilities.

REFERENCES

Bauman, Robert P. (1992). Physics that textbook writers usually get wrong, I. Work. *The Physics Teacher, 30,* 264-9.

Bauman, Robert P. (1992). Physics that textbook writers usually get wrong, II. Heat and energy. *The Physics Teacher, 30,* 353-6.

Brookes, David T. *The role of language in learning physics* (2006). PhD dissertation, Rutgers University, New Jersey.

Brookes, David T. and Etkina, Euginia (2007). Using conceptual metaphor and functional grammar to explore how language used in physics affects student learning. *Phys. Rev. ST Phys. Educ. Res., 3*.

Clerk, D. and Rutherford M. (2000). Language as a confounding variable in the diagnosis of misconceptions. *International Journal of Science Education, 22,* 703-17.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press.

Coletta, Vincent P. and Phillips, Jeffrey A. (2005). Interpreting FCI scores: Normalized gain, preinstruction scores, and scientific reasoning ability. *American Journal of Physics 73,* (12), 1172-1182.

Crouch, Catherine H. and Mazur, Eric (2001). Peer Instruction: Ten years of experience and results. *American Journal of Physics, 69,* (9), 970-77.

deLeeuw, Nicholas & Chi, Michelene T.H. (2003). Self-Explanation: Enriching a Situation Model or Repairing a Domain Model? International Conceptual Change, Sinatra, G. & Pintrich, P. (eds.), 55-78.

diSessa, A. (1993). Towards an epistemology of physics. *Cognition and Instruction,10*, (2 & 3), 105-225.

Goldberg, Fred; Robinson, Steve; and Otero, Valerie. 2006. *Physics for Elementary Teachers*. New York: Its About Time.

Greca, Ileana Maria and Moreira, Marco Antonio (2002). Mental, physical, and mathematical models in the teaching and learning of physics. *Science Education*, 86, 106-121.

Hake, Richard (1998). Interactive-engagement vs traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66,* 64-74.

Hammer, David (1996). Misconceptions or P-Prims: How may alternative perspectives of cognitive structure influence instructional perceptions and intentions? *The Journal of the Learning Sciences, 5,* (2), 97-127.

Hart, Christina (2002). If the sun burns you is that a force? Some definitional prerequisites for understanding Newton's laws. *Physics Education, 37,* 234-38.

Hayakawa, S. I. (1940). *Language in Thought and Action.* New York: Harcourt, Brace and Company.

Hestenes, David and Halloun, Ibrahim (1995). Interpreting the Force Concept Inventory: A response to Huffman and Heller. *The Physics Teacher, 33,* 502-6.

Hestenes, David and Wells, Malcolm (1992). Force concept inventory. *The Physics Teacher, 30,* (3), 141-58.

Hestenes, David; Wells, Malcolm; and Swackhamer, Gregg (1992). A mechanics baseline test. *The Physics Teacher, 30,* (3), 159-66.

Hilborn, Robert C. (2000). Let's ban 'work' from physics! *The Physics Teacher, 38,* 447.

Hilborn, Robert C. (2003). Comment on "Physical and colloquial meanings of the term 'work.' *American Journal of Physics 71, (8),* 743.

Huffman, D. and Heller, P. (1995). What does the force concept inventory actually measure? *The Physics Teacher, 33*, 138-43.

Itza-Ortiz, S; Rebello, N. S.; Zollman, D et al. (2003). The Vocabulary of Introductory Physics and its Implications for Learning Physics. *The Physics Teacher, 41*, 330-336.

Jackson, David P.; Laws, Priscilla W.; and Franklin, Scott V. 2003. *Explorations in Physics: An Activity-Based Approach to Understanding the World.*

Johnson, Mark. 1987. *The Body in the Mind: The Bodily Basis of Meaning, Imagination, and Reason*. Chicago: University of Chicago Press.

Korzybski, Alfred. (1933). *Science and Sanity: An introduction to non-Aristotelian systems and general semantics,* Chapter XXV.

Lakoff, George and Johnson, Mark. (2003). *Metaphors We Live By*. Chicago: University of Chicago Press.

Laws, Priscilla W. 1997. *Workshop Physics*. New York: John Wiley & Sons, Inc.

Lawson, Anton E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching, 15,* (11), 11-24.

Lawson, Anton E. Classroom test of scientific reasoning, revised edition. http://lsweb.la.asu.edu/alawson/LawsonAssessments.htm.

Lawson, Anton E., Banks, D.L., and Logvin, M. (2007). Self-efficacy, reasoning ability, and achievement in college biology. *Journal of Research in Science Teaching, 44*, 706-24.

Lemke, Jay. 1990. *Talking Science: Language, Learning, and Values.* Norwood, NJ: Ablex Publishing Corporation.

Leonard, W.J., Dufresne, R.J., Gerace, W.J., and Mestre, J.P. 2000. *Minds on Physics, Activities and Reader* (6 volumes). Dubuque, Iowa.

Maier, Steven J. and Marek, Edmund A. (2005). The learning cycle: a re-introduction. *The Physics Teacher, 44*, 109 – 113.

Maloney, David P.; Desbian, Dwain; and Beichner, Bob. (2004) *Intro to Standardized Assessment Instruments for Novices*. Workshop given at the summer AAPT meeting, Sacramento, CA. The most recent delivery of this workshop was 2005.

Marek, Edmund A. and Cavallo, Ann M. L. 1997. *The learning cycle: elementary school science and beyond*. New Hampshire: Heinemann.

McClain, John (1990). The grammar of force. *The Physics Teacher, 28,* 519.

McDermott, Lillian C. 1996. *Physics By Inquiry*. New York: John Wiley & Sons, Inc.

McDermott, Lillian C. and Redish, Edward F. (1999). Resource Letter: PER-1: Physics Education Research. *American Journal of Physics, 67,* (0), 755-67.

McDermott, Lillian C. and Shaffer, Peter S. 2002. *Tutorials in Introductory Physics.* New Jersey: Prentice Hall.

McKinnon, Joe W. and Renner, John W. (1971). Are colleges concerned with intellectual development? *American Journal of Physics, 39,* (9), 1047-52.

Mclldowie, Eric. (1995). Energy transfer—where did we go wrong? *Physics Education, 30,* 228-30.

Meador, Granger. 2001. *Inquiry Physics: A Modified Learning Cycle Curriculum, v1.1.*Bartlesville, OK: www.meador.org.

Mendelson, Kenneth S. 2003. Physical and colloquial meanings of the term 'work.' *American Journal of Physics, 71, (3),* 279-81.

Mendelson, Kenneth S. 2003. Author's response in Letters to the Editor: Comment on "Physical and colloquial meanings of the term 'work.'" *American Journal of Physics, 71, (8),* 743.

Novak, Gregor M.; Patterson, Evelyn T.; Garvin, Andrew D.; and Christian, Wolfgang. 1999. *Just-In-Time Teaching*. New Jersey: Prentice Hall.

Palmer, David H. (2003). Investigating the relationship between refutational text and conceptual change. *Science Education*, 87, 663-684.

Phillips, J. 1975. *The Origins of Intellect: Piaget's Theory*, 2nd ed. San Francisco, CA: W.H. Freeman.

Piaget, J. 1975. *Biology and Knowledge Theory: An essay on the relations between organic regulations and cognitive processes*. Chicago, IL: University of Chicago Press.

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. Review of Educational Research, 6, 167-199.

Podolefsky, Noah S. and Finkelstein, Noah D. (2007). Analogical scaffolding and the learning of abstract ideas in physics: Empirical studies. *Phys. Rev. ST Phys. Educ. Res. 3*.

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

Quine, W.V. (1986). The Philosophy of W.V. Quine: an Expository Essay. Tampa University Press of Florida.

Renner, John W. and Lawson, Anton E. (1973). Piagetian theory and instruction in physics. *The Physics Teacher, 11*, 165-69.

Renner, John W. and Marek, Edmund A. (1990). An educational theory base for science teaching. *Journal of Research in Science Teaching*, 241 – 46.

Renner, John W.; Nickel, Jim A.; Westbrook, Susan L.; and Renner, Michael J. 1985. *Investigations in Natural Science: Physics.* Norman, OK: University of Oklahoma Science Education Center.

Smith, Crosbie. 1998. *The Science of Energy: A cultural history of energy physics in Victorian Britain.* Chicago: The University of Chicago Press.

Strike, K. A. & Posner, G. J. (1992). A revisionist theory of conceptual change. In R. Duschl & R. Hamilton (Eds.), Philosophy of Science, Cognitive Psychology, and Educational Theory and Practice (pp. 147-176). Albany, NY:SUNY.

Styer, Daniel F. (2000). The word 'force.' *American Journal of Physics, 69,* 631-632.

Thornton, R. K. and Sokoloff, D. (1999). Force and Motion Conceptual Evaluation. *Tools for Scientific Thinking, CSMT.* Tufts University.

Touger, Jerold. (1991). When words fail us. *The Physics Teacher, 29,* 90-95.

Touger, Jerold. (2000). The role of language in learning physics: beyond semantics. *The American Journal of Physics 68,* (4), 306 – 307.

Vosniadou, Stella (1994). Capturing and modeling the process of conceptual change. *Learning and Instruction, 4*, 45-69.

Vygotsky, Lev (1986). *Thought and Language.* Cambridge, Massachusetts: MIT Press.

Wells, Malcolm; Hestenes, David; and Swackhamer, Gregg (1995). A Modeling Method for high school physics instruction. *American Journal of Physics, 63,* (7), 606-619.

Williams, H. T. (1999). Semantics in teaching introductory physics. *American Journal of Physics, 67,* 670-680.

APPENDICES INDEX

APPENDIX A: Participant Consent Form
APPENDIX B: Participant Copy of Consent Form
APPENDIX C: Pre-Test Administration Instructions
APPENDIX D: Post-Test Administration Instructions
APPENDIX E: The Force Concept Inventory (FCI) Access Information
APPENDIX F: Lawson's Classroom Test of Scientific Reasoning Ability (TSR)
APPENDIX G: Key to TSR
APPENDIX H: The Mechanics Language Usage Instrument (MLU) Pre-Test
APPENDIX I: The Mechanics Language Usage Instrument (MLU) Post-Test

APPENDIX A:
Participant Consent Form

University of Oklahoma
Institutional Review Board
Informed Consent to Participate in a Research Study

Project Title:	Learning Physics Concepts as a Function of Colloquial Language Usage
Principal Investigator:	Steven Maier
Department:	Instructional Leadership and Academic Curriculum (ILAC)

You are being asked to volunteer for this research study. This study is being conducted at the University of Oklahoma and Oklahoma State University. You were selected as a possible participant because you are enrolled in algebra based general physics I.

Please read this form and ask any questions that you may have before agreeing to take part in this study.

Purpose of the Research Study
The purpose of this study: to determine the extent of change in everyday language usage of technical terminology following instruction. Specifically, this study is investigating terminology common to physics and everyday language.

Number of Participants
About 500 people will take part in this study.

Procedures
If you agree to be in this study, you will be asked to do the following things: Complete three different surveys. Two of these surveys will be administered twice during the regular semester. Completion of each survey will take between 20 and 45 minutes.

Length of Participation
Participation in this study will consist of completing five surveys requiring a time of 20 – 45 minutes each. Participation will occur on separate dates within the span of one regular academic semester.

This study has the following risks:
The study has the following risks: aside from class/laboratory time used to complete the surveys, there are no foreseeable “more than minimal” risks involved in participating in this study.

Benefits of being in the study are
None.

Confidentiality
In published reports, there will be no information included that will make it possible to identify you without your permission. Research records will be stored securely and only approved researchers will have access to the records.

There are organizations that may inspect and/or copy your research records for quality assurance and data analysis. These organizations include the Science Education Center at the University of Oklahoma and the OU Institutional Review Board.

Compensation
At the discretion of the instructor, you will be awarded bonus points for participation in this study. These points will be awarded if you participate in the study and will not be pro-rated should you be unable or choose not to complete participation in the study. Therefore, points will not be taken away if you choose to not stop your participation in the study. The extra credit earned for participation will be equivalent in weight to the points earned for completing one assigned group problem.

Names of participants will be presented to the instructor only after all of the data for the study are collected for the sole purpose of awarding extra credit.

Voluntary Nature of the Study
Participation in this study is voluntary. If you withdraw or decline participation, you will not be penalized or lose benefits or services unrelated to the study. If you decide to participate, you may decline to answer any question and may choose to withdraw at any time.

Should you not be able to adequately complete all of the surveys, your participation may be withdrawn without your consent. Circumstances that might warrant this include illegibly written or missing ID numbers on surveys, not completing all of the instruments, and leaving excessive blanks on surveys. Withdrawal from the study simply means that data from the instruments you completed will not be included in the analysis part of the study.

If you participate in the study but are withdrawn from the study, you will still receive compensation for your participation. The instructor for the course will not be informed of individuals who choose to withdraw or who are withdrawn from the study as described above.

Contacts and Questions

If you have concerns or complaints about the research, the researcher(s) conducting this study can be contacted at (580) 327 – 8562, sjmaier@nwosu.edu (PI) or (580) 325 – 1498, eamarek@ou.edu (advisor, Dr. Edmund Marek).

Contact the researcher(s) if you have questions.

If you have any questions about your rights as a research participant, concerns, or complaints about the research and wish to talk to someone other than individuals on the research team or if you cannot reach the research team, you may contact the University of Oklahoma – Norman Campus Institutional Review Board (OU-NC IRB) at 405-325-8110 or irb@ou.edu.

You will be given a copy of this information to keep for your records. If you are not given a copy of this consent form, please request one.

Statement of Consent

I have read the above information. I have asked questions and have received satisfactory answers. I consent to participate in the study.

Signature   Date

APPENDIX B:
Participant Copy of Consent Form

INFORMATION SHEET FOR CONSENT TO PARTICIPATE IN A RESEARCH STUDY

My name is Steven Maier and I am a doctoral student in Science Education within the Department of Instructional Leadership and Academic Curriculum at the University of the Oklahoma. I am requesting that you volunteer to participate in a research study titled Learning Physics Concepts as a Function of Colloquial Language Usage. You were selected as a possible participant because you are enrolled in algebra based General Physics I. Please read this information sheet and contact me to ask any questions that you may have before agreeing to take part in this study.

Purpose of the Research Study: The purpose of this study is: to determine the extent of change in everyday language usage of technical terminology following instruction. Specifically, this study is investigating terminology common to physics and everyday language.

Procedures: If you agree to be in this study, you will be asked to do the following things: Complete three different surveys. Two of these surveys will be administered twice during the regular semester. Completion of each survey will take between 20 and 45 minutes.

Risks and Benefits of Being in the Study: The study has the following risks: aside from class/laboratory time used to complete the surveys, there are no foreseeable “more than minimal” risks involved in participating in this study. There are no direct benefits to participation in this study.

Compensation: At the discretion of the instructor, you will be awarded bonus points for participation in this study. These points will be awarded if you participate in the study and will not be pro-rated should you be unable or choose not to complete participation in the study. Therefore, points will not be taken away if you choose to not stop your participation in the study. The extra credit earned for participation will be equivalent in weight to the points earned for completing one assigned group problem.

Names of participants will be presented to the instructor only after all of the data for the study are collected for the sole purpose of awarding extra credit.

Voluntary Nature of the Study: Participation in this study is voluntary. Your decision whether or not to participate will not result in penalty or loss of benefits to which you are otherwise entitled. If you decide to participate, you are free not to answer any question or discontinue participation at any time without penalty or loss of benefits to which you are otherwise entitled.

Length of Participation: The length of participation for this study is one regular academic semester. Should you not be able to adequately complete all of the surveys, your participation may be withdrawn without your consent.

Confidentiality: The records of this study will be kept private and your supervisor will not have access to your responses. In published reports, there will be no information included that will make it possible to identify you as a research participant. Research records will be stored securely. After all of the data are collected, all identifiable information will be removed from the surveys and destroyed. For purposes of the study, each survey you complete will be assigned a random identification number unique to those involved in the study. This random number will not be associated with your name or student identification number in any after completion of the data collection. All data sets will be stored in a locked filing cabinet at an off campus site. Only the PI (not your instructor) will have access to the data. Your instructor may be provided with aggregate data of the entire class once the surveys have been completed; there will be no identifiable information included in this data. Handwritten responses may be scanned and stored digitally for the purpose of presenting examples in the final research. Identifiable information will not be linked to this form of data. Only approved researchers will have access to the records.

Contacts and Questions: If you have concerns or complaints about the research, the researcher(s) conducting this study can be contacted at (580) 327 – 8562, sjmaier@nwosu.edu (PI) or (580) 325 – 1498, eamarek@ou.edu (advisor, Dr. Edmund Marek). In the event of a research-related injury, contact the researcher(s). You are encouraged to contact the researcher(s) if you have any questions. If you have any questions, concerns, or complaints about the research and wish to talk to someone other than the individuals on the research team, or if you cannot reach the research team, you may contact the University of Oklahoma – Norman Campus Institutional Review Board (OU-NC IRB) at (405) 325-8110 or irb@ou.edu.

Please keep this information sheet for your records. By completing and returning this questionnaire, I am agreeing to participate in this study.

APPENDIX C:
Pre-Test Administration Instructions

Total estimated time for completing both surveys: 40 – 50 minutes.

In-between each pair of colored paper are 35 copies of Instrument 1. An extra 15 copies are grouped separately by a paper binder.

- **This instrument must be administered 1st**.
- It should take 10 – 15 minutes to complete.
- Please make sure students write their student ID numbers on the front page.
- This instrument should be kept intact (students should not remove any pages).
- Students may make marks on this instrument (there is not a separate answer sheet).
- Please collect this instrument back from the students prior to passing out Instrument 2.

In-between each pair of blue colored paper are 35 copies of Instrument 2. An extra 10 copies are grouped separately by a paper binder.

- This instrument should be administered **following Instrument 1**.
- It should take 30 – 35 minutes.
- Please make sure students write their student ID numbers on the front page.
- The front page of this instrument should be removed and used as a separate answer sheet by the students.
- Students should only mark the answer sheet (not the instrument itself).
- Make sure that no copies of this instrument leave with students.

APPENDIX D:
Post-Test Administration Instructions

If time is an issue, the shorter survey (with checkboxes on the front page) may be given to students to complete outside of class.

Please DO NOT allow the longer survey (with the ABCDE answer sheet on top) to leave with students.

A note about the longer survey: These have been recycled to save paper; please instruct students to ignore/disregard remnant pencil markings.

APPENDIX E:
The Force Concept Inventory (FCI) Access Information

The Force Concept inventory (FCI) Access Information

At the request of the authors of the FCI, the latest current version of the FCI (1995) can be accessed by contacting directly the Modeling Instruction Program at Arizona State University. At the time of publication of this study, Jane Jackson was the coordinator of files maintained by this research group.

Current active contact information as of 2008:
Arizona State University Modeling Instruction Program
Website: http://modeling.asu.edu/

Jane Jackson
Phone: (480) 965-8438
Email: jane.jackson@asu.edu

The 1995 version of the FCI used for this research very closely resembles the 1992 version. This version of the FCI has been published and is available by accessing:

Hestenes, David and Wells, Malcolm (1992). Force concept inventory. *The Physics Teacher, 30,* (3), 141-58.

APPENDIX F:
Lawson's Classroom Test of Scientific Reasoning Ability (TSR)

CLASSROOM TEST OF SCIENTIFIC REASONING

Multiple Choice Version

Directions to Students:

This is a test of your ability to apply aspects of scientific and mathematical reasoning to analyze a situation to make a prediction or solve a problem. Make a dark mark on the answer sheet for the best answer for each item. If you do not fully understand what is being asked in an item, please ask the test administrator for clarification.

DO NOT OPEN THIS BOOKLET UNTIL YOU ARE TOLD TO DO SO

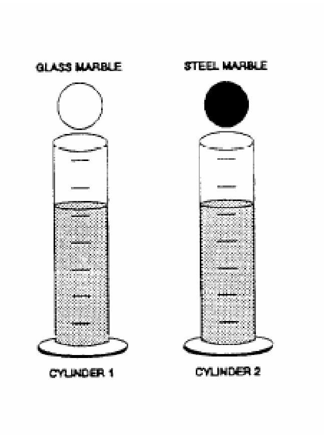
Revised Edition: August 2000 by Anton E. Lawson, Arizona State University. Based on: Lawson, A.E. 1978. Development and validation of the classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1): 11-24.

1. Suppose you are given two clay balls of equal size and shape. The two clay balls also weigh the same. One ball is flattened into a pancake-shaped piece. *Which of these statements is correct?*

    a. The pancake-shaped piece weighs more than the ball
    b. The two pieces still weigh the same
    c. The ball weighs more than the pancake-shaped piece

2. *because*

    a. the flattened piece covers a larger area.
    b. the ball pushes down more on one spot.
    c. when something is flattened it loses weight.
    d. clay has not been added or taken away.
    e. when something is flattened it gains weight.



3. To the right are drawings of two cylinders filled to the same level with water. The cylinders are identical in size and shape.

    Also shown at the right are two marbles, one glass and one steel. The marbles are the same size but the steel one is much heavier than the glass one.

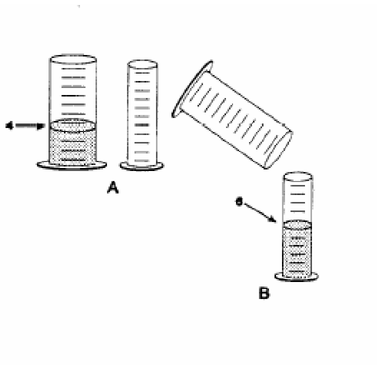
    When the glass marble is put into Cylinder 1 it sinks to the bottom and the water level rises to the 6th mark. *If we put the steel marble into Cylinder 2, the water will rise*

    a. to the same level as it did in Cylinder 1
    b. to a higher level than it did in Cylinder 1
    c. to a lower level than it did in Cylinder 1

4. *because*

    a. the steel marble will sink faster.
    b. the marbles are made of different materials.
    c. the steel marble is heavier than the glass marble.
    d. the glass marble creates less pressure.
    e. the marbles are the same size.

5. To the right are drawings of a wide and a narrow cylinder. The cylinders have equally spaced marks on them. Water is poured into the wide cylinder up to the 4th mark (see A). This water rises to the 6th mark when poured into the narrow cylinder (see B).



Both cylinders are emptied (not shown) and water is poured into the wide cylinder up to the 6th mark. *How high would this water rise if it were poured into the empty narrow cylinder?*

a. to about 8
b. to about 9
c. to about 10
d. to about 12
e. none of these answers is correct

6. *because*

a. the answer can not be determined with the information given.
b. it went up 2 more before, so it will go up 2 more again.
c. it goes up 3 in the narrow for every 2 in the wide.
d. the second cylinder is narrower.
e. one must actually pour the water and observe to find out.

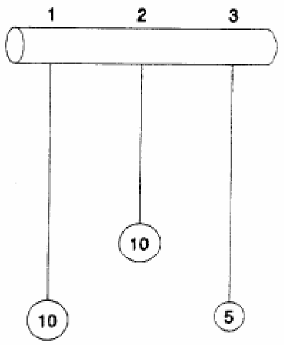
7. Water is now poured into the narrow cylinder (described in Item 5 above) up to the 11th mark. *How high would this water rise if it were poured into the empty wide cylinder?*

a. to about 7 1/2
b. to about 9
c. to about 8
d. to about 7 1/3
e. none of these answers is correct

8. *because*

a. the ratios must stay the same.
b. one must actually pour the water and observe to find out.
c. the answer can not be determined with the information given.
d. it was 2 less before so it will be 2 less again.
e. you subtract 2 from the wide for every 3 from the narrow.

9. At the right are drawings of three strings hanging from a bar. The three strings have metal weights attached to their ends. String 1 and String 3 are the same length. String 2 is shorter. A 10 unit weight is attached to the end of String 1. A 10 unit weight is also attached to the end of String 2. A 5 unit weight is attached to the end of String 3. The strings (and attached weights) can be swung back and forth and the time it takes to make a swing can be timed.



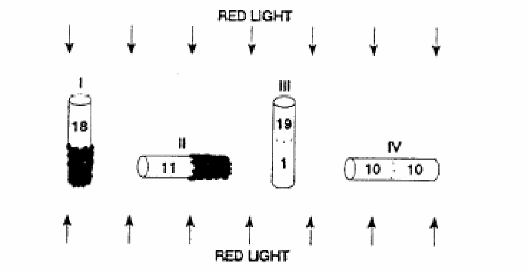
Suppose you want to find out whether the length of the string has an effect on the time it takes to swing back and forth. *Which strings would you use to find out?*

a. only one string
b. all three strings
c. 2 and 3
d. 1 and 3
e. 1 and 2

10. *because*

a. you must use the longest strings.
b. you must compare strings with both light and heavy weights.
c. only the lengths differ.
d. to make all possible comparisons.
e. the weights differ.

11. Twenty fruit flies are placed in each of four glass tubes. The tubes are sealed. Tubes I and II are partially covered with black paper; Tubes III and IV are not covered. The tubes are placed as shown. Then they are exposed to red light for five minutes. The number of flies in the uncovered part of each tube is shown in the drawing.



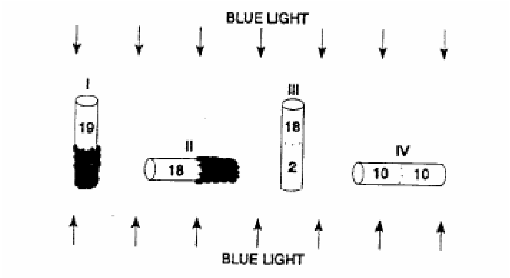
This experiment shows that flies respond to (respond means move to or away from):

a. red light but not gravity
b. gravity but not red light
c. both red light and gravity
d. neither red light nor gravity

12. *because*

a. most flies are in the upper end of Tube III but spread about evenly in Tube II.
b. most flies did not go to the bottom of Tubes I and III.
c. the flies need light to see and must fly against gravity.
d. the majority of flies are in the upper ends and in the lighted ends of the tubes.
e. some flies are in both ends of each tube.

13. In a second experiment, a different kind of fly and blue light was used. The results are shown in the drawing.



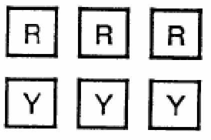
These data show that these flies respond to (respond means move to or away from):

a. blue light but not gravity
b. gravity but not blue light
c. both blue light and gravity
d. neither blue light nor gravity

14. *because*

a. some flies are in both ends of each tube.
b. the flies need light to see and must fly against gravity.
c. the flies are spread about evenly in Tube IV and in the upper end of Tube III.
d. most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
e. most flies are in the upper end of Tube I and the lighted end of Tube II.

15. Six square pieces of wood are put into a cloth bag and mixed about. The six pieces are identical in size and shape, however, three pieces are red and three are yellow. Suppose someone reaches into the bag (without looking) and pulls out one piece. *What are the chances that the piece is red?*

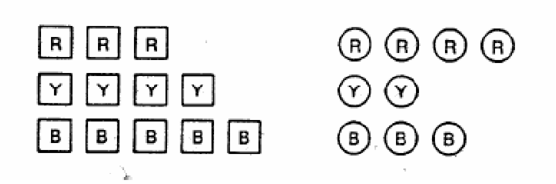


a. 1 chance out of 6
b. 1 chance out of 3
c. 1 chance out of 2
d. 1 chance out of 1
e. cannot be determined

16. *because*

    a. 3 out of 6 pieces are red.
    b. there is no way to tell which piece will be picked.
    c. only 1 piece of the 6 in the bag is picked.
    d. all 6 pieces are identical in size and shape.
    e. only 1 red piece can be picked out of the 3 red pieces.

17. Three red square pieces of wood, four yellow square pieces, and five blue square pieces are put into a cloth bag. Four red round pieces, two yellow round pieces, and three blue round pieces are also put into the bag. All the pieces are then mixed about. Suppose someone reaches into the bag (without looking and without feeling for a particular shape piece) and pulls out one piece.



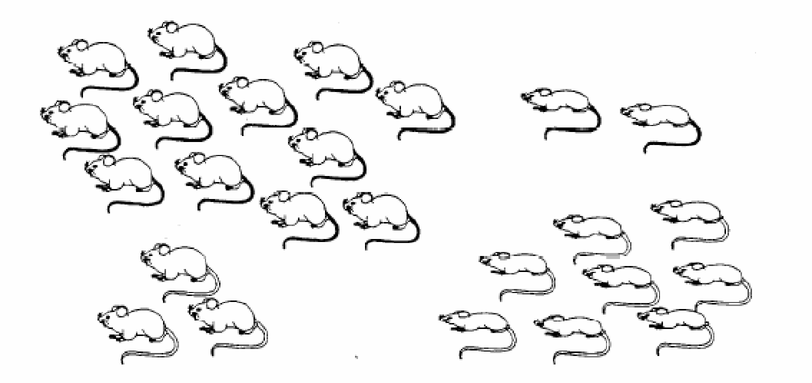
    What are the chances that the piece is a red round or blue round piece?

    a. cannot be determined
    b. 1 chance out of 3
    c. 1 chance out of 21
    d. 15 chances out of 21
    e. 1 chance out of 2

18. *because*

    a. 1 of the 2 shapes is round.
    b. 15 of the 21 pieces are red or blue.
    c. there is no way to tell which piece will be picked.
    d. only 1 of the 21 pieces is picked out of the bag.
    e. 1 of every 3 pieces is a red or blue round piece.

19. Farmer Brown was observing the mice that live in his field. He discovered that all of them were either fat or thin. Also, all of them had either black tails or white tails. This made him wonder if there might be a link between the size of the mice and the color of their tails. So he captured all of the mice in one part of his field and observed them. Below are the mice that he captured.



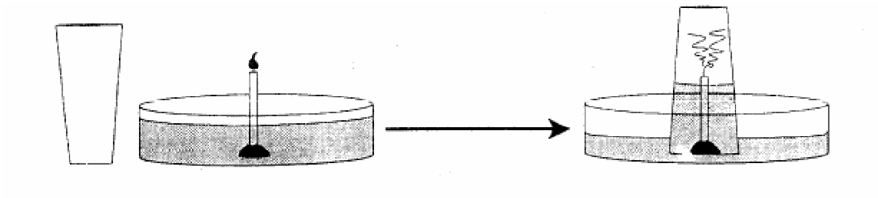
Do you think there is a link between the size of the mice and the color of their tails?

a. appears to be a link
b. appears not to be a link
c. cannot make a reasonable guess

20. *because*

a. there are some of each kind of mouse.
b. there may be a genetic link between mouse size and tail color.
c. there were not enough mice captured.
d. most of the fat mice have black tails while most of the thin mice have white tails.
e. as the mice grew fatter, their tails became darker.

21. The figure below at the left shows a drinking glass and a burning birthday candle stuck in a small piece of clay standing in a pan of water. When the glass is turned upside down, put over the candle, and placed in the water, the candle quickly goes out and water rushes up into the glass (as shown at the right).



This observation raises an interesting question: Why does the water rush up into the glass?

Here is a possible explanation. The flame converts oxygen into carbon dioxide. Because oxygen does not dissolve rapidly into water but carbon dioxide does, the newly formed carbon dioxide dissolves rapidly into the water, lowering the air pressure inside the glass.

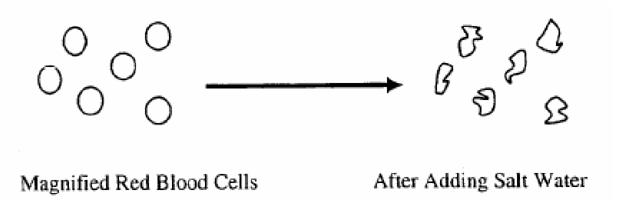
Suppose you have the materials mentioned above plus some matches and some dry ice (dry ice is frozen carbon dioxide). *Using some or all of the materials, how could you test this possible explanation?*

a. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
b. The water rises because oxygen is consumed, so redo the experiment in exactly the same way to show water rise due to oxygen loss.
c. Conduct a controlled experiment varying only the number of candles to see if that makes a difference.
d. Suction is responsible for the water rise, so put a balloon over the top of an open-ended cylinder and place the cylinder over the burning candle.
e. Redo the experiment, but make sure it is controlled by holding all independent variables constant; then measure the amount of water rise.

22. What result of your test (mentioned in #21 above) would show that your explanation is probably wrong?

a. The water rises the same as it did before.
b. The water rises less than it did before.
c. The balloon expands out.
d. The balloon is sucked in.

23. A student put a drop of blood on a microscope slide and then looked at the blood under a microscope. As you can see in the diagram below, the magnified red blood cells look like little round balls. After adding a few drops of salt water to the drop of blood, the student noticed that the cells appeared to become smaller.



This observation raises an interesting question: Why do the red blood cells appear smaller?

Here are two possible explanations: I. Salt ions (Na+ and Cl-) push on the cell membranes and make the cells appear smaller. II. Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

To test these explanations, the student used some salt water, a very accurate weighing device, and some water-filled plastic bags, and assumed the plastic behaves just like red-blood-cell membranes. The experiment involved carefully weighing a water-filled bag, placing it in a salt solution for ten minutes, and then reweighing the bag.

What result of the experiment would best show that explanation I is probably wrong?

a. the bag looses weight
b. the bag weighs the same
c. the bag appears smaller

24. *What result of the experiment would best show that explanation II is probably wrong?*

a. the bag looses weight
b. the bag weighs the same
c. the bag appears smaller

Please:

Do ***not*** *write anything on the questionnaire.*
Circle ***only one*** *answer per item on this answer sheet.*
Do ***not*** *skip any question.*
Avoid guessing. Your answers should reflect what ***you*** *personally think.*

For your convenience, please remove this answer sheet and circle your response for each item below (only circle one answer per item).

1. A B C
2. A B C D E
3. A B C
4. A B C D E
5. A B C D E
6. A B C D E
7. A B C D E
8. A B C D E
9. A B C D E
10. A B C D E
11. A B C D
12. A B C D E
13. A B C D
14. A B C D E
15. A B C D E
16. A B C D E
17. A B C D E
18. A B C D E
19. A B C
20. A B C D E
21. A B C D E
22. A B C D
23. A B C
24. A B C

Student ID: ____________________

Your Student ID number will be removed from this score sheet once all of the data are collected and paired using a random number de-identification system.

APPENDIX G:
Key to TSR

Classroom Test of Scientific Reasoning
Answer Key: Multiple Choice Version
Revised August 2000

1. B
2. D
3. A
4. E

5. B
6. C
7. D
8. A

9. E
10. C
11. B
12. A

13. C
14. D
15. C
16. A

17. B
18. E
19. A
20. D

21. A
22. A
23. A
24. B

Classroom Test of Scientific Reasoning
Answer Key: Free Response Version
Revised August 2000

1. B
2. Clay has not been added or taken away.
3. A
4. The marbles are both the same size, so they will displace the same amount of water.

5. 6
6. 4/6 = 6/x 4x = 36 x=9
   Note: Students do not have to use this method to be considered correct. Any indication of proportional rather than additive reasoning is acceptable.
7. 7 1/3
8. 4/6 = 2/3 = x/11 6x = 44 x = 44/6x = 7 1/3

9. 1 and 2
10. Everything is the same except the length, so you can tell if length makes a difference.
11. B
12. Most flies are in the upper end of Tube III but spread about evenly in Tube II.

13. C
14. Most flies are in the lighted end of Tube II but do not go down in Tubes I and III.
15. 1 chance out of 2.
16. 3 out of 6 pieces are red.

17. 1 chance out of 3.
18. 7 out of 21 (1 out of 3) pieces is a red or blue round piece.
19. A
20. Most of the fat mice have black tails, while most of the thin mice have white tails.

21. Saturate the water with carbon dioxide and redo the experiment noting the amount of water rise.
22. The water rises the same as it did before.
23. Weigh a water-filled bag in a salt solution for ten minutes and then reweigh the bag.
24. Explanation I would be wrong if the bag looses weight. Explanation II would be wrong if the bag stays the same weight.

Classroom Test of Scientific Reasoning
Reasoning Patterns Assessed
Revised August 2000

1. conservation of weight
2.
3. conservation of displaced volume
4.
5. proportional thinking
6.
7. advanced proportional thinking
8.
9. identification and control of variables
10.
11. identification and control of variables and probabilistic thinking
12.
13. identification and control of variables and probabilistic thinking
14.
15. probabilistic thinking
16.
17. advanced probabilistic thinking
18.
19. correlational thinking (includes proportions and probability)
20.
21. hypothetico-deductive thinking
22.
23. hypothetico-deductive reasoning
24.

APPENDIX H:
The Mechanics Language Usage Instrument (MLU) Pre-Test

Please **do not** remove this sheet from the questionnaire.

Please select the class year that most accurately describes your academic progress:

☐ Freshman
☐ Sophomore
☐ Junior
☐ Senior

In general, how confident are you in your knowledge of physics concepts?

☐ 5 (high degree of confidence)
☐ 4
☐ 3
☐ 2
☐ 1 (low degree of confidence)

What is your age?

☐ under 20
☐ 21 - 30
☐ 31+

What is your gender?

☐ Female
☐ Male

If you have declared a major, please indicate what it is here: ___________________________

Please indicate the classes you have taken before by checking one **or more** of the appropriate boxes:

☐ High School Physical Science
☐ High School Physics
☐ College Physical Science
☐ College Physics

Student ID: ____________________

Your Student ID number will be removed from this sheet once all of the data are collected and paired using a random number de-identification system.

This page intentionally left blank.

[This blank page is the backside of the first page of the MLU. This permitted the Student ID to be removed without loss of data once each survey was assigned a random non-identifying participant number.]

Please **write legibly** and use **complete sentences** for the free response items.

Free Response 1

The driver of an automobile accidentally backs her vehicle into a brick wall of a building. Although no significant damage to the brick wall occurs, the car *is* significantly damaged. For analysis, a security camera records the car's sudden stop.

To the best of your ability, how do you account for the car coming to rest and becoming damaged while the wall remained motionless and undamaged?

Free Response 2

During a soccer game, a soccer player breaks free from the other players and scores a goal with an impressive kick. In the excitement of the moment, two comments are made:

Announcer 1: "WOW! That ball had a lot of force!"

Announcer 2: "I agree, it would be hard to stop a ball with that much momentum."

Are these statements in agreement with one another?

☐Yes ☐No ☐Difficult to tell

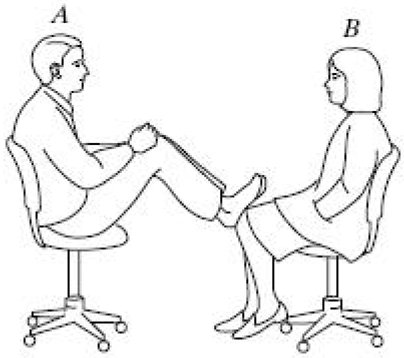
In your own words, briefly explain what you think the announcers mean by their statements.

In your own words, what is another way to accurately say what Announcers 1 and 2 mean by their statements?

MULTIPLE CHOICE

1. Imagine a head-on collision between a large truck and a small compact car. Both are traveling at the same speed before they collide. Which of the following statements is/are true about the car and the truck *due to their collision*? (There may be more than one, pick all that apply)

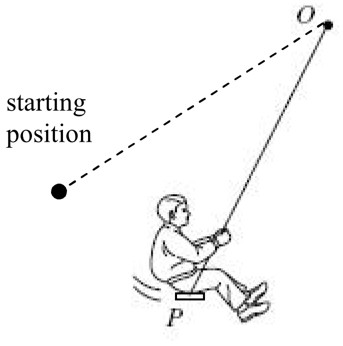
(A) The truck has more force than the car.
(B) The truck has more energy than the car.
(C) The truck has more momentum than the car.
(D) The truck has more power than the car.
(E) The truck has more strength than the car.



2. In the figure at right, student "A" has a mass of 95 kg and student "B" has a mass of 77 kg. They sit in identical office chairs facing each other. Initially, both students are at rest.

Student "A" places his bare feet on the knees of student "B," as shown. Student "A" then suddenly pushes outward with his feet, causing both chairs to move. While moving away from student "A" *and still in contact* with student "A," what property(ies) does student "B" have that she did not have before? (There may be more than one, pick all that apply)

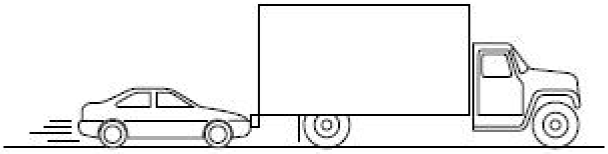
(A) Student "B" now has force.
(B) Student "B" now has energy.
(C) Student "B" now has momentum.
(D) Student "B" now has power.
(E) Student "B" now has strength.



3. The figure to the right shows a boy swinging on a rope, starting at a point higher than "P." What has increased since the instant the boy began the swing? (There may be more than one, pick all that apply)

(A) The boy's power has increased.
(B) The boy's momentum has increased.
(C) The boy's energy has increased.
(D) The boy's force has increased.
(E) The boy's strength has increased.

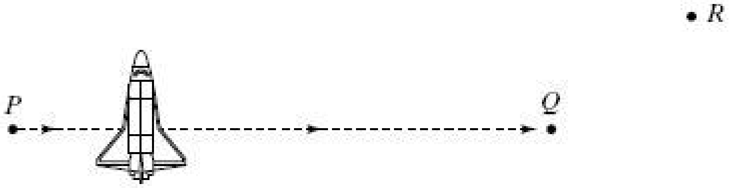
4. A large truck breaks down out on the road and receives a push back in to town by a small compact car as shown in the figure below.



The small compact car has to push hard to get the truck moving due to which of the following *physical properties of the truck*? (There may be more than one, pick all that apply)

(A) The truck's energy.
(B) The truck's force.
(C) The truck's mass.
(D) The truck's momentum.
(E) The truck's power.
(F) The truck's strength.

5. A rocket drifts sideways in outer space from point "P" to point "Q" as shown below. Starting at position "Q", the rocket's engine is turned on and produces a constant thrust (push on the rocket) at right angles to the line "PQ" (upward). This thrust is the only thing acting on the rocket from point "Q" to point "R." The constant thrust is maintained until the rocket reaches the point "R" in space.



The resulting path of the rocket due to the thrust from point "Q" to point "R" will be a direct result of the application of (There may be more than one, pick all that apply)

(A) force.
(B) momentum.
(C) power.
(D) strength.
(E) energy.

APPENDIX I:
The Mechanics Language Usage Instrument (MLU) Post-Test

Please **do not** remove this sheet from the questionnaire.

In general, how confident are you in your knowledge of physics concepts, having nearly completed the course?

- ☐ 5 (high degree of confidence)
- ☐ 4
- ☐ 3
- ☐ 2
- ☐ 1 (low degree of confidence)

In general, how much understanding of physics concepts do you feel you've developed having taken this course of first semester physics?

- ☐ 5 (understand a lot more physics concepts than before taking this class)
- ☐ 4
- ☐ 3
- ☐ 2
- ☐ 1 (understand the same amount of physics concepts than before taking this class)

If next semester, you were asked by another student to tutor them for this course, how confident do you think you would feel in your ability to successfully tutor them on the material of this course? Disregard any shyness or any tutoring experience you may already have.

- ☐ 5 (high degree of confidence; comfortable enough with the material to tutor)
- ☐ 4
- ☐ 3
- ☐ 2
- ☐ 1 (low degree of confidence; not comfortable enough with the material to tutor)

Please note: Regardless of how you respond above, you will not be contacted to be a physics tutor due to your participation in this study.

Do you intend to take additional physics coursework?

- ☐ Yes
- ☐ No

Student ID: ___________________

Your Student ID number will be removed from this sheet once all of the data are collected and paired using a random number de-identification system.

This page intentionally left blank.

[This blank page is the backside of the first page of the MLU. This permitted the Student ID to be removed without loss of data once each survey was assigned a random non-identifying participant number.]

Please **write legibly** and use **complete sentences** for the free response items.

Free Response 1

The driver of an automobile accidentally backs her vehicle into a brick wall of a building. Although no significant damage to the brick wall occurs, the car *is* significantly damaged. For analysis, a security camera records the car's sudden stop.

To the best of your ability, how do you account for the car coming to rest and becoming damaged while the wall remained motionless and undamaged?

Free Response 2

During a soccer game, a soccer player breaks free from the other players and scores a goal with an impressive kick. In the excitement of the moment, two comments are made:

Announcer 1: "WOW! That ball had a lot of force!"

Announcer 2: "I agree, it would be hard to stop a ball with that much momentum."

Are these statements in agreement with one another?

☐Yes ☐No ☐Difficult to tell

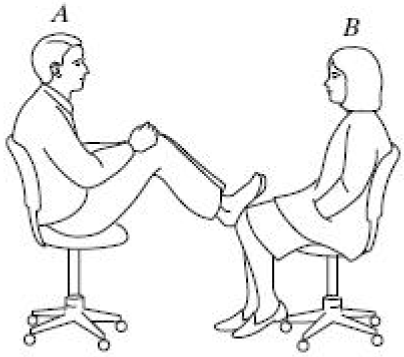
In your own words, briefly explain what you think the announcers mean by their statements.

In your own words, what is another way to accurately say what Announcers 1 and 2 mean by their statements?

MULTIPLE CHOICE

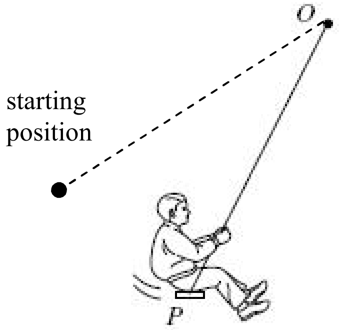
1. Imagine a head-on collision between a large truck and a small compact car. Both are traveling at the same speed before they collide. Which of the following statements is/are true about the car and the truck *due to their collision*? (There may be more than one, pick all that apply)
(A) The truck has more force than the car.
(B) The truck has more energy than the car.
(C) The truck has more momentum than the car.
(D) The truck has more power than the car.
(E) The truck has more strength than the car.

2. In the figure at right, student “A” has a mass of 95 kg and student “B” has a mass of 77 kg. They sit in identical office chairs facing each other. Initially, both students are at rest.

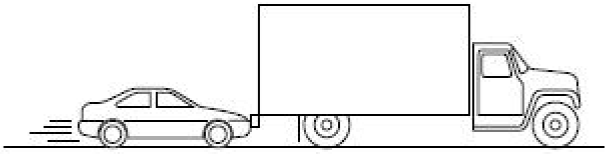


Student “A” places his bare feet on the knees of student “B,” as shown. Student “A” then suddenly pushes outward with his feet, causing both chairs to move. While moving away from student “A” *and still in contact* with student “A,” what property(ies) does student “B” have that she did not have before? (There may be more than one, pick all that apply)
(A) Student “B” now has force.
(B) Student “B” now has energy.
(C) Student “B” now has momentum.
(D) Student “B” now has power.
(E) Student “B” now has strength.

3. The figure to the right shows a boy swinging on a rope, starting at a point higher than “P.” What has increased since the instant the boy began the swing? (There may be more than one, pick all that apply)
(A) The boy’s power has increased.
(B) The boy’s momentum has increased.
(C) The boy’s energy has increased.
(D) The boy’s force has increased.
(E) The boy’s strength has increased.

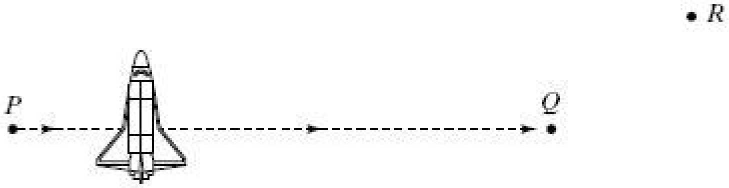


4. A large truck breaks down out on the road and receives a push back in to town by a small compact car as shown in the figure below.



The small compact car has to push hard to get the truck moving due to which of the following *physical properties of the truck*? (There may be more than one, pick all that apply)
(A) The truck's energy.
(B) The truck's force.
(C) The truck's mass.
(D) The truck's momentum.
(E) The truck's power.
(F) The truck's strength.

5. A rocket drifts sideways in outer space from point "P" to point "Q" as shown below. Starting at position "Q", the rocket's engine is turned on and produces a constant thrust (push on the rocket) at right angles to the line "PQ" (upward). This thrust is the only thing acting on the rocket from point "Q" to point "R." The constant thrust is maintained until the rocket reaches the point "R" in space.



The resulting path of the rocket due to the thrust from point "Q" to point "R" will be a direct result of the application of (There may be more than one, pick all that apply)
(A) force.
(B) momentum.
(C) power.
(D) strength.
(E) energy.