The development of elementary and middle school teacher science knowledge instruments for the evaluation of a professional development program

Jacob Burgoon

The University of Toledo

Follow this and additional works at: http://utdr.utoledo.edu/theses-dissertations

Recommended Citation

Burgoon, Jacob, "The development of elementary and middle school teacher science knowledge instruments for the evaluation of a professional development program" (2008). Theses and Dissertations. 1175.
http://utdr.utoledo.edu/theses-dissertations/1175

This Thesis is brought to you for free and open access by The University of Toledo Digital Repository. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of The University of Toledo Digital Repository. For more information, please see the repository's About page.
A Thesis

Entitled

The Development of Elementary and Middle School Teacher Science Knowledge

Instruments for the Evaluation of a Professional Development Program

By

Jacob Burgoon

Submitted as partial fulfillment of the requirements for

The Master of Science in Biology

Advisor - Patricia Komuniecki

College of Graduate Studies

The University of Toledo

May 2008
Professional development programs are important in helping teachers to obtain the knowledge and skills that are necessary to overcome students’ lack of achievement in science. Effective measures of teachers’ science knowledge are essential for successfully evaluating the programs’ impact on teachers’ knowledge. This study explores the development of science knowledge instruments for elementary and middle school teachers participating in the second cohort of a professional development program called NWO-TEAMS (Teachers Enhancing Achievement in Mathematics and Science). The instruments that were used for cohort one of the program were found to be too easy and thus not able to assess the effectiveness of the program. New instruments were created to be more difficult by using Bloom’s taxonomy and increasing the effectiveness of the
items’ distracters. The second year instruments included more items with effective
distracters and more items that measured higher order cognitive abilities. As a result, the
second year instruments were better able to separate teachers based on their science
knowledge and every grade level in the second cohort demonstrated significant increases
in science knowledge on the posttests. The development of the instruments in this study
is presented as a model for the evaluation of professional development programs which
seek to improve teachers’ science knowledge.
Acknowledgements

First, I’d like to thank my wife, Rebecca, for her constant support and encouragement throughout this entire process. Next, I’d like to thank my committee: Dr. Emilio Duran, Dr. Christine Fox, Dr. Patricia Komuniecki, and Dr. John Plenefisch for their suggestions and guidance that helped to make this thesis better. I’d also like to thank Dr. Mandy Heddle with whom I collaborated to create the instruments used in this study. Next, I’d like to thank Toni Sondergeld and Xi Zhang who assisted in the analysis of the items. Lastly, I’d like to thank the entire NWO-TEAMS staff whose hard work and pursuit of perfection resulted in the creation and delivery of a successful professional development program.
# Table of Contents

Abstract ii

Acknowledgements iv

Table of Contents v

List of Figures vi

List of Tables vii

I. INTRODUCTION 1

   The State of Education in America 1

   Science Inquiry 3

   The Impact of Teacher Competency on Student Learning 6

   Professional Development 8

   Importance of Evaluation 10

   Purpose of the Study 11

II. METHODS 15

III. RESULTS 26

IV. DISCUSSION 40

V. REFERENCES 55

VII. Appendix I: Science Indicators Taught During Summer Institute I 61

VIII. Appendix II: Example of a Table of Specifications 64

IX. Appendix III: Modifying an Item’s Cognitive Level 65

X. Appendix IV: Linking Multiple-Choice and Open-ended Items 67

XI. Appendix V: List of Abbreviations Used in this Study 68

XII. Appendix VI: Life Sciences Item from the Third Grade AY Instrument 69
LIST OF FIGURES

Figure 1: Mean pretest and posttest scores for year one of NWO-TEAMS ........... 27

Figure 2: Change in percentages of items measuring higher and lower order cognitive abilities between the first and second year instruments .......................... 30

Figure 3: Correlation between earned point percentage and cognitive level ........ 31

Figure 4: Correlation between earned point percentage and percentage of distracters not chosen .......................................................... 34

Figure 5: Changes in earned point percentage for lower order, higher order, and total item difficulties from year one to year two ................................. 36

Figure 6: Mean pretest and posttest scores for year two of NWO-TEAMS ........... 39

Figure 7: NWO-TEAMS science knowledge instrument development model ........ 54
## LIST OF TABLES

Table 1: Summary of analyses of year one science knowledge instruments .......... 26

Table 2: Percentages of first year test items measuring Bloom’s levels of cognitive ability ................................................................. 28

Table 3: Percentages of second year test items measuring Bloom’s levels of cognitive ability ................................................................. 29

Table 4: Higher order item scores for year two instruments ...................... 32

Table 5: Lower order item scores for year two instruments ...................... 32

Table 6: Description of distracter effectiveness for the first and second year instruments ................................................................. 33

Table 7: Number of open-ended, multiple-choice, and total items on the first and second year instruments .................................................. 37

Table 8: Range of pretest scores from the first and second year instruments ...... 38

Table 9: Summary of analyses of year two science knowledge instruments ........ 38

Table 10: Misconceptions held by NWO-TEAMS teacher participants ............. 47
I. INTRODUCTION

The State of Education in America

The lack of student achievement in science has been a national and international concern for over 30 years. Academic performance, as measured by the National Assessment of Educational Progress (NAEP), steadily declined after the administration of the first assessment in 1969. The publication of A Nation at Risk in 1983 by the National Commission on Excellence in Education (NCEE) increased public awareness of the educational crisis and pushed educational reform to the top of the policy agenda (NCEE, 1983). Reform efforts based on suggestions made by the NCEE helped to increase student achievement scores during the late 1980’s. The improvements, however, were short lived and scores reached a plateau in the early 1990’s. Among the recommendations made by the NCEE for educational improvement were the adoption of rigorous academic standards, the improvement of teacher quality through teacher preparation, and the improvement of academic curricula.

The suggestion for the adoption of academic standards began the movement known as standards-based reform. The idea behind this reform is to create standards that express what students should know and be able to do after formal education. National efforts to create a system of standards-based accountability resulted in the publication of national academic standards (AAAS, 1993; NRC, 1996). This was followed by a
dramatic increase in state efforts and by 2001 49 states had adopted academic standards (Achieve, 2002). However, generally the quality of the standards written by most states is regarded as poor; they are either intentionally vague or overly extensive (Finn & Kanstroom, 2001). For example, a state science standards review by the Fordham Institute gave 22 states a letter grade of D or F regarding the quality of their science standards (Gross, 2005). However, the same review states that the weaknesses in the standards are easily correctable. Collaboration with bench scientists and a careful re-evaluation of science content and desired outcomes should help most states to create standards that effectively address the initial goal of accountability.

Although student achievement scores have improved since the 1980’s, today they only reflect a basic understanding of science. For example, only 29% of 4th and 8th grade students and 18% of 12th grade students achieved a “proficient” score on the 2005 NAEP science assessment (ACC, 2007). Proficiency is characterized by the ability to apply knowledge to real-world situations and demonstrate analytic skills and mastery of subject-matter knowledge (Brown, 2000). Further evidence of a lack of proficiency comes from NAEP long-term assessment scores that show that from 1977 to 1999, less than half of 17-year-olds possessed the skills to analyze scientific procedures and data (Campbell et al., 2000).

International assessment scores also demonstrate a deficiency in critical thinking skills. In 2003, American students scored below the international average on the Program of International Student Assessment (PISA), which assesses the ability of 15-year old students to apply scientific and mathematical knowledge to real-world situations (NSB, 2006). PISA assesses three aspects of scientific literacy: 1) scientific knowledge
or concepts, which is assessed by application to subject matter; 2) scientific processes; and 3) situations or context, in which science-based issues are used to assess knowledge (OECD, 2003). Poor performance on the PISA is associated with low proficiency on national assessments because critical thinking skills are measured on both tests.

**Science Inquiry**

As a result of students’ inability to apply and analyze scientific knowledge, curricular and instructional changes have been made to focus on the process of science and its real-world applicability. Emphasis on the process of science requires that students play a more central role in the learning of science by actively engaging in science investigations, and teachers spend less time simply presenting information, a practice associated with the so-called traditional method of science teaching. The National Science Education Standards promote the active learning of science and describe the ideal learning environment in which “students describe objects and events, ask questions, acquire knowledge, construct explanations of natural phenomena, test those explanations in many different ways, and communicate their ideas to others” (NRC, 1996). Advocates of inquiry-based teaching de-emphasize the importance of older learning methods including memorization of vocabulary and facts, and instead focus on methods of doing science that promote students’ conceptual understanding of science (Mestre & Cocking, 2002).

Learning science by doing science has been a theme in educational reform since the 1920’s. The Progressive reform movement in the 1920’s and 1930’s revolved around the inquiry philosophy popularized by John Dewey (Cremin, 1961), and the inquiry-
based reform movement of the late 1950’s and 1960’s emphasized the importance of classroom investigations that were similar to those done by professional scientists (Dow, 1991). Another era of inquiry-based reform began in the early 1990’s. This current reform movement has a larger following and more support than the others, and with its prevalence in state and national standards (see also AAAS, 1993) and research literature, science inquiry has become the national standard for teaching science. The National Science Teaching Standards state that “inquiry into authentic questions generated from student experiences is the central strategy for teaching science” (NRC, 1996). In addition, the importance of inquiry-based teaching methods can be demonstrated by the fact that most state standards include science inquiry as a separate standard independent of any other content (Gross, 2005).

The active learning process of science inquiry is supported by research and literature about how students learn science. Research about science misconceptions held by students (Driver et al., 1994) demonstrates that students are often reluctant to give up their prior conceptions about science because they function well in the real world (Glynn et al., 1991; Smith, 1991). Presenting students with real-world examples that refute their prior conceptions can help them to overcome and discard their inaccurate conceptions. However, if students are not presented with plausible alternatives they may go back to believing their old conceptions (Smith, 1991). Inquiry-based lessons allow teachers to create a learning environment where students can develop scientifically accurate ideas, thus allowing them to fully discard their prior conceptions. In this regard, science inquiry may be more beneficial to students than simply reading textbooks or listening to lectures in that instead of merely taking the teacher’s word for it, they are given the opportunity to
validate their new conceptions for themselves through experimentation (see Ball, 1991 for an explanation of student knowledge validation as a result of reasoning through mathematical quandaries). Classroom science investigations are beneficial to student learning because they encourage students to pursue their natural curiosity. Even before formal schooling, children try to make sense of their environment by creating their own personal theories that explain natural phenomena; even infants have the ability to identify patterns and learn about some physical and biological concepts (NRC, 2000).

Although science inquiry has been widely popularized, debate about how much inquiry should take place in a classroom and what the best methods of inquiry are exist. Project-based instruction is an example of an inquiry-based teaching method in which students are provided with opportunities to investigate real-world questions, make predictions, design experiments, communicate with fellow students, and draw conclusions (Blumenfeld, 1991; Krajcik, 1994). This type of “open inquiry” puts science investigations entirely in the hands of the students, allowing them to build knowledge that can be applied to other situations (Blumenfeld, 1991). Educating students with this approach helps them to understand science practice and has been regarded as a practical way to achieve national science literacy goals (O’Neill & Polman, 2004). However, this inquiry method in particular has been argued to be ineffective in classroom settings. It has even been proposed that science educators abandon this method to focus on “genuine problems” (Settlage, 2007). In addition, some science educators believe that the science standards place too much emphasis on inquiry and not enough on disciplinary science content, such as astronomy, biology, physics, and chemistry (Gross, 2005; Finn & Kanstoroom, 2001). Gross (2005) argues that inquiry will not be effective if children do
not first have an acquired knowledge base. This debate presents a misconception that effective science teaching comes from the use of one teaching method only. Science inquiry cannot simply be viewed as all-or-nothing. Science education reformers need to embrace the middle ground by helping teachers use multiple teaching methods to develop their students’ science knowledge and abilities (NRC, 1996).

**The Impact of Teacher Competency on Student Learning**

The competency of classroom teachers is an important factor in the success of students. The National Commission on Teaching and America’s Future (1996) argued that “what teachers know and can do is the most important influence on what students learn”. In order for students to achieve the success expected by state governments and local school boards, they need to be taught by professionals who are fully competent in subject-matter knowledge and able to motivate, encourage, and facilitate student learning. The importance of teacher competency is exemplified by a study comparing low-achieving and high-achieving elementary schools, which found that over 90% of the variance in student achievement could be explained by differences in teacher qualifications (Armour-Thomas *et al*., 1989). Despite the known importance of high quality teachers, thousands of under qualified teachers currently teach in schools throughout the nation. In fact, 40 states allow their school districts to hire teachers that have not met basic requirements (NCTAF, 1996). Furthermore, Darling-Hammond (2000) showed that teacher characteristics, such as certification status and having a degree in the subject, have a positive impact on student achievement. However, the number of fully certified science teachers declined from 1990 to 2002, and in 2000, 23%
of 7th-12th grade science teachers did not have a major or a minor in the subject they were teaching (NSB, 2006). In addition, many teachers currently teach subjects that are outside their area of expertise. Out-of-field teaching is most likely the result of a shortage of qualified teachers. When the demand for highly qualified teachers surpasses the supply, school districts are forced to fill teaching positions with less qualified teachers who have not had the pedagogical and subject-matter training that highly qualified teachers have had (Howard, 2003). Educational policies that mandate the use of new curricula, standards, or assessments are meaningless without competent teachers who are able to effectively use them in the classroom (NCTAF, 1996). Because teachers exist as a link between policy and practice (Cohen and Hill, 1998), it is critical to invest in teacher education to ensure that they have the skills required to meet national teaching demands.

The national emphasis on science inquiry has created many challenges for classroom teachers, who frequently lack the knowledge and skills in specific content areas to deliver the challenging instructional approaches called for by the standards (Fuhrman, 2003). Inquiry science lessons are designed to give students more freedom than traditional lessons, so teachers must be prepared to answer student questions about subject-matter that may be slightly beyond the scope of the original lesson. For this reason, inquiry-based teaching requires deeper and broader subject-matter knowledge than traditional teaching (Fishman, 2003). Teachers who are less competent in subject-matter knowledge may actually be harmful to their students by passing on inaccurate ideas or uncritically using or inappropriately altering textbooks (Ball & McDiarmid, 1990). Furthermore, teachers who are less competent in subject-matter knowledge may
have misconceptions similar to those held by their students. For example, many teachers have misconceptions about the motion of objects, seasonal changes, and aggregate changes of matter (Kikas, 2004). Misconceptions, in students and teachers alike, can come from different sources including everyday experiences (Yip, 1998), textbooks (Barrass, 1984), and improper usage of analogies (Kikas, 2004). It is unlikely that teachers with science misconceptions will be able to help their students to overcome misconceptions. An understanding of teacher subject-matter knowledge is needed for teacher educators to create professional development programs that specifically address the teachers’ needs.

### Professional Development

Professional development programs provide teachers with opportunities to acquire and improve professional skills such as subject matter knowledge and pedagogical knowledge. Researchers (e.g., Fishman, 2003; Mizell, 2003) agree that the first step in designing professional development programs should be to identify areas where students need improvement. For example, national data on student achievement reveal that students lack the ability to apply and analyze scientific data. Since this ability is improved by inquiry-based curricula and teaching methods, professional development programs should train teachers to effectively use inquiry-based methods. Currently, many professional development programs do, in fact, aim to increase teachers’ subject matter knowledge and inquiry-based teaching skills (e.g., Marx, 2004; Lotter, 2006; Supovitz, 2000) due to a demand for teachers able to use inquiry-based teaching methods. The need for highly competent teachers can be satisfied by better preparation of pre-
service and in-service teachers by increasing the quality of college courses and professional development programs, respectively. The importance of professional development is reflected by current state policies. In 2002, 48 states had policies that required teachers to attend professional development to renew their teaching license (NSB, 2006). Linking professional development to student achievement is an important but difficult task for professional development evaluators. Many studies do not directly assess the effect of teacher professional development on student achievement (Loucks-Horsley & Matsumoto, 1999). In fact, of the 450 studies reviewed by the Middle Grades Initiative program, over 90 percent did not include any measure of student achievement (Killion, 1998). This may be because professional development programs do not directly affect students; they directly affect teachers by improving some aspect of their teaching practice and students are indirectly affected by the improvement in the teachers (Guskey & Sparks, 1996).

Professional development program design differs depending on the purpose of the program and the targeted population. Programs may take many forms (e.g., seminars, summer institutes, workshops) and cover a variety of topics (e.g., subject-matter, teaching methods, student learning). Reformers almost unanimously agree that programs sustained over a long period of time have a greater impact than short-term programs, such as one or two-day workshops. For example, one professional development program that lasted for six weeks focused on content knowledge and inquiry teaching and produced long-lasting increases in teachers’ attitudes, preparation, and use of inquiry-based teaching methods (Supovitz, 2000). Further, programs that provide teachers
opportunities for “hands-on” work and focus on subject matter and student learning are more likely to enhance knowledge and skills (Kennedy, 1996; Garet et al., 2001).

**Importance of Evaluation**

Evaluation of professional development programs is an important factor in long-term program success since reflection on the results of the program contributes to the program’s continuous improvement (Loucks-Horsley et al., 1998). Many Science, Technology, Engineering, and Mathematics (STEM) education programs lack scientifically rigorous evaluations that result in evidence of their effectiveness. In fact, the Department of Education’s Institute of Education Science’s What Works Clearinghouse reviewed 75 middle school mathematics programs and found that only three had well-designed experimental studies that led to strong evidence of their effectiveness (ACC, 2007). Furthermore, despite the efforts of federal and state funding agencies to provide educational programs to improve student learning, the Academic Competitiveness Council states:

> It is unclear which programs or activities are effective in generating positive outcomes. While many ideas have been tested in small case studies, few have been evaluated at the necessary scale to prove their efficacy for a broad range of students in an array of instructional settings. Without such evidence, it is nearly impossible for educators or administrators to know which activities, curricula, or materials to use to achieve the results that our nation demands (ACC, 2007).

The first step in evaluating the effectiveness of professional development should be to assess teacher learning because improvement in teacher knowledge and skills result
in increases in student achievement (see Marx et al., 2004). However, evaluating teacher learning (specifically content knowledge) in science and math is difficult, due to the lack of effective assessment instruments (Basile, 2006). Most tests that assess teacher knowledge use a multiple-choice format, which is “not very useful for assessing teachers’ ability to analyze and apply knowledge” (Darling-Hammond, 2000). Open-ended items (e.g., essay items) are better than multiple-choice items at assessing higher order thinking skills like organization, integration, and application of knowledge (Gronlund, 2003; Kubiszyn & Borich, 2003), but the use of these items on teacher tests is rare due to grading difficulties (Kubiszyn & Borich, 2003). The alignment of an assessment to course curriculum is also an important factor in an instrument’s effectiveness. General knowledge assessments published by national centers are commonly used in the evaluation of professional development programs. However, content varies between programs, so the assessments may not align with the program’s curriculum, thus not adequately measuring what is being taught. Locally developed assessments that are directly targeted at course content result in more accurate demonstrations of teacher knowledge as compared to nationally developed assessments that are not directly tied to course content (Basile et al., 2006).

Purpose of the Study

This study focuses on the development of assessment instruments designed to improve the quality of professional development for elementary and middle school science teachers teaching grades three through six. The instruments were created to assess the subject-matter knowledge of third through sixth grade science teachers
participating in a professional development program called NWO-TEAMS (Northwest Ohio Teachers Enhancing Achievement in Mathematics and Science). Before this particular study was done, instruments had already been developed to assess the knowledge of teachers participating in the first year of the NWO-TEAMS program. This study examines the effectiveness of developing science knowledge instruments for the second year of the NWO-TEAMS program by using Bloom’s taxonomy as a theoretical framework coupled with a rigorous evaluation of the quality of the previously used test items.

The Taxonomy of Educational Objectives (Bloom et al., 1956), commonly called Bloom’s taxonomy, is comprised of three domains – cognitive, affective, and psychomotor. The cognitive domain is used to classify educational objectives and, in this study, was used to classify test items; the items were classified by the cognitive ability they measured. The cognitive domain (hereafter referred to as Bloom’s taxonomy) is hierarchical and is comprised of six levels – knowledge, comprehension, application, analysis, synthesis, and evaluation.

The knowledge level describes objectives/items that require students to remember information such as specific facts, principles, methods, and theories. Since the recall of facts is the cognitive ability required, this level was renamed “remember” when the Taxonomy Handbook was revised (Anderson et al., 2001).

The comprehension level describes objectives/items that require students to translate, interpret, or extrapolate information. For example, a comprehension level item might require students to describe scientific principles in their own words or use information from a table or graph to answer questions. These two levels, knowledge and
comprehension, comprise what will be termed the “lower order” cognitive abilities. Test items that measure content at these two levels are the most common items found on the majority of science content tests.

The application level describes objectives/items that require students to apply knowledge to new situations. Objectives/items at this level may require students to identify which scientific principle should be used to find the solution to a given problem, or explain a scientific phenomenon in terms of scientific principles. In order to answer items at this level, students must first have knowledge and comprehension of the scientific principles in question.

The analysis level describes objectives/items that require students to break down ideas and find relationships between concepts. For example, analysis level objectives for reading a scientific paper may include identifying the hypotheses and discussing the theoretical framework on which it is based (if it is not delineated in the paper).

The synthesis level describes objectives/items that require students to create something new, such as a story incorporating scientific principles or a protocol for a lab experiment.

Lastly, the evaluation level describes objectives/items that require students to make judgments about a body of work, such as a scientific paper. For example, after reading a scientific paper, students might be asked to judge whether or not the author did a good job of presenting their argument, including identifying the author’s point of view and pointing out any underlying assumptions that the author failed to recognize. The previous four levels, application, analysis, synthesis, and evaluation, comprise what will be termed the “higher order” cognitive abilities.
The research question explored in this study was: “What are the most effective methods of developing science knowledge instruments for elementary and middle school teachers?” The specific aims for this study included demonstrating 1) whether the cognitive ability that an item measures affects the difficulty of the item, 2) whether the effectiveness of an item’s distracters affect the difficulty of that item, and 3) whether instruments that include a balanced number of “higher order” and “lower order” items with effective distracters are better able to separate teachers based on their science knowledge.
II. METHODS

*NWO-TEAMS*

NWO-TEAMS is a three year long (2006-2009) grant funded professional development program that aims to increase the content knowledge and inquiry-based teaching skills of elementary and middle school teachers in northwest Ohio. The topics covered during the program are aligned with state and national standards, and were chosen by a local focus group comprised of curriculum experts and experienced teachers who selected the concepts that teachers have the most reluctance and/or difficulty teaching (e.g., electricity, physical/chemical changes, forces and motion).

Physical, Earth, and Life Science topics are taught over the course of the program with lessons designed to align with Ohio grade-level indicators. The teachers are separated into groups by grade-level (third through sixth) and receive instruction in the selected topics from facilitators, who are experienced teachers familiar with inquiry-based teaching methods, and scientists. The facilitator/scientist team teaching model (National Science Resources Center, 1997; Ballone-Duran *et al.*, 2005) allows teachers to experience science instruction that models how they should teach science in their own classrooms. Also, by allowing the teachers to ask the scientists complex scientific questions, the teaching model helps the teachers to develop the deep subject-matter knowledge that is needed to successfully use inquiry-based teaching methods. Program instruction is based on modified OSCI (Ohio Science Institute) modules, which are
science inquiry lessons, supplemented by inquiry-based science kits (e.g., FOSS, STC), which are effective in preparing teachers to use inquiry-based methods and increasing their students’ science achievement (Mangrubang, 2004; Young & Lee, 2005).

NWO-TEAMS is comprised of three sessions over a year long period: Summer Institute I (SI-I), Academic Year (AY), and Summer Institute II (SI-II). Teachers are provided a total of 168 professional development hours throughout the year. During SI-I, teachers attend eight full days of science instruction that includes lessons that cover concepts in Physical and Earth Sciences, including non-contact forces, weathering and erosion, electricity, and physical and chemical changes. During the AY phase, teachers attend eight monthly sessions (September to April) covering concepts in Life, Physical, and Earth Sciences. During SI-II, teachers attend four days of instruction in Life, Physical, and Earth Sciences as well as educational field trips to the Toledo Zoo and Fossil Park in Sylvania, Ohio. Teachers are encouraged to use these area resources in their own classrooms to enhance their students’ learning experience.

Different indicators are taught during each phase of the program to maximize the amount of science content that the teachers receive. At the conclusion of the program, teachers will have received instruction in 11 to 17 Ohio grade-level indicators, depending on which grade they teach. At the conclusion of SI-I, which is the focus of this study, teachers will have received instruction in four to ten indicators (see Appendix I).

**Participants**

The participants in this study include two cohorts of in-service elementary and middle school science teachers involved in NWO-TEAMS. The first NWO-TEAMS
cohort was comprised of 65 teachers, including 14 third grade, 22 fourth grade, 16 fifth grade, and 11 sixth grade teachers. There were 20 different school districts represented, including two that are currently rated “academic watch” by the Ohio Department of Education. School districts that are rated “academic watch” do not meet the state’s performance standards and thus are important to be represented in NWO-TEAMS because their students are the most in need. The second NWO-TEAMS cohort was comprised of 64 teachers, including 17 third grade, 20 fourth grade, 17 fifth grade, and 10 sixth grade teachers. There were 22 different school districts represented, including two that are currently rated “academic watch”.

**Content Test Development**

Content tests for each group (third grade through sixth grade) were created for every phase of the program to assess the program’s effect on teachers’ science knowledge. This study specifically focuses on the development of the Summer Institute one (SI-I) tests for cohorts one and two. The SI-I content tests for cohort one were created using items from state achievement tests and locally developed classroom tests. Items were chosen from achievement tests from Texas, California, Oregon, and Ohio, among others, that aligned with the grade-level indicators on which the instructional content was based. Developing the tests in this way ensured that the teachers would not be tested on a concept that was not covered during the program. After the administration of the first year tests, a software program called ClearStat (Stone, 2003) was used to analyze the test items. ClearStat provides information such as item difficulty according the Rasch model (Rasch, 1960), item point biserials (i.e., discrimination), the proportion
of teachers answering the item correctly, and the proportion of teachers choosing each of the item’s distracters. Analysis revealed that many of the items were too easy for the teacher participants, and for some items, improvement on the posttest was impossible because all of the teachers answered correctly on the pretest. Upon further analysis, two major factors were identified that lead to the instruments’ lack of difficulty. First, many of the multiple-choice items included distracters that were not chosen by any teachers, thus not effectively “distracting” teachers from the correct response. Second, most of the test items measured lower order cognitive abilities, such as knowledge and comprehension.

Due to their lack of difficulty, the SI-I content tests were extensively modified before the second year of the program. Most of the previously used items taken from student achievement tests were replaced with new items that aligned with the program content. However, some items from the first year instruments were changed and used on the second year instruments. Changes included rewording the item stems, which are the parts of the items where the question is posed, so teachers would better understand what the question was asking. Changes also included reformatting items on the first year tests that measured lower order cognitive abilities so they measured higher order cognitive abilities on the second year tests.

The content tests for cohort two were developed to establish higher difficulty by using Bloom’s taxonomy, tables of specifications, lesson plans from the first year of the program, and literature about science misconceptions.
Bloom’s Taxonomy, Tables of Specifications, and Lesson Plans

Bloom’s taxonomy was used as a guide in the creation of “higher order” items that require teachers to apply scientific knowledge learned during the program. The reasoning behind the use of Bloom’s taxonomy was twofold: 1) because higher order items require more critical thinking skills than lower order items, the addition of these items will increase the content tests’ difficulty, and 2) because the teachers are instructed with the use of inquiry-based teaching methods, the use of application items could evaluate the program’s ability to develop critical thinking skills.

Tables of specifications were used during the development of the tests to ensure the items were aligned with the instructional content and that lower order items did not comprise the majority of the test, as was the case with the first year tests. A table of specifications serves as a blueprint for a test in the form of a two-dimensional table, with content on one side, and behavior or skill on the other. In this study, behavior was defined by the cognitive levels of Bloom’s taxonomy. The use of these tables ensures that a wide range of content is represented in the set of items, as well as higher order cognitive abilities (Notar et al., 2004). See Appendix II for an example of a table of specifications.

Lesson plans from the first year of the program were attained from the facilitators and consulted during test development. Some of the grade-level indicators contain a number of more specific concepts not taught during the program. Newly written test items were compared to the lesson plans to make sure that the teachers were only being tested on the concepts taught during the program.
**Misconceptions**

Research literature about student and teacher misconceptions supplemented the use of Bloom’s taxonomy during test development. For the purposes of this study, science misconceptions were defined as ideas or concepts that are not scientifically accurate, but are commonly held by teachers before and after science instruction. One factor contributing to the first tests’ lack of difficulty was the use of ineffective distracters for multiple-choice items.

In order to increase the effectiveness of the distracters for the new tests, student and teacher misconceptions previously identified in literature were used as distracters for some multiple-choice items. The purpose of doing this was to create distracters that seemed more plausible to teachers, therefore distracting those who do not possess the knowledge to correctly answer the item. Some of the open-ended items were also based on student or teacher misconceptions previously identified by research. The purpose of developing these items was to identify teacher misconceptions that were similar to those held by students. Other open-ended items were written to explore concepts about which little is known regarding student or teacher misconceptions. The open-ended items provided participants with the opportunity to explain the reasoning behind their answers, which would help us to better understand teachers’ knowledge base for a particular concept. In addition, several open-ended items were aligned with multiple-choice items that assessed the same concept, so we would be able to explore the reasoning behind the participants’ answers to the multiple-choice items.
Data Analysis

Grading rubrics for the open-ended items from the cohort one SI-I content tests were developed by the science facilitators. Answer keys and rubrics for the cohort two SI-I content tests were developed by graduate students, scientists, and science educators.

The results of the first year instruments were analyzed by comparing the pretest and posttest means using dependent t-tests. Effect sizes were represented by Cohen’s $d$, where $d = .2$ is a small effect size, $d = .5$ is a medium effect size, and $d = .8$ is a large effect size (Cohen, 1988). The effect size shows how far apart the pretest mean is from the posttest mean in standard deviation units and demonstrates how important the difference is between the two means. Effect size, in contrast to statistical significance, is not affected by sample size so it provides a more practical interpretation of the difference between the means.

In order to answer the research question for this study, three hypotheses were explored. The first hypothesis was that items that measured higher order cognitive abilities would be more difficult than those that measured lower order cognitive abilities. Based on this hypothesis, it was predicted that instruments that contain a balanced number of higher and lower order items would be more difficult than instruments that contain a majority of lower order items.

The number of items measuring higher and lower order cognitive abilities was compared from the first to the second year. Then the mean item difficulty of the first year tests was compared to mean item difficulty of the second year tests using an independent t-test. Lastly, a point biserial correlation was calculated to find if there was a relationship between item cognitive level and item difficulty. The effect size for this...
test was represented by $r^2$, which provided the percentage of variance in item difficulty that could be accounted for by the cognitive level that the item measured.

Difficulty was represented in this study by the percentage of total possible points earned by the teacher participants; this percentage is termed the earned point percentage (EPP). For example, if an item was worth five points and there were 20 teachers answering the item, there would be 100 total possible points that the teachers could earn. If the teachers collectively earned 40 points, the EPP would be 40%. Therefore, lower earned point percentages were associated with more difficult items because fewer teachers answered the items correctly. Because the multiple-choice items were worth one point, 20 teachers answering an item would represent 20 total possible points, so the proportion of teachers answering correctly would equal the earned point percentage. Therefore, the proportion of teachers answering correctly (given by the ClearStat item analysis) was used to determine the earned point percentages of the multiple-choice items. The proportion of teachers answering correctly could not be used to calculate the earned point percentages of the open-ended items because the dichotomy of correct versus incorrect could not be applied to items worth multiple points for which teachers could potentially answer an item partially correct. Therefore, for each open-ended item, the total possible points were calculated by multiplying the number of teachers answering the item by the number of points the item was worth. Then, all of the teachers’ scores for the item were added together and divided by the total possible points.

In addition to increasing instrument difficulty, the higher order items developed for the second year tests were also used to assess whether the NWO-TEAMS program effectively helped the teacher participants to acquire critical thinking skills. In order to
find whether teachers acquired critical thinking skills during the program, higher and lower order item scores were calculated for each teacher from their total scores. The pre and posttest scores for both higher and lower order items were then compared using dependent t-tests.

The second hypothesis was that multiple-choice items with effective distracters would be more difficult than those with ineffective distracters. Based on this hypothesis, it was predicted that instruments comprised of “effective” multiple-choice items would be more difficult than instruments that were comprised of “ineffective” items. Teachers who do not have the knowledge required to answer an item should essentially have to guess between all of the answer options in a multiple-choice item. If the distracters do not make sense or are not plausible, teachers could disregard those distracters, thus increasing their chances of correctly guessing the answer. The ClearStat item analysis software provided the percentage of teachers who chose each of an item’s answer options, and ineffective distracters were seen as the options that were chosen by none of the teacher participants. Therefore, the more ineffective distracters a multiple-choice item possessed, the more ineffective the item would be. So, distracter effectiveness was represented in this study as the percentage of an item’s distracters that were not chosen by any teacher participants. For example, if an item had three distracters and two were not chosen by any teachers, the percentage of distracters not chosen would be 67% for that item. Therefore, the higher the percentage of distracters not chosen, the more ineffective the distracters were, and vice versa.

The percentage of distracters not chosen was calculated for every multiple-choice item on the first and second year tests, and the number of items in each of the percentage
categories (e.g., 100%, 67%, 33%, 0%) was compared from first to second year. To find if there was a relationship between distracter effectiveness and earned point percentage (item difficulty), a Pearson correlation was calculated. The effect size for this test was again represented as $r^2$, which provided the percentage of variance in item difficulty that could be explained by distracter effectiveness.

To reinforce the importance of distracter effectiveness, items on the first and second year tests were again separated into lower and higher order items. Mean difficulties for those item groups were found and compared from first to second year using independent t-tests. This test was performed because item/test difficulty (represented by EPP) was the dependent variable for this and the previous hypothesis, so cognitive level needed to be held constant to see the individual effect of distracter effectiveness on item difficulty.

The third hypothesis was that instruments with a balanced number of higher and lower order items and effective distracters are better able to separate teachers based on their science knowledge. Based on this hypothesis, it was predicted that the second year tests, because of the modifications that were made in cognitive level and distracter effectiveness, would have a higher range of scores than the first year tests. A frequency table of the final pretest scores was created for every instrument for the first and second years and the range was found by subtracting the lowest score from the highest.

Misconceptions that were held by teachers were identified by analyzing the responses to the open-ended items to look for incorrect answers that were common among the participants. Also, because misconceptions were used as distracters for multiple-choice items, pre and posttest multiple-choice items were analyzed to find the
proportion of teachers that chose each distracter. In addition to identifying
misconceptions in each of the two item formats, the tests were analyzed to find
connections between multiple-choice and open-ended items that assessed the same
concept.
III. RESULTS

This study began by analyzing the Summer Institute I instruments used during the first year of the program. The pretest and posttest scores for all of the grades (three through six) in cohort one were compared using dependent t-tests. No significant changes in test scores were seen for grades three and five, but a significant increase in test scores was seen in grades four and six. The mean scores, standard deviations, t values, and effect sizes for each grade are presented in Table 1. Figure 1 shows the pretest and posttest scores attained by cohort one on the SI-I tests.

Table 1. Summary of analyses of year one science knowledge instruments

The pretest and posttest means from the first year SI-I instruments were compared using dependent t-tests. Effect sizes were measured by Cohen’s $d$ and calculated by using the means and standard deviations of each grade’s tests.

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Total Possible Points</th>
<th>Mean Pretest Score ± SD</th>
<th>Mean Posttest Score ± SD</th>
<th>t</th>
<th>Effect Size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>14</td>
<td>21</td>
<td>18.0 ± 1.4</td>
<td>18.8 ± 1.0</td>
<td>1.55</td>
<td>0.66$^M$</td>
</tr>
<tr>
<td>4</td>
<td>20</td>
<td>21</td>
<td>15.5 ± 2.4</td>
<td>17.8 ± 1.4</td>
<td>5.88***</td>
<td>1.17$L$</td>
</tr>
<tr>
<td>5</td>
<td>20</td>
<td>22</td>
<td>15.4 ± 1.9</td>
<td>15.9 ± 3.7</td>
<td>0.63</td>
<td>0.17</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>26</td>
<td>18.3 ± 2.6</td>
<td>21.1 ± 1.5</td>
<td>4.37***</td>
<td>1.31$L$</td>
</tr>
</tbody>
</table>

*** p < .001, n = number of participants in each grade, L = large, M = medium
Using Bloom’s Taxonomy in Instrument Development

The first hypothesis tested in this study was that items that measured higher order cognitive abilities would be more difficult than those that measured lower order cognitive abilities. The tests used during the first year were made up primarily of items that measured lower order cognitive abilities. In fact, over 80% of the items on every instrument measured lower order cognitive abilities. Almost half of the items for each grade...
grade were classified as comprehension level items. Table 2 shows the percentages of first year test items that are classified in each of the cognitive levels.

**Table 2. Percentages of first year test items measuring Bloom’s levels of cognitive ability**

The percentages of items measuring each cognitive level were calculated for each grade by counting the number of items measuring each cognitive level and dividing by the instruments’ total number of items. The total percentages of items measuring each cognitive level were calculated by combining the number of items from all four grades that measure each cognitive level and dividing by the total number of items on the first year tests.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Cognitive Level</th>
<th>Total Low</th>
<th>Total High</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>38.9 (7)</td>
<td>55.6 (10)</td>
<td>5.6 (1)</td>
</tr>
<tr>
<td>4</td>
<td>30.0 (6)</td>
<td>50.0 (10)</td>
<td>15.0 (3)</td>
</tr>
<tr>
<td>5</td>
<td>29.4 (5)</td>
<td>52.9 (9)</td>
<td>17.7 (3)</td>
</tr>
<tr>
<td>6</td>
<td>41.2 (7)</td>
<td>47.1 (8)</td>
<td>11.7 (2)</td>
</tr>
<tr>
<td>Total</td>
<td>34.7 (25)</td>
<td>51.4 (37)</td>
<td>12.5 (9)</td>
</tr>
</tbody>
</table>

The numbers in parentheses represent the number of items measuring each cognitive level.

In order to balance the percentages of lower and higher order items, new items that measured higher order cognitive abilities were developed for the second year SI-I tests, and some of the lower order items were removed. In some cases, however, lower order items on the first year tests were re-written as higher order items on the second year tests. For example, item six on the third grade test for the first year was classified at the knowledge level. The item asked teachers to identify the force that pulls a skateboarder down a ramp (See Appendix III for the complete item). Item seven on the third grade test for the second year is the modified form of the previously described item and is classified at the analysis level. In order to change the item so it measured higher order cognitive
ability, the ramp from the first year item was divided into four sections and teachers were asked to choose the forces that acted on the skateboarder at each stage of his descent.

Some of the forces that the teachers could choose were gravity, friction, and air resistance (See Appendix III for the complete item).

After the modifications were made and new items were developed for the second year tests, the percentages of items measuring higher order cognitive abilities on every test had increased from the first year. The percentages of second year test items that are classified in each of the cognitive levels are presented in Table 3, and the change in the percentages between the first and second year tests are depicted in Figure 2.

Table 3. Percentages of second year test items measuring Bloom’s levels of cognitive ability

The percentages of items measuring each cognitive level were calculated for each grade by counting the number of items measuring each cognitive level and dividing by the instruments’ total number of items. The total percentages of items measuring each cognitive level were calculated by combining the number of items from all four grades that measure each cognitive level and dividing by the total number of items on the second year tests.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Cognitive Level</th>
<th>Total Low</th>
<th>Total High</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>7.7 (1)</td>
<td>38.5 (5)</td>
<td>46.1 (6)</td>
</tr>
<tr>
<td>4</td>
<td>17.4 (4)</td>
<td>43.5 (10)</td>
<td>39.1 (9)</td>
</tr>
<tr>
<td>5</td>
<td>18.7 (3)</td>
<td>31.3 (5)</td>
<td>43.8 (7)</td>
</tr>
<tr>
<td>6</td>
<td>0.0</td>
<td>50.0 (6)</td>
<td>33.3 (4)</td>
</tr>
<tr>
<td>Total</td>
<td>12.5 (8)</td>
<td>40.6 (26)</td>
<td>40.6 (26)</td>
</tr>
</tbody>
</table>

The numbers in parentheses represent the number of items on each instrument measuring each cognitive level
Based on the hypothesis, it was predicted that the second year tests would be more difficult because they contained a higher percentage of higher order items. The mean earned point percentage (EPP) for all of the first year items was found to be 78.7%. The mean EPP for the second year items was 46.0%, which is significantly lower (and therefore more difficult) than the first year items ($t = 8.06, p < .001$). Also, a significant correlation, depicted in Figure 3, was found between item cognitive level and EPP ($r = -0.408, p < .001$).

**Figure 2. Change in percentages of items measuring higher and lower order cognitive abilities between first and second year instruments**

The percentages of lower and higher order items were calculated by counting the number of items measuring lower and higher order cognitive abilities and dividing by the total number of items. The percentages were then compared from the first to the second year tests.

Bars one and two, three and four, five and six, and seven and eight represent the first and second year tests, respectively, and are connected by brackets, see Tables 3 and 4 for the number of items on each instrument measuring each cognitive level.
Another reason for using Bloom’s taxonomy in this study was to measure the teacher participants’ acquisition of critical thinking skills. Test items were separated into two categories of cognitive ability – lower order and higher order – in order to measure the program’s effect on developing those abilities. Significant increases between pretest and posttest “higher order item” scores were seen in all of the second year tests. Only the fifth and sixth grade tests showed significant increases between pretest and posttest “lower order item” scores. Tables 4 and 5 present a summary of mean higher order and lower order test scores with $t$ values and effect sizes.

**Figure 3. Correlation between earned point percentage and cognitive level**

Items from the first and second year instruments were assigned a cognitive level based on the cognitive ability that was measured by the item. Earned point percentages were calculated for each item. A point biserial correlation was calculated between cognitive level and earned point percentage.

![Figure 3](image)

$r = .408, p < .001$

$r^2 = .166$

$1 = $knowledge$, 2 = $comprehension$, 3 = $application$, 4 = $analysis$, 5 = $synthesis$, 6 = $evaluation$
Table 4. Higher order item scores for year two instruments

Higher order item scores were calculated by counting the points earned from higher order items. The pretest and posttest higher order item score means for each grade were compared using dependent t-tests. Effect sizes were measured by Cohen’s $d$ and calculated by using the means and standard deviations for each grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>$n_P$</th>
<th>$n_I$</th>
<th>Pretest Mean $\pm$ SD</th>
<th>Posttest Mean $\pm$ SD</th>
<th>$t$</th>
<th>Effect Size ($d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
<td>7</td>
<td>6.69 ± 2.79</td>
<td>9.42 ± 2.44</td>
<td>3.81**</td>
<td>1.04L</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>9</td>
<td>9.67 ± 3.38</td>
<td>12.22 ± 3.69</td>
<td>2.88**</td>
<td>0.72M</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>8</td>
<td>4.94 ± 1.71</td>
<td>8.24 ± 3.09</td>
<td>4.12***</td>
<td>1.32L</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6</td>
<td>3.29 ± 2.48</td>
<td>5.33 ± 2.37</td>
<td>2.32*</td>
<td>0.84L</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$, *** $p < .001$, $n_P$ is the number of teachers in each grade and $n_I$ is the number of higher order items on each instrument, L = large, M = medium.

Table 5. Lower order item scores for year two instruments

Lower order item scores were calculated by counting the points earned from lower order items. The pretest and posttest higher order item score means for each grade were compared using dependent t-tests. Effect sizes were measured by Cohen’s $d$ and calculated by using the means and standard deviations for each grade.

<table>
<thead>
<tr>
<th>Grade</th>
<th>$n_P$</th>
<th>$n_I$</th>
<th>Pretest Mean $\pm$ SD</th>
<th>Posttest Mean $\pm$ SD</th>
<th>$t$</th>
<th>Effect Size ($d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
<td>6</td>
<td>3.23 ± 1.74</td>
<td>3.61 ± 1.39</td>
<td>1.24</td>
<td>0.24S</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>14</td>
<td>9.94 ± 2.52</td>
<td>10.61 ± 2.09</td>
<td>1.24</td>
<td>1.18L</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>8</td>
<td>5.00 ± 1.17</td>
<td>5.65 ± 0.81</td>
<td>2.18*</td>
<td>0.65M</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6</td>
<td>5.96 ± 3.08</td>
<td>8.25 ± 2.24</td>
<td>2.81*</td>
<td>0.85L</td>
</tr>
</tbody>
</table>

* $p < .05$, $n_P$ is the number of teachers in each grade and $n_I$ is the number of lower order items on each instrument, L = large, M = medium, S = small.
Creation and Modification of Item Distracters

The second hypothesis tested in this study was that multiple-choice items with effective distracters would be more difficult than those with ineffective distracters. In this study, we looked at the overall “effectiveness” of an item’s distracters by measuring the percentage of distracters not chosen by teacher participants. The higher the percentage of the distracters not chosen, the more ineffective the item’s distracters were, and vice versa.

Table 6 presents the number of items from each year that were calculated to have 0%, 25%, 33%, 67%, and 100% of their distracters not chosen. Most of the multiple-choice items had four options – one correct answer and three distracters. There were some items on the second year tests that consisted of one correct answer and four distracters, which is why there are two items for which 25% of the distracters were not chosen. A significant correlation, depicted in Figure 4, was found between percentage of distracters not chosen and EPP ($r = .727$, $p < .001$).

**Table 6. Description of distracter effectiveness for the first and second year instruments**

The percentage of distracters not chosen was calculated by dividing the number of an item’s unchosen distracters by the item’s total number of distracters. The number of items in each percentage category were counted and compared from the first to the second year.

<table>
<thead>
<tr>
<th>Percentage of distracters not chosen</th>
<th>Number of Items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Year 1</td>
</tr>
<tr>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>33</td>
<td>14</td>
</tr>
<tr>
<td>67</td>
<td>19</td>
</tr>
<tr>
<td>100</td>
<td>14</td>
</tr>
</tbody>
</table>
The previous results in this study showed that both cognitive level and distracter effectiveness affect item difficulty. We wanted to find how distracter effectiveness alone affected test difficulty, so items were separated into higher and lower order items and their mean EPP’s were calculated and compared from the first to the second year. Both higher order and lower order items demonstrated a significant decrease in EPP from the first to the second year. For the first year tests, the mean EPP for the lower order items was 80.7% and the mean EPP for the higher order items was 52.9%. For the second year
tests, the mean EPP for the lower items was 68.9% and the mean EPP for the higher order items was 38.0%. Significant differences in this case could not be attributed to cognitive level because it remained constant for each group. Therefore, these data show that the modifications made to the tests between years one and two (increasing distracter effectiveness being the primary modification here) increased lower and higher order item difficulty as well as overall test difficulty. Figure 3 depicts the changes in earned point percentage for lower and higher order items from year one to year two (remember that lower EPP’s represent more difficult tests).

**Other Modifications to the Instruments**

Modifications were made to the SI-I instruments between the first and second years in an attempt to increase their effectiveness to assess teachers’ science knowledge. Some items from the first year tests were modified by changing the wording of the stem of the item. Other items from the first year tests were simply removed and replaced with new items. For the third grade, five items on the second test were modified from first year items. For the fourth grade, two items on the second test were carried forward with no changes from the first test, while six other items were modified from first year items. For the fifth grade, four of the items on the second test were modified from first year items. Lastly, for the sixth grade, all of the items on the second test were newly created; none were modified from first year items.

The number of items on every test, with exception of fourth grade, decreased from the first year to the second year. The number of open-ended items,
however, increased or remained the same for three of the four tests. The changes in the numbers of open-ended, multiple-choice, and total items can be seen in Table 7.

Figure 5. Changes in earned point percentage for lower order, higher order, and total item difficulties from year one to year two

Items on the first and second year instruments were separated into lower order and higher order items based on the cognitive ability that they measured. The mean earned point percentages for lower order, higher order, and total items were compared using independent t-tests.

*** p < .001, there were 58 lower order first year and 34 lower order second year items, 10 higher order first year and 29 higher order second year items, and 68 total first year and 63 total second year items, bars represent mean earned point percentages ± standard deviation
Table 7. Number of open-ended, multiple-choice, and total items on the first and second year instruments

The number of multiple-choice and open-ended items included in each instrument was counted and compared from the first to the second year.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Open-Ended</th>
<th>Multiple-Choice</th>
<th>Total</th>
<th>Open-Ended</th>
<th>Multiple-Choice</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>15</td>
<td>18</td>
<td>3</td>
<td>10</td>
<td>13</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>19</td>
<td>20</td>
<td>9</td>
<td>14</td>
<td>23</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>10</td>
<td>17</td>
<td>5</td>
<td>11</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>12</td>
<td>17</td>
<td>8</td>
<td>4</td>
<td>12</td>
</tr>
</tbody>
</table>

Use of Science Knowledge Instruments as Evaluation Tools

The third hypothesis tested in this study was that instruments with effective distracters and a balanced number of higher and lower order items are better able to separate teachers based on their science knowledge. In this study, the range of pretest scores earned by teacher participants determined how well the instrument was able to separate teachers. Total score frequencies for every grade’s test were compared from the first to the second year. In all but one of the tests, the range of scores increased. For the fifth grade test, the range slightly decreased from the first to second year. Table 8 shows the lowest, highest, and total score and the range of pretest scores for each of the SI-I instruments.
Table 8. Range of pretest scores from the first and second year instruments

The range for each instrument was calculated by subtracting the lowest score from the highest score. Ranges were calculated for each instrument and compared from the first to the second year.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Low Score</th>
<th>Highest Score</th>
<th>Total Score</th>
<th>Range</th>
<th>Low Score</th>
<th>Highest Score</th>
<th>Total Score</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>15</td>
<td>20</td>
<td>21</td>
<td>5</td>
<td>3.5</td>
<td>18.5</td>
<td>22</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>19</td>
<td>21</td>
<td>9</td>
<td>12</td>
<td>27</td>
<td>38</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>12</td>
<td>19</td>
<td>22</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>25</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>23</td>
<td>26</td>
<td>10</td>
<td>2</td>
<td>19.5</td>
<td>25</td>
<td>17.5</td>
</tr>
</tbody>
</table>

The pretest and posttest scores for all of the second year instruments were compared using dependent t-tests. Significant increases in test scores were found for all of the instruments. The mean scores, standard deviations, t values, and effect sizes for each grade are presented in Table 9. Figure 4 shows the pretest and posttest scores attained by cohort two on the SI-I tests.

Table 9. Summary of analyses of year two science knowledge instruments

The pretest and posttest means from the second year SI-I instruments were compared using dependent t-tests. Effect sizes were measured by Cohen’s d and calculated by using the means and standard deviations of each grade’s tests.

<table>
<thead>
<tr>
<th>Grade</th>
<th>n</th>
<th>Mean Pretest Score ± SD</th>
<th>Mean Posttest Score ± SD</th>
<th>t</th>
<th>Effect Size (d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>13</td>
<td>9.9 ± 4.04</td>
<td>13.0 ± 3.32</td>
<td>3.65**</td>
<td>0.84L</td>
</tr>
<tr>
<td>4</td>
<td>18</td>
<td>19.6 ± 4.84</td>
<td>22.8 ± 4.72</td>
<td>3.04**</td>
<td>0.67M</td>
</tr>
<tr>
<td>5</td>
<td>17</td>
<td>9.9 ± 1.92</td>
<td>13.9 ± 3.44</td>
<td>4.54***</td>
<td>1.41L</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>9.2 ± 5.31</td>
<td>13.6 ± 4.14</td>
<td>2.66*</td>
<td>0.91L</td>
</tr>
</tbody>
</table>

* p < .05, ** p < .01, *** p < .001, L = large, M = medium
Figure 6. Mean pretest and posttest scores for year two of NWO-TEAMS

Total test scores for each grade were calculated by counting the total points earned by the teacher participants. The pretest and posttest mean test scores from the first year SI-I instruments were then compared using dependent t-tests.

* p < .05, ** p < .01, *** p < .001, bars represent raw score means ± standard deviations, see Table 9 for the number of participants in each grade.
IV. DISCUSSION

This study was performed to find the most effective ways to develop science knowledge instruments for elementary and middle school teachers. This question was explored by testing the following hypotheses:

1) Test items that measure higher order cognitive abilities are more difficult than those that measure lower order cognitive abilities,

2) Multiple-choice items with effective distracters are more difficult than those with ineffective distracters, and

3) Instruments with a balanced number of higher and lower order items and effective distracters are better able to separate teachers based on their science knowledge.

Instrument Alignment

Before discussing the results of the tested hypotheses, it is necessary to first discuss the importance of aligning instruments with instruction. Alignment of an instrument to instructional content is important and necessary for that instrument to serve its designed purpose. Regarding the measurement of progress in standards-based reform, Barton (2004) argues that “if alignment is out of whack, there can be no confidence that changes in the [test] scores are a valid measure of accountability”. This alignment is also referred to as content validity, defined as the degree to which an instrument measures the content that it is designed to measure. There are three aspects of content validity: domain definition, domain representation, and domain relevance (Sireci, 1998a). Domain
definition involves defining the content and cognitive areas included in the content domain, while domain representation and domain relevance involves judging how well a test matches the content domain and how relevant the items are to the content domain (Sireci, 1998b). The first year instruments were created strictly by focusing on domain representation, without definition of the cognitive areas involved in program instruction. These instruments were developed by choosing items from state student achievement tests and locally developed student tests that measured the content taught during Summer Institute I. However, the chosen items were originally created to measure children’s knowledge, so although the tests were aligned with the program content, they were too easy for the teacher participants. The instruments’ lack of difficulty resulted in high pretest scores for teacher participants (see Figure 1), which left little room for improvement on the posttest. Non-significant increases in test scores for third and fifth grade teachers resulted from the closeness of their pretest and posttest scores. And although the fourth and sixth grade teachers showed significant increases in test scores, their mean pretest scores represented 70% or higher of the total possible points, reinforcing the instruments’ lack of difficulty.

The second year tests were created after considering the content and cognitive areas included in program instruction and were written to represent and be relevant to the content domain. In order to create effective measures of science knowledge, it is not enough to simply align the instruments with program instruction. It is also necessary to define cognitive areas included in the content domain and ensure that the items are reasonably difficult for the participants.
**Measuring Higher Order Cognitive Abilities**

Bloom’s taxonomy was used to as a guide for the creation of new items that measured higher order cognitive abilities. The reasoning behind the use of this method was twofold: 1) because higher order items require more critical thinking skills than lower order items, the addition of these items will increase the content tests’ difficulty, and 2) because the teachers are instructed with the use of inquiry-based teaching methods, the use of higher order items could evaluate the program’s ability to develop critical thinking or problem solving skills.

The results of the point biserial correlation between cognitive level and item difficulty supported the hypothesis that higher order items are more difficult then lower order items (see Figure 3). It was found that 17% of the variance in item difficulty could be explained by cognitive level (higher and lower order). Therefore, because the second year tests contained a higher percentage of higher order items than the first year tests (see Figure 2), it makes sense that the second year tests were also more difficult than the first year tests. The results also showed that the second year tests were effective in assessing the program’s ability to develop “meaningful learning” where teachers use the knowledge learned during the program to solve new problems (Anderson et al., 2001). All grade levels in the second cohort attained significant increases in higher order item scores (see Figure 6), meaning that the teacher participants developed knowledge and skills during SI-I that enabled them to correctly answer more higher order items on the posttest. Only the fifth and sixth grade teachers attained significant increases in lower order item scores. It was expected for there to be no significant increases in lower order item scores because during the first year, the teachers answered many of the lower order items correctly on
the pretests, leaving no room for growth on the posttests. However, the lack of significant increases for the third and fourth grade teachers was not due to high pretest scores (see Table 5). It can be concluded from these results that the lower order items on the second year tests were more difficult than those on the first year tests, most likely due to the creation of more effective distractors for second year items.

_Evaluation of Item Distracters_

The analysis of the effectiveness of item distracters was an important method for increasing the difficulty of the second year instruments. Constructing good multiple-choice items depends greatly on the creation of effective distracters (Gronlund, 2003). The Pearson correlation between item difficulty and percentage of distracters not chosen was significant (see Figure 4), which supported the hypothesis that items with effective distracters are more difficult than items with ineffective distracters. It was found that about 53% of the variance in item difficulty could be explained by the item’s percentage of distracters not chosen. In addition, Figure 5, which shows the changes in mean EPP for higher and lower order items from year one to year two, supports the conclusion that the lower order items on the second year tests are more difficult than those on the first year tests. Because lower order items from year one are being compared to lower order items from year two, the increase in difficulty could not have come from a change in cognitive level. Modification and creation of item distracters is the only other major change that was made to the lower order items, so it can be concluded that the use of effective distracters played a significant role in the increased difficulty of the lower order items on the second year tests.
In this study, distracter effectiveness was represented by the percentage of distracters not chosen. However, in addition to quantifying the number of unchosen distracters, the point biserial statistic associated with each distracter should be analyzed. This statistic, calculated by ClearStat, is a correlation between whether or not teachers chose a distracter and the teachers’ overall test score. Point biserials associated with effective distracters should be negative since teachers who choose the distracters should score lower on the overall test while teachers who do not choose the distracters should score higher on the overall test. In contrast, the point biserial associated with the correct answer should be positive since teachers who choose that response should also do well on the overall test. A distracter with a positive point biserial would mean that teachers who chose that distracter did well on the overall test. Therefore, that distracter somehow does not match the stem of the item and should be changed or removed. Due to the large number of items on the first year tests for which 100% of the distracters were not chosen, this study only focused on writing plausible distracters that more teachers would choose, thus lowering the percentage of distracters not chosen. However, future evaluations of the instruments will include analyses of the point biserials associated with each answer option.

Examples of items with ineffective distracters could easily be found on the first year instruments, which included 14 items for which none of the distracters were chosen (see Table 6). Teacher participants who do not have the knowledge required to answer a question should ideally have to guess between all of the answer options. However, the use of an ineffective distracter essentially increases the chances of teachers correctly guessing the answer by decreasing the number of logical options. One of Gronlund’s
(2003) 18 rules for writing effective multiple-choice items is to “make the distracters plausible and attractive to the uninformed”. One way that is suggested for doing this is to “use the common misconceptions or errors of students as distracters”. This is highly supportive of the method that was used in this study to increase the effectiveness of distracters. Misconceptions that have been researched for students and teachers were used as distracters for some of the multiple-choice items. When known misconceptions could not be found, effective distracters were developed by imagining what teachers would think if they had an incomplete understanding of the content being measured. These distracters could be called “predicted misconceptions” because even though these misconceptions have not been found in prior research, teachers would be predicted to answer this way if they did not fully understand the concept.

Including misconceptions as distracters not only makes the instruments more effective, but also guides the program’s curriculum development by identifying what specific concepts the teacher participants struggle with. If it is observed that a majority of teachers are choosing one distracter over another, that data can be used to inform the program facilitators who will then alter the program’s curriculum to address that specific misconception. For example, the responses to a multiple-choice question that asked teachers “what happens to water molecules when liquid water changes to vapor?” showed that many teachers thought that the molecules change by becoming lighter. The next time this concept is addressed, the results of this example will guide curriculum development to make sure that teachers understand that the water molecules are not altered during a phase change. In addition to using misconceptions as distracters for multiple-choice items, they were also used to guide the creation of open-ended items.
For the most part, these items were written to probe for new misconceptions. They were written in a way that gave teachers the opportunity to thoroughly explain their thought process about a certain concept. The responses for these items given by teachers provided insight into what teachers were thinking about a concept, and how they went about solving problems. These responses, like those of the multiple-choice items, can be used to guide program curriculum that will improve the quality of the program. In addition, several open-ended items were linked to multiple-choice items assessing the same concept. For example, item 19 from the fourth grade test was linked to the multiple-choice item mentioned above. Item 19 asked the teachers to compare the masses of three sealed beakers – one at room temperature, one frozen, and one that had been heated. Although the item stated that the room temperature and heated beakers were filled with the same amount of water, teachers answered that the heated beaker had least mass, supporting the results of the multiple-choice item (See Appendix IV for these items).

In this study, the use of misconceptions in creating test items lead to the discovery of several misconceptions held by teachers, some of which were previously identified in students or teachers, and some that had not been previously identified. It is especially important that the misconceptions that have been previously identified in students are corrected because teachers who hold those misconceptions will likely pass them on to their students and reinforce incorrect ideas about scientific concepts. Yip (1998) emphasizes that because teachers demonstrate some of the same errors as students, they are likely to be a direct source of student misconceptions. Table 10 presents the misconceptions identified in teachers during SI-I.
Table 10. Misconceptions about SI-I content held by NWO-TEAMS teacher participants

Items that probed for misconceptions were based on misconceptions that had already been identified in students, already been identified in teachers, or had not already been identified. Teacher responses were analyzed to see if teachers held the misconceptions.

<table>
<thead>
<tr>
<th>Previously identified in students</th>
<th>Previously identified in teachers</th>
<th>Previously unidentified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravity acts only on a body in motion</td>
<td>Chemical changes are not reversible</td>
<td>Gravity acts more when an object is further from the ground</td>
</tr>
<tr>
<td>All metals are attracted to magnets</td>
<td>Density is directly proportional to density and/or volume</td>
<td>Heat and pressure provide the energy stored in fossil fuels</td>
</tr>
</tbody>
</table>

**AY and SI-II Results**

The same methods that were used to develop science knowledge instruments for SI-I were also used to develop instruments for the AY and SI-II sessions. The data collected from these instruments provided more examples of teacher misconceptions, including misconceptions about Life Sciences concepts that were not taught during SI-I. Items on the AY third grade instrument measured the teachers’ knowledge of animal classification, life cycles, and adaptation. For example, one item required teachers to classify pictures of animals as vertebrates or invertebrates, and as mammals, birds, reptiles, amphibians, or fish (See Appendix VI). Nearly a quarter of the teachers classified a penguin as a mammal, a crab as a fish, and a sea turtle as an amphibian. In addition, over half of the teachers classified a snake and a sea turtle as invertebrates.

Items on the SI-II third grade instruments measured teachers’ knowledge of fossils, extinction, and adaptation, and the sixth grade instruments measured teachers’ knowledge of characteristics of animal cells and genetic inheritance. Results showed that
many sixth grade teachers lacked knowledge of how traits are inherited in asexually reproducing organisms.

_Evaluating the Effectiveness of NWO-TEAMS_

The science knowledge instruments in this study were developed to better assess the effectiveness of NWO-TEAMS and identify ways to increase the quality of the program. All grade levels in cohort two demonstrated significant increases in posttest scores (see Figure 4 and Table 9). Because program instruction did not change from year one to year two, it can be concluded that the significant increases in test scores are a result of the increased effectiveness of the instruments. The scores on the first year tests were all fairly high and thus not informative or representative of the teacher participants’ true science knowledge. Because the instruments were not difficult enough, teachers with low science knowledge acquired similar scores to teachers with high science knowledge. Therefore, determining the effectiveness of NWO-TEAMS based on the results from these tests would be inaccurate. The range of pretest scores for each test was used to determine the tests’ ability to separate teachers. If all of the teachers receive a similar score, the range will be small, and it can be deduced that the test is not effective at separating levels of science knowledge. A large range indicates there is a large difference between the highest and lowest scores, signifying that the test may be effective at separating levels of science knowledge. Between the first and second year tests, the range of pretest scores tripled for the third grade test and nearly doubled for the fourth and sixth grade tests (see Table 8). These data show that there is more variability in the pretest scores for the second year tests, which means that they are more able to identify
teachers who have different levels of science knowledge. The only test that did not show improvement in pretest score range was the fifth grade test, which actually decreased by one. A small range of scores may result when all of the teachers score highly, as was seen in the pretest scores for the first year instruments. However, a small range may also result when all of the teachers score poorly. It may be that the methods used in this study overcorrected the problems in the first year instruments, resulting in the fifth grade test for the second cohort being too difficult for the teacher participants. The items on the fifth grade test will be evaluated to identify possible reasons for the small range, and modified for the next cohort. A weakness of defining separation using range is that range only depends on two of the test scores (highest and lowest). This could potentially give an inaccurate picture of the dispersion of the test scores if most of the scores from a particular test are high but one score is unusually low. To ensure that was not the case for the results of this study, the frequency tables for each test were consulted before making any conclusions about the dispersion of the test scores, and it was found that the scores were evenly dispersed between the highest and lowest scores. An alternative method to define separation would be to use the Rasch measure of separation that would provide how many “levels” of teacher participants based on science knowledge could be identified by each test. A high number would indicate a test that effectively assesses science knowledge by identifying teachers at differing skills levels. Due to time constraints, this method was not used. However, future evaluation of science knowledge instruments for NWO-TEAMS will include Rasch separation.

Based on the results of second year tests, it can be concluded that they are not only useful for assessing the overall quality of NWO-TEAMS instruction, but are also
useful in showing which content areas need the most improvement. The results have shown that the instruments developed for the second cohort were appropriately challenging for the teacher participants, thus resulting in scores that were representative of actual science knowledge. Total scores could be broken down into Physical science and Earth Science scores, and dependent t-tests could be run to find the differences between pretest and posttest scores. Then, for example, if a significant difference was not found between the pretest and posttest Earth Science scores, the curriculum for that particular grade could be modified to better address the earth science indicators.

Conclusion

The results of this study are presented here as a model for the development of effective science knowledge instruments for professional development programs. The results have shown that developing instruments by focusing on cognitive abilities and common misconceptions is an effective way to assess the quality of professional development programs and inform decisions about curricular changes. Figure 7 presents the science knowledge instrument development model that was used for NWO-TEAMS.

The model is presented in three parts: the preparation phase, the continuous improvement cycle, and outcomes. The first step of the preparation phase is the definition of the content and cognitive domains. Defining the content domain involves determining the science content that will be measured with the instruments. The lesson plans developed by program facilitators provide the specific content that will be taught during the program, so consultation of these plans ensures that the created items accurately align to program content. In addition, the lesson plans also show how the
content is taught, which is necessary for the definition of the cognitive domain because ideally, instruments should measure program content at the same cognitive level as it was taught (Notar, 2004). Also, in order for test items to be classified as higher order, they need to measure the participants’ ability to apply their science knowledge to new situations that were not encountered during instruction; knowledge of how the content is taught guides the creation of these items. Tables of specifications are created during this step to ensure that the content and cognitive domains are accurately defined and aligned with the program’s instructional objectives.

The next step in the preparation phase is the creation of test items, which are written with the guidance of Bloom’s taxonomy to ensure that they measure the desired cognitive ability. Also, literature is searched for common misconceptions about the program content. These misconceptions can be used in the instruments as distracters for multiple-choice items or as a basis for open-ended items. As items are created, they are inserted into a table of specifications. When the instrument is completed, the table of specifications will show how many items measure each content and cognitive objective (see Appendix I for an example).

The next part of the model is the continuous improvement cycle, so named because the steps facilitate the improvement of the instruments and can be repeated as many times as the instrument is administered. The results are analyzed after each administration of the instrument to collect data such as the proportion of teachers answering each item correctly, the number of distracters not chosen for each multiple-choice item, and the range of total scores obtained by the participants. In this study,
ClearStat software was used to obtain these data, but other software can be used, or it can be done by hand, depending on the number of participants in the program.

Next, the data that come from the analyses are used to modify the instruments. For example, if the analyses show that there are several multiple-choice items that contain distracters that were not chosen by any of the participants, more plausible distracters should be considered before the items are used again. Also, if responses to an open-ended item are drastically different from what was expected, the item should possibly be reworded to make the intentions of the item clearer to the participants. The modified instruments can then be administered again to the next group of participants to begin the cycle again.

The last part of the model is the outcomes. The data that come from the analyses of the test results are used to inform the program staff about the quality of the program. Total scores are broken into several sub scores that represent each of the major content areas measured by the instrument, such as physical, earth, and life sciences. The participants’ scores in each of the content areas show the program staff which areas need the most work, thus leading to changes in the program’s curriculum. Also, the misconceptions that were identified by the instruments are presented to the facilitators so that those concepts can be better addressed for the next group. The curriculum is modified to meet the specific needs of the participants, thus improving the quality of the program by addressing concepts that are difficult to understand or commonly misunderstood. Therefore, continuous improvement of the instruments leads to continuous improvement of the program. The improvements in program quality will most likely lead to increases in the participants’ science subject matter knowledge, which
plays a central role in their effectiveness as teachers (Ball and McDiarmid, 1990; Kennedy, 1998). The use of the model, as this study shows, results in science knowledge instruments that effectively assess professional development programs.
Figure 7. NWO-TEAMS science knowledge instrument development model

**PREPARATION PHASE**

- Define content and cognitive domains
  - Tables of specifications
  - Lesson plans

- Create test items
  - Bloom’s taxonomy
  - Common misconceptions

**CONTINUOUS IMPROVEMENT CYCLE**

- Administer instruments

- Modify instruments based on results

- Analyze results

**OUTCOMES**

- Improvements in program quality
- Changes in curriculum
REFERENCES


Appendix I: Science Indicators Taught During Summer Institute I

Grade 3

*Physical Science*

1. Describe an object's position by locating it relative to another object or the background.
2. Describe an object's motion by tracing and measuring its position over time.
3. Identify contact/noncontact forces that affect motion of an object (e.g., gravity, magnetism and collision).
4. Predict the changes when an object experiences a force (e.g., a push or pull, weight and friction).

*Earth Science*

1. Compare distinct properties of rocks (e.g., color, layering and texture).
2. Observe and investigate that rocks are often found in layers.
3. Describe that smaller rocks come from the breakdown of larger rocks through the actions of plants and weather.
4. Observe and describe the composition of soil (e.g., small pieces of rock and decomposed pieces of plants and animals, and products of plants and animals).
5. Investigate the properties of soil (e.g., color, texture, capacity to retain water, ability to support plant growth).
6. Investigate that soils are often found in layers and can be different from place to place.

Grade 4

*Physical Science*

1. Identify characteristics of a simple physical change (e.g., heating or cooling can change water from one state to another and the change is reversible).
4. Explain that matter has different states (e.g., solid, liquid and gas) and that each state has distinct physical properties.

**Earth Science**

1. Explain that air surrounds us, takes up space, moves around us as wind, and may be measured using barometric pressure.

2. Identify how water exists in the air in different forms (e.g., in clouds, fog, rain, snow and hail).

3. Investigate how water changes from one state to another (e.g., freezing, melting, condensation and evaporation).

8. Describe how wind, water and ice shape and reshape Earth's land surface by eroding rock and soil in some areas and depositing them in other areas producing characteristic landforms (e.g., dunes, deltas and glacial moraines).

9. Identify and describe how freezing, thawing and plant growth reshape the land surface by causing the weathering of rock.

10. Describe evidence of changes on Earth's surface in terms of slow processes (e.g., erosion, weathering, mountain building and deposition) and rapid processes (e.g. volcanic eruptions, earthquakes and landslides).

**Grade 5**

**Physical Science**

1. Define temperature as the measure of thermal energy and describe the way it is measured.

2. Trace how thermal energy can transfer from one object to another by conduction.
3. Describe that electrical current in a circuit can produce thermal energy, light, sound and/or magnetic forces.

4. Trace how electrical current travels by creating a simple electric circuit that will light a bulb.

**Grade 6**

*Physical Science*

2. Describe that in a chemical change new substances are formed with different properties than the original substance (e.g., rusting, burning).

3. Describe that in a physical change (e.g., state, shape and size) the chemical properties of a substance remain unchanged.

5. Explain that the energy found in nonrenewable resources such as fossil fuels (e.g., oil, coal and natural gas) originally came from the sun and may renew slowly over millions of years.

6. Explain that energy derived from renewable resources such as wind and water is assumed to be available indefinitely.

7. Describe how electric energy can be produced from a variety of sources (e.g., sun, wind and coal).

8. Describe how renewable and nonrenewable energy resources can be managed (e.g., fossil fuels, trees and water).
Appendix II: Example of a Table of Specifications

Cohort Two: Grade 4

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Item Type</th>
<th>Bloom’s Taxonomy</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>PS 1</td>
<td>MC</td>
<td>(3)(7) (12)</td>
<td>3</td>
</tr>
<tr>
<td>PS 4</td>
<td>MC SA</td>
<td>(5)(9) (19)</td>
<td>3</td>
</tr>
<tr>
<td>ES 1</td>
<td>SA MC</td>
<td>(21) (22)</td>
<td>2</td>
</tr>
<tr>
<td>ES 2</td>
<td>SA</td>
<td>(20)</td>
<td>1</td>
</tr>
<tr>
<td>ES 3</td>
<td>MC SA</td>
<td>(2)(4) (1) (18)</td>
<td>4</td>
</tr>
<tr>
<td>ES 8</td>
<td>MC SA</td>
<td>(10) (8) (16)</td>
<td>4</td>
</tr>
<tr>
<td>ES 9</td>
<td>SA</td>
<td>(23)</td>
<td>1</td>
</tr>
<tr>
<td>ES 10</td>
<td>SA MC</td>
<td>(15) (6)(13) (11)(17)</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>4 10 9</td>
<td><strong>23</strong></td>
</tr>
</tbody>
</table>

PS = Physical Science
ES = Earth Science
SA = Short Answer
MC = Multiple Choice
Appendix III: Example of Modifying an Item’s Cognitive Level

Year One: Knowledge Level Item

6. What force acts to pull Chris down the ramp once he begins his downward motion?
   A. Gravity
   B. Friction
   C. Static Electricity
   D. Magnetism
Year Two: Analysis Level Item

7. In the diagram below, Chris pushes off the top of the half pipe and rides down the ramp. He stops at point D (indicated by an arrow). Check the MAJOR forces that are acting at each stage of his descent, labeled A, B, C, and D.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gravity</td>
<td>Gravity</td>
<td>Gravity</td>
<td>Gravity</td>
</tr>
<tr>
<td>Air Resistance</td>
<td>Air Resistance</td>
<td>Air Resistance</td>
<td>Air Resistance</td>
</tr>
<tr>
<td>Friction</td>
<td>Friction</td>
<td>Friction</td>
<td>Friction</td>
</tr>
<tr>
<td>Velocity</td>
<td>Velocity</td>
<td>Velocity</td>
<td>Velocity</td>
</tr>
<tr>
<td>Acceleration</td>
<td>Acceleration</td>
<td>Acceleration</td>
<td>Acceleration</td>
</tr>
</tbody>
</table>
Appendix IV: Linking Multiple-Choice and Open-ended Items

Item 12

When liquid water is heated to vapor, the water molecules are:

A. changed, and they become lighter
B. unchanged, and they exist as single molecules in the air
C. changed, and they form compounds with other air molecules
D. unchanged, and they are bonded together in groups

Item 19

Look at the picture above. Water is poured into Beaker 1 and frozen. Water is poured into Beakers 2 and 3 to match the level of the ice in Beaker 1. Each beaker is tightly sealed. Beaker 2 is heated until all the water has evaporated. Beaker 3 is left at room temperature. Describe how the mass of each beaker relates to the others, and explain why. Be specific. Be sure to compare EVERY beaker.
Appendix V: List of Abbreviations Used in this Study

NAEP – National Assessment of Educational Progress
NCEE – National Commission on Excellence in Education
AAAS – American Association for the Advancement of Science
NRC – National Research Council
ACC – Academic Competitiveness Council
PISA – Program of International Student Assessment
OECD – Organization for Economic Cooperation and Development
NCTAF – National Commission on Teaching and America’s Future
NSB – National Science Board
STEM – Science, Technology, Education, and Mathematics
NWO-TEAMS – Northwest Ohio-Teachers Enhancing Achievement in Mathematics and Science
OSCI – Ohio Science Institute
FOSS – Full Option Science System
STC – Science and Technology for Children
SI-I – Summer Institute I
AY – Academic Year
SI-II – Summer Institute II
EPP – Earned Point Percentage
SD – Standard Deviation
Appendix VI: Life Sciences Item from the Third Grade AY Instrument

Drawings of several organisms are shown below. Classify the organisms by correctly circling one or more of the options below them. Circle all groups to which the organism belongs.

<table>
<thead>
<tr>
<th>A. Vertebrate</th>
<th>A. Vertebrate</th>
</tr>
</thead>
<tbody>
<tr>
<td>B. Invertebrate</td>
<td>B. Invertebrate</td>
</tr>
<tr>
<td>C. Fish</td>
<td>C. Fish</td>
</tr>
<tr>
<td>D. Amphibian</td>
<td>D. Amphibian</td>
</tr>
<tr>
<td>E. Reptile</td>
<td>E. Reptile</td>
</tr>
<tr>
<td>F. Bird</td>
<td>F. Bird</td>
</tr>
<tr>
<td>G. Mammal</td>
<td>G. Mammal</td>
</tr>
</tbody>
</table>

![Bat](image1.png) ![Frog](image2.png)

![Penguin](image3.png) ![Snake](image4.png)

![Turtle](image5.png) ![Crab](image6.png)