2015

Analyzing the role of science practices in general chemistry courses and assessments

Jessica Jewett Reed
Iowa State University

Analyzing the role of science practices in general chemistry courses and assessments

by

Jessica Jewett Reed

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Chemical Education

Program of Study Committee:
Thomas A. Holme, Major Professor
Joseph W. Burnett
Thomas J. Greenbowe
Joanne K. Olson
Theresa L. Windus

Iowa State University

Ames, Iowa

2015

DEDICATION

This work is dedicated to those I love.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

It is with much gratitude that I thank my advisor, Dr. Tom Holme, for his advice, stories, and support. I am grateful for the many opportunities I have been afforded in the Holme group in conjunction with the ACS Exams Institute. Additionally, I would like to thank the members of my committee, Dr. Joe Burnett, Dr. Tom Greenbowe, Dr. Joanne Olson, and Dr. Theresa Windus for their interest in and support of my research.

I am so grateful to Dr. Kathy Burke for her advice, encouragement, enthusiasm, and mentorship. Also, many thanks to Dr. Johna Leddy for her continued guidance.

The Holme Research Group, past and present, deserves much appreciation for fostering an environment where research ideas could be exchanged freely and laughter was a daily occurrence. Special thanks to postdoctoral scholars, Dr. Allie Brandriet, Dr. Kim Linenberger, Dr. Cindy Luxford, and Dr. Jeff Raker for their guidance and assistance with my research. Their advice has been invaluable.

I am thankful to have had the support of wonderful family and friends during this journey. Mom and Dad, thanks for instilling in me the value of education and the importance of lifelong learning. Jason, I am so glad for your smiles and hugs. The advice and encouragement I have received from all of my family and friends over the years has meant a lot to me.

To Jordan, I am incredibly humbled by your patience and support during this journey. I am rarely certain of the future, but I am always certain of my future with you. Let us not lose sight of Galatians 6:10 in our journey together. I love you.

CHAPTER 1: GENERAL INTRODUCTION

This chapter serves as a general introduction to overarching themes of curriculum and assessment reform that are presented throughout the dissertation.

Goals of Science Education

Science, technology, engineering, and math (STEM) education has received a great deal of public attention in recent years as calls to foster more interest in STEM fields have persisted. While the implementation of new pedagogies and curricula to support goals and skills in science classrooms is relatively dynamic, the goals themselves have remained fairly static. DeBoer writes, "Ultimately what we want is a public that finds science interesting and important, who can apply science to their own lives, and who can take part in the conversations regarding science that take place in society" (2000, p. 598). In this sense, at least at many levels, the goals of science education remain centered around development of scientific literacy (American Association for the Advancement of Science, 1993; Bybee & DeBoer, 1994; DeBoer, 2000, 2011; Hofstein & Yager, 1982; Millar & Osborne, 1998). Additional goals relate to development of skills beyond the cognitive realm of content proficiency. Skills such as problem solving, critical thinking, and scientific reasoning are apt to be included in discussions of goals of science education (Hodson, 1988; Resnick, 1987). While these sentiments are directed, in general, toward the K-12 science classroom, it stands to reason that they should also apply to introductory undergraduate science courses (DeHaan, 2005). After all, the populations of such courses are often recent high school graduates taking their first, and perhaps only, college-level science course. Yet, students are often unaware of the goals beyond content proficiency in their science courses, especially when assessments are

misaligned with classroom instruction. The research herein aims to support improved alignment between instruction and assessment measures within general chemistry courses at the college level, as is suggested by recent reform efforts in science education. Thus, it is important to consider the reform efforts of science education that have led to a focus on creating a broader knowledge base by providing evidence of student engagement with science practices and emphasizing content knowledge depth over shallow understanding of a breadth of topics.

Reform Efforts Within Science Education

Efforts to reform science education have been around nearly as long as the field has existed it seems, yet as Tobias criticizes, little has changed (1992). Reform efforts seemingly changed trajectory in the late 1950s with the launch of Sputnik (DeBoer, 1997). Suddenly there was political interest to reform science curricula based upon the notion that American students were behind in mathematics and science learning as compared to their Soviet peers (DeBoer, 1997, 2011; Yager, 2000). This is not to say that research to support educational reforms had not been present in previous eras, rather that science became a mainstream component of general education during the 1950s and 1960s due to the political assets garnered from a scientifically literate public (DeBoer, 1997). Science was no longer a subculture accessible only to the intellectually elite and conveyed in an encyclopedic manner, but was now part of mainstream popular culture and related to everyday life (DeBoer, 1991). The purpose of this brief emphasis on science history is to set the stage for reform efforts that led to the integration of theory and practice in science instruction.

It would not be feasible to outline every reform effort that has occurred since the Sputnik era, so for a more historic review of these efforts, consult DeBoer (1991) or Atkin and Black (2003). For practical purposes, the reforms considered herein relate more directly to the research at hand as it relates to development and incorporation of scientific practices. In this regard, highlighted reforms are limited to a select number of national reform efforts that have shaped how the concept of science practices has developed.

This analysis of reform efforts can begin in the 1980s and early 1990s where science education efforts began to shift to encompass greater understanding of how people learn as obtained from advances in cognitive science (Yager, 2000). Science education moved beyond simply what students know to how they know it. Efforts such as those from the American Association for the Advancement of Science embodied within the reports *Science for All Americans* (1989) and *Benchmarks for Science Literacy* (1993) aimed to expand the focus of science education to encompass the development of skills that would make students scientifically literate citizens capable of understanding and managing the scientific events encountered in everyday life. The National Science Education Standards (NSES) (National Research Council, 1996) followed soon after, and provided criteria for what students should know and be able to do with that knowledge at various grade levels of science instruction through a lens of inquiry instruction. Many states used these documents to create their own state standards for science education. While the intentions of these documents were to aid in the creation of a more cohesive framework for unifying science content with skills related to the 21st-century, they received criticism as being "a mile wide and an inch deep" in terms of content, and

lacking definition in terms of what is meant by engaging students in inquiry to aid in the development of science practices (Hodson, 2003; Pellegrino, 2012). Therefore, a more unified and consistent framework supported by research endeavors in science education was necessary. This sparked the development of *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas* (National Research Council, 2012a) approximately 15 years later. The Next Generation Science Standards (NGSS) (Achieve, 2013) were derived from the *Framework* and build upon constructs found in the NSES. Science learning is structured around three dimensions in the *Framework: practices* that enable scientists to do their work, *crosscutting concepts* which link science disciplines, and *core ideas* related to the discipline. These documents build upon the prior suggestions and calls for reform by providing discrete learning outcomes that intertwine content and skills. Sadler writes "If the primary goal of science education is to support student abilities to engage in scientific practices, then educational opportunities should be designed such that they maximize student engagement in those practices" (2011, p. 3). In this manner, the NGSS seek to move science education away from rote memorization of isolated facts and cookbook laboratory experiments to dynamic engagement with content and scientific practices. The NGSS aims to provide fewer, higher, and deeper goals than predecessors criticized for being a "mile wide and an inch deep."

It is important to clarify that while these reforms are supported by government agencies, the construction of the documents to support these efforts was created through the counsel of teams of scientists, outstanding teachers, and science education researchers. They are not curricula, but support the development of specific curriculum

materials by providing description of what a student is expected to know and be able to do with that knowledge. Additionally, there are no prescribed methods for assessing the standards.

It should also be noted that while the efforts discussed previously relate to reform efforts within the United States, science education reforms are occurring on a global scale. For an extensive discussion of the global efforts of science education reform refer to Deboer (2011) or obtain highlights in Osborne and Dillon (2008). Many countries are now implementing more rigorous measures of student outcomes in science courses. For example, in Australia implementation of Threshold Learning Outcomes (TLOs) has aided in providing structure to undergraduate science courses by defining what a student must know and be able to do in order to pass the course (Lim, 2013; Schultz, Crow, & O'Brien, 2013). In Singapore, an overhaul of the assessment system has led to increasingly ambitious performance assessments where students are required to produce sophisticated written, oral, mathematical, physical, or multimedia products (Darling-Hammond & Adamson, 2010). Additionally, in Singapore, England, and Australia, high-stakes science tests include measures of experimental design and performance (Darling-Hammond & Adamson, 2010). All of these examples point to a common theme, the desire to not only engage students in higher-order cognitive practices, but to also measure student learning of those practices beyond traditional content knowledge.

Assessments to Measure Beyond Content

The National Research Council's Discipline-Based Education Research (DBER) report states: "Learning and becoming adept at science and engineering practices should not be separated from content learning" (2012b, p. 143). In order to measure student

success in the development of these practices as supported by constructs of disciplinary content, appropriate assessments which intertwine measures of content proficiency with measures of science practices will likely be the most efficient method to measure student progress. In this paradigm, "teaching to the test" would have a positive connotation as it would mean promoting higher order transfer skills and science practices in instruction, not just disjointed recollection of facts. Pellegrino describes how effective transformation of assessment can have additional positive effects such as supporting student learning and promoting advanced competency when integrated properly with course instruction (2014).

Some reform efforts in assessment are already implementing measures of skills beyond content. For example, courses and tests associated with Advanced Placement® Chemistry, Biology, and Physics have been redesigned to include measures of scientific practices and inquiry-based learning (College Board, 2011a, 2011b, 2014). Additionally, the Medical College Admission Test® (MCAT), has been redesigned to have measures of skills related to scientific inquiry and reasoning (Association of American Medical Colleges, 2014; Kirch, Mitchell, & Ast, 2013). Student performance data on these assessments is not readily available at this time since the first national iteration of the assessments has occurred only within the past year. While it is not yet possible to speculate on the success of these efforts to measure skills independent of content, anecdotal evidence suggests that measurements of this nature are quite difficult to make within a standardized testing environment. Nevertheless, the fact that high-stakes tests are moving to more sophisticated measures of students' learning and abilities related to

scientific practices and skills speaks volumes to the importance of these practices and skills in science education.

These exam reforms are couched within an evidence-based framework known as Evidence Centered Design (ECD). ECD exemplifies aspects of "backward design" in which the goals and outcomes to be measured, and the evidence to be accepted to represent mastery, are determined prior to the design of any assessment materials (Brennan, 2010; Huff, Steinberg, & Matts, 2010; Mislevy, Almond, & Lukas, 2003; Zieky, 2014). In this assessment model, emphasis is given to demonstration of ability to use knowledge appropriately in an assessment situation which requires transfer beyond recall of factual knowledge, which is a relatively novel approach in science assessments (Brennan, 2010). These efforts add to the clear and consistent message that assessment reforms are necessary in order to provide measures of student learning beyond traditional measures of content knowledge.

Purpose and Significance of the Research Herein

In order for the proposed reforms to make measurable headway, the manner in which assessments measure science learning will need to be reconsidered. The National Research Council highlights this need in its report entitled *Developing Assessments for the Next Generation Science Standards* (Pellegrino, Wilson, Koenig, & Beatty, 2014). The report describes how future assessments will need to be designed in order to support student learning within the realm of the NGSS. Three key components are suggested to be integral parts of this assessment system: "assessments designed to support classroom instruction, assessments designed to monitor science learning on a broader scale, and a series of indicators to monitor that the students are provided adequate opportunity to

learn science in the ways laid out in the *Framework* and the NGSS" (Pellegrino, et al., 2014, p. 4). The purpose of the research efforts reported herein was identify the goals and skills valued by chemistry instructors, and then determine how current forms of assessments are incorporating those skills and practices into measurements. Through the analysis of chemistry exam items developed on the levels of classroom instruction and broader, larger-scale assessment this project investigates two of the three components necessary to create an assessment system for the NGSS. It should be noted that the research herein focuses on efforts at the college level, particularly within general chemistry, even though the NGSS is designed for K-12 science education.

Students are often apt to subscribe to the notion that what is important to learn is what is assessed (Liu, Bridgeman, & Adler, 2012), yet by and large the majority of traditional assessments have yet to incorporate measures of cognitive domains beyond content knowledge. By analyzing current large-scale chemistry assessment materials for the incorporation of science practices, and additionally creating items to measure content and practices in a general chemistry course, the research herein investigates the efforts to transform chemistry assessments to align more closely with course instruction that values the development of scientific practices. By creating items to measure science practices which require students to use skills and knowledge beyond recall of content knowledge, the project supports the notion that assessment is to "educate and improve student performance, not merely audit it" (Wiggins, 1998). In this regard, the research project has significance because it not only informs the chemistry community about the current status of incorporation of science practices in large-scale assessment, but it also provides

evidence to support the construction of quality multiple-choice items as viable measures of science practice development.

Theories of Learning to Support This Research

The work herein was guided by two theories of learning: meaningful learning and the unified learning model.

Meaningful Learning

Meaningful learning stems from the work of Joseph Novak and his theory of education, human constructivism (Bretz, 2001; Novak, 1977, 1993). Human constructivism is derived from the ideas of psychologist and philosopher David Ausubel's *assimilation theory* (Ausubel, 2000; Ausubel, Novak, & Hanesian, 1968). Ausubel's theory describes the differences between rote and meaningful learning, outlines the conditions necessary for learning, and suggests that meaningful learning occurs when the learner is afforded opportunities in the domains of cognitive, affective, and psychomotor learning (Ausubel, 2000; Ausubel, et al., 1968; Bretz, 2001). Novak's theory asserts that humans construct knowledge individually, and thus it is incumbent upon the education system to support learners as they construct knowledge (Bretz, 2001). Additionally, meaningful learning integrates thinking, feeling, and acting to empower the learner, and thus encourages the learner to be committed and responsible for his or her own learning (Novak, 1977).

Meaningful learning is an important framework within science education in general, and more specifically within chemistry, especially when considering that one of the primary aims of science education is empower the learner to become scientifically literate and engage with science constructs in his or her life. Students often are apt to

memorize individual facts rather than purposefully connecting them to prior knowledge, which constitutes rote learning. Meaningful learning, rather, occurs when the student is afforded experiences in each of the three learning domains mentioned previously, and is able to incorporate new concepts encountered within the course to his or her existing mental structure of knowledge. In chemistry, Bretz (2001) describes the domains of cognitive, affective, and psychomotor learning and highlights how each relates to specific facets of chemistry. These non-arbitrary connections between old and new ideas are essential for meaningful learning (Novak, 1977), and support the individual's development of scientific literacy. The transferability of knowledge becomes a key component of meaningful learning, and as such, meshes well with the ideas of the *Framework* (National Research Council, 2012a) which support the transfer of knowledge and skill between core content, crosscutting concepts, and science practices.

Assessment serves as a powerful mechanism for rewarding and encouraging meaning making from the constructivist viewpoint (Mintzes, Wandersee, & Novak, 2005). Unfortunately, an overstuffed science curriculum often promotes rote learning as the most viable means to attaining high scores on course assessments. The focus of the research herein is on moving beyond assessments of rote memorization to assessments that measure skills and transfer of knowledge; assessments that represent consideration of meaningful learning in their design.

Unified Learning Model

An additional framework for analyzing this research is the Unified Learning Model (ULM) (Shell et al., 2009). The ULM combines the underlying ideas of several theories of learning to create one model for learning. By drawing upon principles of

cognitive science and psychology, the ULM provides a model of how people learn and describes a resultant paradigm for teaching and instruction. Within the ULM, working memory, knowledge, and motivation are central to understanding how people learn. Learning is influenced by the individual's working memory capacity, prior knowledge, or the concepts or skills the individual already knows, and the individual's motivations which drive him or her to put forth effort. The interconnection of content knowledge and procedural skills is an important component of the ULM (Shell, et al., 2009), and therefore relates to the research herein which focuses on the learning and assessment of skills and practices beyond traditional content knowledge. In this sense, learning goes beyond knowledge of concepts or facts (declarative knowledge) to encompass knowledge of skills, behaviors, and thinking processes (procedural knowledge). The practical implication of the ULM is that the learner is aided by the instructor who directs the learner's attention (working memory) to the concept or skill to be learned. The instructor also aids in the construction of connections between prior knowledge and new concepts or skills, and aids the learner in establishing goals to support the motivation to learn. In this sense, the instructor serves as a facilitator for individual learning that is multi-faceted; learning that includes the development of skills beyond content knowledge. Thus, this framework is useful when considering the broad scope of this research is to understand, analyze, and develop assessments that support how people learn practices and concepts beyond the realm of traditional declarative knowledge.

Dissertation Outline

The dissertation chapters reflect the progression of the research project, and intertwine to create a cohesive depiction of how science practices are valued and assessed

in general chemistry courses. Chapters 2 and 3 identify the goals and skills beyond content that general chemistry instructors value. Chapter 4 builds upon the work of the previous to chapters by analyzing large-scale chemistry assessments for the presence of science practices to better understand how the skills that instructors value are being assessed. The results of this analysis are then implemented in Chapter 5 where the research examines student performance on assessment items designed to intentionally incorporate measures of science practices. While the chapters are intertwined, each chapter serves as an independent publication and as such has its own literature review, theoretical framework, research questions, and analyses. The chapters combine to bring the research project full circle.

Chapter 2 highlights qualitative research conducted during the early stages of the project. Interviews were conducted with chemistry instructors from high schools, community colleges, and state-funded universities to identify the goals and skills valued in general chemistry courses.

Chapter 3 expands upon the work conducted in Chapter 2 and presents results from a national quantitative survey about non-content learning goals. The chapter offers more generalizable results about the status of goals and skills within general chemistry courses and assessments.

Chapter 4 describes the need for assessments to measure beyond content proficiency and includes a study in which a pre-designed rubric was modified to analyze chemistry multiple-choice assessments for incorporation of science practices.

Chapter 5 presents a pragmatic approach to the use of the assessment rubric detailed in Chapter 4 to create multiple-choice items containing science practices for general chemistry course exams. Student performance on items written to incorporate science practices is analyzed.

The concluding chapter, Chapter 6, provides a synopsis of general findings, implications, and sources of future work.

References

Achieve. (2013). Next generation science standards. Washington, DC: National Academies Press.

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology* (Vol. 1): AAAS.

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*: Oxford University Press.

Association of American Medical Colleges. (2014). What's on the MCAT 2015 Exam? Retrieved March 26, 2015, from https://www.aamc.org/students/download/377882/data/mcat2015-content.pdf

Atkin, J. M., & Black, P. (2003). *Inside science education reform: A history of curricular and policy change*. New York, NY: Teachers College Press.

Ausubel, D. P. (2000). *The acquisition and retention of knowledge : a cognitive view*. Boston: Kluwer Academic Publishers.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). Educational psychology: A cognitive view. New York, NY: Holt, Rinehart, and Winston.

Brennan, R. L. (2010). Evidence-Centered Assessment Design and the Advanced Placement Program®: A Psychometrician's Perspective. *Applied Measurement in Education*, *23*(4), 392-400.

Bretz, S. L. (2001). Novak's theory of education: Human constructivism and meaningful learning. *Journal of Chemical Education*, *78*(8), 1107.

Bybee, R. W., & DeBoer, G. E. (1994). Research on goals for the science curriculum. *Handbook of research on science teaching and learning*, 357-387.

College Board. (2011a). The AP biology curriculum framework. New York: The College Board.

College Board. (2011b). The AP chemistry curriculum framework. New York: The College Board.

College Board. (2014). The AP physics curriculum framework. New York: The College Board.

Darling-Hammond, L., & Adamson, F. (2010). Beyond basic skills: The role of performance assessment in achieving 21st century standards of learning. *Stanford Center for Opportunity Policy in Education (SCOPE)*, Stanford University. Retrieved from https://edpolicy. stanford. edu/sites/default/files/beyond-basic-skills-role-performance-assessment-achieving-21st-centurystandards-learning-report_0. pdf.

DeBoer, G. E. (1991). *A History of Ideas in Science Education: Implications for Practice*: New York, NY: Teachers College Press.

DeBoer, G. E. (1997). *What we have learned and where we are headed: Lessons from the Sputnik Era*.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of research in science teaching, 37*(6), 582-601.

DeBoer, G. E. (2011). The globalization of science education. *Journal of Research in Science Teaching, 48*(6), 567-591.

DeHaan, R. L. (2005). The impending revolution in undergraduate science education. *Journal of Science Education and Technology, 14*(2), 253-269.

Hodson, D. (1988). Toward a philosophically more valid science curriculum. *Science Education, 72*(1), 19-40.

Hodson, D. (2003). Time for action: Science education for an alternative future. *International Journal of Science Education, 25*(6), 645-670.

Hofstein, A., & Yager, R. E. (1982). Societal issues as organizers for science education in the ‘80s. *School science and mathematics, 82*(7), 539-547.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Kirch, D. G., Mitchell, K., & Ast, C. (2013). The new 2015 MCAT: testing competencies. *JAMA, 310*(21), 2243-2244.

Lim, K. (2013). Threshold learning outcomes. *Chemistry in Australia*, 35.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: motivation matters. *Educational Researcher, 41*(9), 352-362.

Millar, R., & Osborne, J. (1998). *Beyond 2000: Science education for the future: A report with ten recommendations*: King's College London, School of Education.

Mintzes, J. J., Wandersee, J. H., & Novak, J. D. (2005). *Assessing science understanding: A human constructivist view*: Academic Press.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academies Press.

National Research Council. (2012a). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

National Research Council. (2012b). *Discipline-based education research : understanding and improving learning in undergraduate science and engineering*. Washington, DC: National Academies Press.

Novak, J. D. (1977). *A theory of education*. Ithaca, NY: Cornell University Press.

Novak, J. D. (1993). Human constructivism: A unification of psychological and epistemological phenomena in meaning making. *International Journal of Personal Construct Psychology, 6*(2), 167-193.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections* (Vol. 13): London: The Nuffield Foundation.

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831-841.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65-77.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*: National Academies Press.

Resnick, L. B. (1987). *Education and learning to think*: National Academies.

Sadler, T. D. (2011). Situating socio-scientific issues in classrooms as a means of achieving goals of science education *Socio-scientific Issues in the Classroom* (pp. 1-9): Springer.

Schultz, M., Crow, J. M., & O'Brien, G. (2013). Outcomes of the chemistry Discipline Network mapping exercises: are the Threshold Learning Outcomes met? *International Journal of Innovation in Science and Mathematics Education, 21*(1), 81-91.

Shell, D. F., Brooks, D. W., Trainin, G., Wilson, K. M., Kauffman, D. F., & Herr, L. M. (2009). *The unified learning model: How motivational, cognitive, and neurobiological sciences inform best teaching practices*: Springer Science & Business Media.

Tobias, S. (1992). *Revitalizing Undergraduate Science: Why Some Things Work and Most Don't. An Occasional Paper on Neglected Problems in Science Education*. Tucson: Research Corporation.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance* (Vol. 1). San Francisco, CA: Jossey Bass.

Yager, R. E. (2000). The history and future of science education reform. *Clearing House, 74*(1), 51-54.

Zieky, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicología Educativa, 20*(2), 79-87.

CHAPTER 2: A QUALITATIVE INVESTIGATION OF GENERAL CHEMISTRY INSTRUCTORS' GOALS FOR DEVELOPMENT OF STUDENTS' SKILLS BEYOND CONTENT PROFICIENCY

Jessica J. Reed and Thomas A. Holme

A paper to be submitted to the *Journal of Chemical Education*

Abstract

It is important to understand what goals and skills general chemistry instructors value beyond content proficiency in their courses to better understand how these values align with current reform efforts in science education. These reform efforts at the K-12 level aim to widen the focus of science curricula to encompass development of practices and skills beyond content. Qualitative interviews were conducted with a total of 18 general chemistry instructors from high schools, community colleges, and state-funded universities to identify what types of goals they held for students. Interviews were then analyzed to reveal patterns in the types of skills valued across institution types.

Introduction

"Well, I think that the broad statement is for students to be able to think like a scientist or think like a chemist." (Richard, State-funded university instructor)

Richard's description of the primary goal for his general chemistry course aligns with goals of science education previously articulated in the literature (DeBoer, 2000; Duschl, 2008; Hodson, 2003; Longbottom & Butler, 1999; Norris, 1997). As technology continues to make access to facts and information easier, mere factual recall is no longer

enough to demonstrate that a student has mastered a particular concept, and it provides little evidence that the student can "think like a scientist." Rather, it is of greater importance to demonstrate that the student understands the concept within the appropriate context and can use it appropriately. The ability to think critically, evaluate problems, and critique information from a scientific perspective becomes increasingly more valuable from the instructional standpoint.

The focus of science education, particularly at the K-12 level, is widening to encompass development of science practices that describe what students should be able to do with the knowledge they develop in their science courses (Achieve, 2013; Cooper, 2013; National Research Council, 2012a). These practices are intended to be developed with the same level of instruction and emphasis as concepts associated with the content discipline. Integration of content knowledge with development of skills is central to modern curriculum reform efforts (Achieve, 2013; College Board, 2011a, 2011b, 2014). The premise of these reform efforts is often associated with the idea of data-driven and evidence-based curriculum and assessments (Cooper, 2010, 2013, 2014; Lloyd & Spencer, 1994; National Research Council, 2012a, 2012b).

While these reform efforts are primarily aimed at courses within the realm of K-12 education, it would be imprudent to ignore or dismiss them in higher education. Should these reform efforts be widely implemented, for example, their effects promise to change how Science, Technology, Engineering, and Math (STEM) courses are taught and assessed at the post-secondary level, as students entering the college classroom may be prepared to engage with science content in a variety of fashions. The chemistry instructional community has a tremendous opportunity to aid in these reform efforts, as

general chemistry courses often enroll a large and diverse portion of the undergraduate population at many institutions. Thus, even though research efforts have not yet established how a curriculum of this nature might look in practice at the post-secondary level, it is important that the goals of general chemistry instruction and assessment at the collegiate level are commensurate with efforts to encompass development of skills and practices that students can transfer to other courses and disciplines, and additionally, apply to everyday life events and decision-making processes.

In addition to changing the type of instruction encountered in a science course, these reform efforts will also change the face of assessment. Assessments of student learning will encompass not only content knowledge, but also measures of development of skills and practices. Further discussion of the status of assessment as it relates to chemistry is included in the following chapter.

In order to inform the design of new assessments to measure development of science practices and skills, it is important to understand what goals and skills are valued in general chemistry instruction. Other studies have evaluated the non-content learning goals of the chemistry laboratory in the United States (Bretz, Fay, Bruck, & Towns, 2013; A. D. Bruck & Towns, 2013; L. B. Bruck, Towns, & Bretz, 2010), but few have examined the status of learning goals in the context of chemistry instruction on the whole.

The aim of this study is to investigate what are the content independent goals that general chemistry instructors hold for their students. Numerous studies have confirmed that what instructors say they value, and what they actually do in practice are quite

frequently mismatched (Bol & Strage, 1996; Crooks, 1988; Goodlad, 1984). Thus, it is important to denote that the goal of this study was not to make sweeping generalizations about the status of the collegiate chemistry classroom, but rather it was to gain insight into what specific goals might be valued for development in general chemistry courses for the purpose of informing future assessment design. In addition to identifying such goals in an open format qualitative study, the data obtained in this work were also used to inform development of quantitative survey items for more quantitative measures reported in Chapter 3.

Theoretical Framework

Novak's theory of education, human constructivism, aids in the design and analysis of this research (Bretz, 2001; Novak, 1977). In this theory, Novak relates the ideas of psychologist and philosopher David Ausubel to how learners construct knowledge. Ausbel's assimilation theory describes differences between rote and meaningful learning, outlines conditions necessary for meaningful learning, and asserts that meaningful learning occurs when the learner is afforded experiences in each of the three learning domains (cognitive, affective, and psychomotor) (Ausubel, et al., 1968). Human constructivism builds upon Ausubel's theory by suggesting that it is the duty and role of the educational system to provide supports and opportunities for learners as they construct their own knowledge. Meaningful learning integrates thinking, feeling, and acting, and therefore empowers students to commit to and be responsible for their own learning. Use of this framework to analyze the learning goals of general chemistry instructors provides insights into how the learning goals provide opportunities for meaningful learning in a general chemistry course (Bretz, et al., 2013).

Research Question

The nature of the study was to serve as an open-ended query to identify the types of goals and skills independent of content that general chemistry instructors value for development in their students. The findings reported herein answered the question: *What themes emerge about the non-content goals and skills that general chemistry instructors value?*

Methods

Human Subjects Research Approval

This research involved human subjects and as such was required to meet the specifications of the Institutional Review Board at Iowa State University. Approval to conduct this research was granted (IRB 12-402) under the condition that participants would not be identified. The informed consent document given to participants is available in Appendix A following this chapter.

Participant Recruitment and Description

Participants (N = 18) in the study consisted of general chemistry instructors from high schools (N = 5), community colleges (N = 7), and state-funded universities (N = 6) within the state of Iowa. Iowa provides a reasonable case study for this work because the defining characteristics of the institutions are readily distinguished. This allows the possibility that both similarities and differences might be identified in the types of goals and skills instructors value in general chemistry. The participants taught courses that had been deemed equivalent by state articulation agreements, meaning that students could

transfer the credit earned at one institution to another without question. At the high school level, this consisted of Advanced Placement® (AP) and dual-enrollment (DE) courses. Equivalent courses at the community college and state university level were two-semester general chemistry courses typically designed for science majors. All participants had taught this type of general chemistry course within the past five years.

Identifying eligible participants was difficult because course instructor information is not often made readily available on the websites of high schools and community colleges. Therefore, sampling methods varied by institution type.

A method of snowball sampling was employed to identify high school teachers who could be potential interview participants. Snowball, or chain, sampling relies on the use of knowledge or suggestions from people who are familiar with cases that meet the desired sample criteria to identify sample participants (Creswell, 2012). In this study, high school instructors were identified as potential participants via word-of-mouth references made by their colleagues or other interview participants.

In regard to community colleges, identification of chemistry instructors was still difficult, but at least one instructor of general chemistry at each of Iowa's fifteen community colleges was identified. This was done through information available on institutions' websites. Criterion sampling was used to ensure that all persons selected for the sample met the necessary criteria of type of general chemistry course taught, and had taught the course within the past five years (Creswell, 2012).

Information contained on the websites of Iowa's three state-funded universities allowed for the identification of general chemistry instructors relatively easily. Use of

criterion sampling ensured that only instructors with proper course experience were invited to participate.

Once identified, prospective participants were contacted via email and invited to participate in the study. Interviews were scheduled at the convenience of the participants and were conducted via telephone near the middle of the fall semester of 2012. It should be noted that participants did not receive any incentives or compensation for participating in the interview.

Table 1 provides information about each of the participants, including a pseudonym and institution type. It has been denoted whether a high school instructor taught AP or DE general chemistry. No other demographic information is included in order to protect the identity of instructors at small institutions who may be easily identifiable.

Instructors' experience teaching general chemistry ranged from 1 to 33 years. Course sizes and pedagogies varied widely across the participants, but generally courses were primarily lecture based regardless of size. High school and community college instructors were typically responsible for teaching both lecture and laboratory associated with their course, whereas the majority of state-funded university instructors taught only one or the other. Instructors often described the students enrolled in their courses as "diverse," whether through ethnicity, college major, academic achievement, mathematics ability, or age. One exception to this was that high school instructors primarily had high performing students due to the nature of AP or Dual-enrollment chemistry courses being electives, and generally taken after a required introductory chemistry course.

Additionally, community college instructors described students with lower mathematics abilities than high school and state-funded university instructors, which may be in part due to the difference in admission requirements between community colleges and state-funded universities.

Data Collection and Analysis Methods

Interviews were semi-structured in nature. This meant that the interviewer followed a guide while conducting the interview, but was free "to request elaboration of additional details and examples in response to the personal views offered by the interviewee" (Bretz, 2008, p. 84). The questions were generally open-ended in nature to permit the interviewee the freedom to respond (Patton, 2002).

Interviewees were informed of their rights as a participant prior to the actual interview, and were asked to give consent at the beginning of the interview. All interviews were digitally recorded and conducted by the same interviewer. Each participant was interviewed only once, and interviews ranged from 48 to 93 minutes in length.

The interview had three main sections. The first section asked participants to describe their teaching background, the demographics of the course they teach, and their pedagogy for the course. The interviewer then provided a transition to the second part of the interview in which the interviewee discussed his or her learning goals for the course. In order to ensure that participants understood what a learning goal was for purposes of the interview, each was asked to give his or her definition of the term. If the participant was using a different definition than what was expected by the researcher and used by

other participants, the researcher would clarify this with the participant. Next, the participant would be asked the broad question of what goals did he or she have for the course. The interviewer would then prompt for discussion of content specific learning goals or non-content related skills depending on the responses of the interviewee. The final portion of the interview asked participants to compare their learning goals to those of general chemistry instructors at other types of institutions. For example, a community college instructor would be asked questions about how his learning goals might be similar or different compared to those of high school teachers and university faculty. For the purpose of survey item development, this paper focuses on the analysis of the second section of the interview guide which focuses on instructors' learning goals.

All interviews were transcribed by the interviewer using InqScribe software (Inquirium, 2011), and then uploaded into Dedoose (SocioCultural Research Consultants, 2014), an online software program for analysis. The constant comparative method, found in Grounded Theory, was used to analyze the data (Glaser & Strauss, 1967; Strauss & Corbin, 1998). The interviews were first coded using open coding methods. The codes were then compared, and each interview recoded using themes that emerged from the open coding (Creswell, 2012; Lincoln & Guba, 1985; Patton, 2002; Strauss & Corbin, 1998). The codebook for the qualitative analysis can be found in Appendix C following this chapter. In particular, the interviews were analyzed to examine what learning goals instructors held for their students.

Throughout the analysis of the interviews, the researcher used peer debriefing to ensure that valid inferences were being drawn from the data (Creswell & Miller, 2000; Lincoln & Guba, 1985). A group of chemistry education researchers served as an external

check by asking questions and providing feedback on the methods of data collection and interpretation of results (Creswell, 2012). During the interviews the researcher also validated participants' responses by asking follow-up questions or asking for clarification when a response was ambiguous. Additionally, the results of the interviews are supported by quantitative survey data described in the following chapter.

Due to the small number of participants, it important to consider that any saturation that is present is in total, and not necessarily at the institutional level. Recall that the IRB called for anonymity of the participants so it is not possible to provide any additional identifiers beyond pseudonym and institution type when reporting results.

Results and Discussion

First, it was important to verify that the participant understood the term "learning goal," and could appropriately apply it to the discussion at hand. All participants gave similar responses when asked to define a learning goal. A representative participant definition:

"*A learning goal would be something that I either want a student to know, understand, or be able to do." (Darrin, HS-DE)*

The interview progressed to have the participant use his or her definition of a learning goal in order to describe what learning goals he or she held for the course. When asked this question, 11 out of the 18 participants began their responses by stating goals related to knowledge of specific content topics. Participants who first discussed content knowledge goals were then prompted to discuss what non-content goals they held for students. The opposite was true for those participants that began with a discussion of non-

content goals. A probing question at this point in the interview would have been something like, "Now that you've highlighted what content knowledge goals you have for students, do you have any non-content goals for students in your course?"

Content knowledge goals were comparable for all participants regardless of institution type, and were consistent with the content associated with a full-year general chemistry course.

Summary of Observations about Non-Content Goals

Variances occurred when considering the non-content goals instructors held for students. Figure 1 displays the counts for the most common non-content learning goals referenced by instructors. The most common goals might best be categorized as development of life skills and appreciation of chemistry, with 10 out of 18 respondents stating these were goals for their general chemistry course. Many of the goals are multi-faceted and participants described specific aims that were considered to be a sub-category of a larger goal. Sub-categories of the non-content goals data analysis, problem solving, laboratory skills, and life skills are displayed in Table 2 through Table 4.

Not surprisingly, all five of the high school teachers described developing life skills as a goal of their course since students would soon be transitioning to independence at college or in a career. Burt described it as the most important non-content learning goal of his AP chemistry course:

> *"And so we basically try to motivate them. We try to find ways to get them to be self-directed more than anything else. Um. Because that's the*

one key thing I've ever seen that decides students' ability to survive in the college setting." (Burt, High School AP Chemistry Instructor)

A breakdown of the most frequently described subcategories of life skills and code counts can be seen in Table 2.

Perhaps of greater interest were the differences between institution types. Community college instructors tended to focus on data analysis, appreciation of chemistry, laboratory goals, and developing understanding of the nature of science. State-funded university faculty focused primarily on developing life skills and an appreciation of chemistry. Perhaps this is due to the nature of general chemistry courses at the state-funded institutions as being fairly large courses that funnel students into more specific science courses, whereas the educational and career goals of the community college population may necessitate the development of more specific scientific skills (data analysis and laboratory skills) at an earlier stage in course progression.

While it is important to note these differences, it is of greater benefit to understand the factors that contribute to why these differences exist. An advantage of a qualitative study is that it allows for greater depths of explanation about why these goals were important to the instructors and how the type of institution influences implementation. For example, a community college instructor described the diversity of student backgrounds in his general chemistry course:

"There is a significant component that are people who are intent on going on in a four-year program somewhere. And then there is a component that is only interested in a two-year program, but they want

> *a more advanced level of chemistry […] And then there are a significant portion of the class is older, has been in the workforce, and is going about trying to re-train for new job opportunities. Uh. So that's a significant part of it. And then there's occasionally people who are, you know, it's not that their program requires it, or anything like it. Maybe they're liberal arts major, but they're just interested and want to see what it's about." (Gerald, Community College Instructor)*

From this perspective, it is understandable that the non-content goals of a general chemistry course at a community college might differ from other institutions due to the perceptions instructors have of the diverse needs of the population served by the community college.

Additionally, when the size and diversity of student majors represented in a state-funded university general chemistry course are considered, it is understandable why developing an appreciation for chemistry is a top learning goal for these instructors.

> *"The vast majority of students we have in our course are not chemistry majors. So, you know, we want them to, um, be able to, you know, have a positive viewpoint looking backwards and saying, 'Sure. Now I see a connection to this.' Or, you know, have some connection to their life so it doesn't seem like it is some foreign entity that they were just thrown into the class because someone told them to sit there. And so, I hope that we are having a positive influence on them […]" (Alice, State-funded University Instructor)*

Two-thirds of the state-funded university instructors described appreciation of the subject of chemistry, particularly as it relates to other disciplines and everyday life, as a goal of their course. Overall, 56% of participants across institution types described developing an appreciation for the subject of chemistry as a learning goal for their general chemistry course.

High school and community college instructors also focused on development of an understanding of the nature of science.

> *"I try to do a little bit of a nature of science focus as well where we'll talk throughout the semester about how scientists came to understand this concept, what we know, what we don't know...umm...you know, where there have been changes in our body of knowledge […] And to give them more of an idea that, if they choose to continue in science, it's something that they could contribute to that body of knowledge." (Daniel, Community College Instructor)*

In total, 6 (3 HS, 2 CC, and 1 SFU) out of the 18 participants described goals related to the understanding of how science works and has developed over time.

Interestingly, only 5 out of 18 participants described goals related to the development of skills traditionally associated with chemistry courses, problem solving and critical thinking, respectively, without being prompted. Perhaps the instructors who did not mention these goals have become so accustomed to using problem solving and critical thinking in chemistry that they have the perception that these goals are just inherently part of chemistry learning. If an instructor did not discuss development of

problem solving skills, the interviewer prompted the participant by asking how development of problem solving skills might fit into their course objectives, if at all. Of the 13 instructors prompted about this construct, all but one described incorporation of problem solving skills to varying degrees. One community college instructor said he did not feel that students were cognitively ready to learn how to solve complex problems until they reached upper-level university courses. It is unclear as to whether in practice the students are really learning problem solving skills, or merely constructing algorithms through repetition of exercises. The representative quotes below are from instructors who mentioned problem solving without being prompted about it.

> *"I would really like students to have a sense of how to approach a problem, and a true problem, not just an exercise. It's really something that they have absolutely no idea how to go about solving." (Dianne, State-funded University Instructor)*

> *"You know, problem solving, critical thinking, um, and pretty much all of the transferrable skills that they can learn when they are in school or college, it's going to be life applicable skills."(Kumar, State-funded University Instructor)*

General chemistry courses often provide opportunities for students to learn and hone laboratory skills. The importance of learning in the laboratory is well documented in relation to chemistry (Hofstein, 2004, 2004; Hofstein & Lunetta, 2004; Hofstein & Mamlok-Naaman, 2007; Lunetta, Hofstein, & Clough, 2007; Reid & Shah, 2007). Five participants, three of which were Community College instructors, readily described goals

related to development of specific laboratory skills. A breakdown of the goals and skills categorized as "Laboratory Skills" and the frequency counts can be seen in Table 3. Data analysis was often described in conjunction with the laboratory, but some instructors described it as goal of the lecture portion of the course as well, so it was considered as an independent skill from laboratory skills. Four out of the five participants who held data analysis as a goal for their course wanted to develop skills related to graphing with their students. Data analysis by graphing ranged from knowing how to construct a graph and draw conclusions from it, to analyzing graphical data to make inferences and decisions. Table 4 outlines how goals related to data analysis were described by five participants.

What I've moved to mostly for non-content goals is, uh, the analysis of data. Being able to, you know, graph the data, you know, get slopes. That sort of thing. And understanding what the data mean.

> *"In the labs that I do, many of them require them to share data with the whole class and analyze a larger data set, instead of just looking at the individual data that they've gathered. Often times that involves some kind of a graphical analysis. We use Microsoft Excel to make graphs of the lab data and talk about data that doesn't fit in, that's outliers. We try to discern trends in the data. We certainly can get at specific chemistry concepts with it, but to me, it's translatable to a much broader area. Many of my students are not going to go on and study chemistry specifically, but if they have that ability to do some data analysis and look at data as a whole, and look for trends, then it's going to be*

transferrable to whatever field they're going into." (Daniel, Community College Instructor)

Analysis of Response Patterns

In the context of meaningful learning, the interview participants provided descriptions of learning goals that bridged the cognitive, affective, and psychomotor domains. Instructors described goals related to chemistry content which lie within the cognitive domain, but they also described the development of other cognitive skills such as problem solving and critical thinking. Goals that lie predominately in the psychomotor domain, with some overlap with the cognitive domain, such as laboratory skills, data analysis (graphing), and communication skills, were also described. The largest percentage of responses were in relation to goals which are predominantly in the affective domain, such as appreciation of chemistry, life skills, and understanding and appreciating the nature of science, which also overlaps significantly with the cognitive domain. How the instructors provided opportunities for meaningful learning in practice, and whether their students made meaningful connections between concepts and domains, remains unknown. Yet, the fact that the instructors described goals that span all three domains indicates that meaningful learning is a valuable construct for approaching the development of non-content goals and skills in evidence-based assessments and curricula.

Even though instructors valued development of these goals and skills, it did not mean that there was not push-back from students. Community college instructor Gerald describes students' reluctance to think critically and conceptually in his course. "So I think my biggest problem is that I want to help them understand concepts, not just be able

to solve problems mechanically, but there's a strong resistance on the students' part to ever have to learn concepts." He continues on to say "[…] And to varying degrees I get some success in that, but it's, you know, sometimes just one little course is not going to change people." Additionally, state-funded university instructor Richard stated that his students "probably view it as nagging" when he routinely emphasizes problem solving skills and conceptual understanding in his course.

Assessment of Non-Content Learning Goals

Discussion of how the learning goals were typically made known to students and assessed yielded results that were not unexpected. Primarily, instructors do not explicitly state the non-content learning goals to students, and do not assess skills beyond what may be encompassed by traditional formative and summative assessments. The majority of instructors described using traditional forms of assessment (exams, homework, quizzes, laboratory reports, and student response systems (clickers)) to measure chemistry content knowledge, and did not mention using these types of assessments for measures of other skills. None of the participants described using any published assessment that is specifically designed to measure the status of a particular non-content goal or skill. This aligns with previous research that suggests that faculty are unaware of the prevalence of additional forms of assessment of students' development beyond content knowledge (Emenike, Raker, & Holme, 2013; Raker, Emenike, & Holme, 2013; Towns, 2009)

Overall, only 4 out of the 18 instructors interviewed described expressing the goals for skill development to students. In these cases the instructors admitted that the goals were typically expressed informally, often verbally. All 18 participants described

formally expressing and assessing the content knowledge goals of the course to their students. Three instructors described using exams to measure problem solving skills, critical thinking, and/or conceptual understanding of typically algorithmic concepts. Additionally, three instructors described using lab reports to assess development of graphing and data analysis skills. Other instructors described informal observation of students as an assessment tool. For example, when asked about how she measures or assesses non-content learning goals, high school AP teacher Jackie says "I don't know, I guess, just making sure that I'm keeping an eye on them […] just kind of paying attention to them as an individual." All of these observations suggest that researchers and practitioners who wish to develop and implement new assessment methods need to be aware of the existence of barriers to such implementation, specifically that instructors may not be ready to admit that anything is missing from current assessment practices.

In this regard, it is not unexpected that the instructors did not formally express and assess students' development of content independent skills, since assessment in chemistry has traditionally focused solely on content knowledge. It is useful to understand how these instructors incorporated, or more frequently did not incorporate, measures of non-content goals and skills into their assessments. This information provides evidence that development of future measures of non-content skills will need to be embedded in content rich assessments to be viable for use in the chemistry classroom.

Use of Qualitative Results to Inform Quantitative Survey Design

Data collected from the qualitative interviews were used to inform the development of survey items for a national survey conducted by the ACS Examinations

Institute (ACS EI). The purpose of the survey was to inform examination development committees about the status of conceptual understanding in general chemistry. Items regarding the use and development of non-content goals and skills were included at the end of the survey. The goal of these items was to provide additional, and more generalizable, data to support the findings of the qualitative interviews presented here. The results of these items are discussed in the following chapter. The survey items are included in the Appendix following Chapter 3.

The recurring themes in the interviews were appreciation of chemistry in everyday life, development of communication skills, laboratory skills, graphing of data, interpreting and drawing conclusions from data, life skills (e.g., study skills, responsibility, time management), problem solving skills, nature of science (i.e., how science works and has developed), critical thinking, and conceptual understanding of traditionally algorithmic problems. Additionally, data from the interviews suggested that some instructors felt that students were not meeting their expectations for achievement of the non-content learning goals, even though they did not formally assess these skills. Therefore it was of interest to understand not only what goals and skills instructors value, but also how they assess the skills and perceive student performance.

Three questions were developed to assess instructors' perspectives on ten non-content skills that were the most common themes during the qualitative interviews. Items were developed to be multiple-choice. The first question related to frequency of intentional and explicit incorporation of the learning objectives into their course. The premise behind this question being that if an instructor frequently incorporates a goal or skill into the course, then the instructor likely values the skill to be developed in students.

The second question related to how the learning goals were assessed in the course. Participants were given a variety of formative and summative assessment methods and asked to select all modes of assessment that applied to each learning goal. Additional response options were created for those that do not assess or incorporate a particular learning goal. The third question developed related to how instructors' perceived their students to be meeting the expectations for development of each learning goal.

Conclusions

The qualitative interviews provided insight into how meaningful learning opportunities are afforded in the general chemistry classroom. These insights were used to design quantitative survey items and also inform the future development of assessments. As curriculum reform efforts necessitate the development of assessments to measure the advancement of goals and skills independent of content, it is important to understand how the general chemistry community values these skills and what barriers may exist to the development and implementation of assessments to measure them.

Limitations of the Study

The limitations of this study stem primarily from the sampling methods used. In order to ensure comparable course standards, the researchers limited the sample population to general chemistry instructors within one geographical state. This allowed for the focus of the interview to be fairly independent of discussion of chemistry content, but limited the potential sampling population. By providing a general description of the participants and their courses, the reader is able decide whether the findings are

applicable to another general chemistry instructor population in which he or she may be part.

Closing Remarks

Overall, the general consensus of instructors was that regardless of how non-content goals and skills were being implemented into the curriculum and assessment, they are valuable skills for students to develop within the realm of general chemistry.

> *"[…] But school's about more than just learning the content, to me at least. We also have an education system to prepare people to be thinkers and to be productive members of society. […] It's partly about knowing the content, and being able to relate that to what's going on in everyday life, but also just being able to realize that critical thinking skills and things like that are not just necessarily just consigned to science. They're skills for everything else in your life, too." (Lisa, High School AP Chemistry Instructor)*

Regardless of a student's career or educational trajectory, these instructors desire to make their students scientifically-literate members of society by developing the students' critical thinking and problem solving skills, skills transferrable beyond chemistry content.

The role of curriculum reform efforts in chemistry education at the post-secondary level is not entirely yet known. It is evident that independent of a particular reform movement, the instructors interviewed place the development of non-content skills in high regard in their general chemistry courses, yet do very little to express the value of these skills or assess the development of these skills beyond the realm of

traditional content assessments. Thus, as chemistry educators and assessment designers prepare to meet the goals of the curriculum reform efforts, special consideration should be given to how the value of non-content skills will be portrayed to students through instruction and assessment.

References

Achieve. (2013). Next generation science standards. Washington, DC: National Academies Press.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). Educational psychology: A cognitive view. New York, NY: Holt, Rinehart, and Winston.

Bol, L., & Strage, A. (1996). The contradiction between teachers' instructional goals and their assessment practices in high school biology courses. *Science Education, 80*(2), 145-163.

Bretz, S. L. (2001). Novak's theory of education: Human constructivism and meaningful learning. *Journal of Chemical Education, 78*(8), 1107.

Bretz, S. L. (2008). Qualitative research designs in chemistry education research. In D. M. Bunce & R. S. Cole (Eds.), *Nuts and Bolts of Chemical Education Research*: Oxford University Press.

Bretz, S. L., Fay, M., Bruck, L. B., & Towns, M. H. (2013). What faculty interviews reveal about meaningful learning in the undergraduate chemistry laboratory. *Journal of Chemical Education, 90*(3), 281-288.

Bruck, A. D., & Towns, M. (2013). Development, implementation, and analysis of a national survey of faculty goals for undergraduate chemistry laboratory. *Journal of Chemical Education, 90*(6), 685-693.

Bruck, L. B., Towns, M., & Bretz, S. L. (2010). Faculty perspectives of undergraduate chemistry laboratory: Goals and obstacles to success. *Journal of Chemical Education, 87*(12), 1416-1424.

College Board. (2011a). The AP biology curriculum framework. New York: The College Board.

College Board. (2011b). The AP chemistry curriculum framework. New York: The College Board.

College Board. (2014). The AP physics curriculum framework. New York: The College Board.

Cooper, M. M. (2010). The case for reform of the undergraduate general chemistry curriculum. *Journal of Chemical Education, 87*(3), 231-232.

Cooper, M. M. (2013). Chemistry and the next generation science standards. *Journal of Chemical Education, 90*(6), 679-680.

Cooper, M. M. (2014). Evidence-based reform of teaching and learning. *Analytical and bioanalytical chemistry, 406*(1), 1-4.

Creswell, J. W. (2012). *Qualitative inquiry and research design: Choosing among five approaches*: Sage.

Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitative inquiry. *Theory into practice, 39*(3), 124-130.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of educational research, 58*(4), 438-481.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of research in science teaching, 37*(6), 582-601.

Duschl, R. (2008). Science education in three-part harmony: Balancing conceptual, epistemic, and social learning goals. *Review of research in education, 32*(1), 268-291.

Emenike, M., Raker, J. R., & Holme, T. (2013). Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *Journal of Chemical Education, 90*(9), 1130-1136.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research*. Chicago: Aldine Publishing Company.

Goodlad, J. I. (1984). *A place called school : prospects for the future*. New York: McGraw-Hill Book Company.

Hodson, D. (2003). Time for action: Science education for an alternative future. *International Journal of Science Education, 25*(6), 645-670.

Hofstein, A. (2004). The laboratory in chemistry education: Thirty years of experience with developments, implementation, and research. *Chemistry Education Research and Practice, 5*(3), 247-264.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science education, 88*(1), 28-54.

Hofstein, A., & Mamlok-Naaman, R. (2007). The laboratory in science education: the state of the art. *Chemistry education research and practice, 8*(2), 105-107.

Inquirium, LLC. (2011). InqScribe (computer software).

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry* (Vol. 75): Sage.

Lloyd, B. W., & Spencer, J. N. (1994). The forum: New directions for general chemistry: Recommendations of the task force on the general chemistry curriculum. *Journal of chemical education, 71*(3), 206.

Longbottom, J. E., & Butler, P. H. (1999). Why teach science? Setting rational goals for science education. *Science Education, 83*(4), 473-492.

Lunetta, V. N., Hofstein, A., & Clough, M. P. (2007). Learning and teaching in the school science laboratory: An analysis of research, theory, and practice. *Handbook of research on science education*, 393-441.

National Research Council. (2012a). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

National Research Council. (2012b). *Discipline-based education research : understanding and improving learning in undergraduate science and engineering* (S. R. Singer, N. R. Nielsen, & H. A. Schweingruber Eds.). Washington, DC: National Academies Press.

Norris, S. P. (1997). Intellectual independence for nonscientists and other content-transcendent goals of science education. *Science Education, 81*(2), 239-258.

Novak, J. D. (1977). *A theory of education*. Ithaca, NY: Cornell University Press.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage Publications.

Raker, J. R., Emenike, M. E., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education, 90*(8), 981-987.

Reid, N., & Shah, I. (2007). The role of laboratory work in university chemistry. *Chemistry Education Research and Practice, 8*(2), 172-185.

SocioCultural Research Consultants, LLC. (2014). Dedoose. from https://www.dedoose.com

Strauss, A. L., & Corbin, J. M. (1998). *Basics of qualitative research : techniques and procedures for developing grounded theory* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Towns, M. H. (2009). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education, 87*(1), 91-96.

Table 1. *Participant Pseudonyms and Corresponding Institution Type*

Participant Pseudonym	Institution Type
Anne	HS-AP
Burt	HS-AP
Darrin	HS-DE
Jackie	HS-AP
Lisa	HS-DE
Daniel	Community College
Gerald	Community College
Ivy	Community College
Jabaar	Community College
Jack	Community College
Jacob	Community College
Todd	Community College
Alice	State-funded University
Dianne	State-funded University
Eric	State-funded University
Kumar	State-funded University
Susan	State-funded University
Richard	State-funded University

Table 2. Frequency counts by institution type for subcategories of the goal "Life Skills." Counts represent responses from the 10 instructors who described goals related to life skills without being prompted.

Life Skills	High School (N=5)	Community College (N=7)	State-funded University (N=6)	Total
Citizenship	1	2	0	3
Effort	2	1	2	5
Organization	0	1	1	2
Responsibility	3	1	0	4
Self-Directed	3	0	1	4
Study Skills	1	1	1	3
Time Management	0	1	1	2

Table 3. Frequency counts by institution type for the subcategories of the goal "Laboratory Skills."

Laboratory Skills	High School (N=5)	Community College (N=7)	State-funded University (N=6)	Total
Appropriate use of laboratory equipment	0	0	1	1
Measurement Skills	1	1	0	2
Observation	0	0	1	1
Predicting laboratory measurements and outcomes	1	0	0	1
Understanding laboratory and safety	1	1	0	2
Understanding and demonstrating appropriate laboratory technique	0	3	0	3

Table 4. Frequency counts by institution type for the subcategories of the goal "Data Analysis."

Data Analysis	High School (N=5)	Community College (N=7)	State-funded University (N=6)	Total
Graphing (Construction and/or Interpretation)	1	3	0	4
Interpretation of Data	0	1	0	1
Using Data as Evidence	0	0	1	1

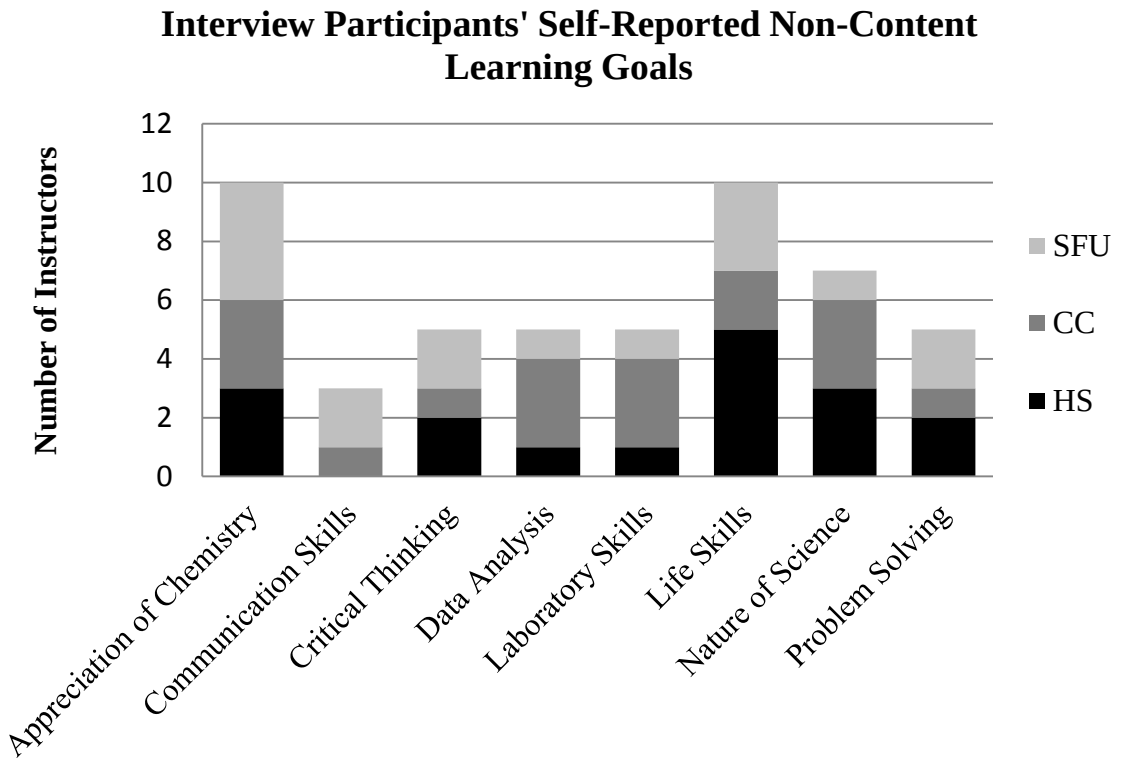


Figure 1. Interview participants self-reported non-content learning goals for their general chemistry course. Frequency counts for each goal are also broken down by institution type. State-funded university (SFU) is in light gray, community college (CC) is in dark gray, and high school (HS) is in black.

APPENDIX A: INFORMED CONSENT DOCUMENT

Dear Participant,

Below is a statement of informed consent indicating your rights as a participant in this study examining perspectives of learning goals in general chemistry courses. Please take a moment to read and review this statement prior to the start of the interview. You will be asked whether or not you agree to participate in the study at the beginning of the interview.

Statement of Informed Consent: "I understand that I will be asked questions about myself, my teaching, and my department. I understand that participation in this study is completely voluntary and I may choose not to answer a question at any time. I may withdraw from this study at any time. However, any responses I give prior to my withdrawal may be used in the study unless otherwise stated. I understand that all data collected will be de-identified and that there will be no way in which I can be linked to participating in this study. I understand that this interview will be tape recorded and transcribed in order to assure that my statements are represented accurately. I understand that the researchers may use quotes from this interview in publications or presentations, but my name will not be associated. I understand that this study has IRB human subjects approval (ISU IRB 12-402) and that if I should have any questions about my rights as a participant I may ask the researcher or the IRB office at Iowa State University."

Thank you. We look forward to speaking with you in the future.

Jessica J. Reed, Graduate Student Researcher

Thomas A. Holme, Professor and Principal Investigator

APPENDIX B: QUALITATIVE INTERVIEW GUIDE

Informed Consent Script:

Good morning/afternoon. Thank you again for agreeing to participate in this interview regarding faculty perspectives on learning goals and learning environments. Before we begin, I need to take a moment to explain your rights as a participant and get your consent.

Did you get a chance to read the statement of consent that I sent you via email?

If not, would you like me to read it to you now or read it yourself?

Now that I have read this statement of consent, please state your name, the date, and whether or not you agree to participate in the interview.

Do you have any questions before we begin the interview itself?

I would like to clarify that any questions or responses that contain specific identifying information are for my purposes only, and will be blackened out on the written transcript.

Faculty interview questions:

1) First, I'd like to ask you some questions about your teaching background.

   A) How long have you been teaching at your school?

   B) How long have you been teaching this particular general chemistry course? Could you give me the course number for my reference?

C) Are you currently teaching this course? If not, when was the last time you taught the course?

D) Are you teaching the lecture component, lab component, or both?

i. If teach lab….

How many labs do you do in the semester?

Are those labs traditional, guided inquiry, or a blend of both?

How do available resources influence what types of labs you are able to do?

E) What are the pre-requisites for this course? Co-requisites (e.g. are students required to enroll in lab or a certain math course)?

2) How would you describe the population of students that enroll in your course?

A) Average age and year in school?

B) Typical majors? (if applicable)

C) Do the majority of students take this course as a requirement for a specific major, or as an elective?

D) Typically, what is the chemistry background of students enrolled in this course? (never taken chem., taken AP, etc)

E) How would you describe the ability of the majority these students? (high achievers/honors, average achievement, low achievement, broad range across the class/too difficult to tell)

3) How often does the course meet and how long is each meeting?

4) Do you teach multiple sections of the same course? (How many?) Approximately how many students do you have in each section? How many total students are enrolled in the course?
5) Are there multiple instructors for this course?
    A) If yes, how many?
    B) If yes, is one instructor designated as the "lead instructor"? (Is that you?)
6) How does your department influence how this course is taught?
    A) Do they set the syllabus? Book? Content? Time on content?
7) How do you assess chemistry knowledge?
    A) Other than your own tests, do you use any external assessments?
8) What teaching pedagogies do you use in the course?
    A) Do you primarily lecture? Include group work? Use POGIL? Daily self quizzes? etc

Now, let's shift gears a little and talk about learning goals.

9) How would you describe a learning goal?
10) What learning goals do you have for this course?
    A) What content knowledge goals do you have for this course? (cognitive goals=content)
        i. What are the top 5 or 10 concepts students should master in your course?
        (e.g. After taking my course, students should know…)

B) Often we think of general chemistry courses as preparing students for higher level chemistry courses, or for other disciplines and career paths. As such, what non-content goals do you have for this course? In other words, how is this course preparing them for future endeavors in areas not related to content?

(psychomotor=skills, affective=attitude/emotional ties)

i. Prompt problem solving skills

(*I've heard a lot about problem solving skills recently, how does that fit into your objectives, if at all?*)

ii. Prompt attitude

(*How important are students' attitudes toward the subject of chemistry in your course?)*

iii. Prompt beliefs/confidence in self

(*What goals might you set for students' confidence in chemistry?*)

iv. Prompt expectations

(*How do students' expectations about learning chemistry influence the learning goals you set for the course?)*

11) Why do you feel these goals are important?

12) How did you determine that these would be the learning goals for your course?

A) Does your department indicate what the learning goals should be, or are you free to make your own?

B) If you have taught this course more than once, do you keep the same learning goals each time? Why?

13) If there is more than one instructor, how do you decide on learning goals for the course? Do you all share these same goals?

14) How are these learning goals expressed to students?

15) What strategies do you implement to ensure these learning goals are achieved? How do you implement these strategies?

16) How well do you feel that students meet your set of learning goals? (Do you measure this in any fashion?)

17) What do you feel are students' motivations for the course?

As mentioned in the recruitment email, part of my project is comparing the learning goals of courses that are deemed equivalent by several institutions. I am now going to ask you some questions comparing your course to other equivalent courses.

18) How might your learning goals be the same compared to:

A. High school AP or dual enrollment instructor's goals? How might they be different?

B. Community college instructor's goals (for the same course)? Different?

C. Large university? Different?

19) Why might you have similar learning outcome goals? Why might you have different learning outcome goals?

20) What do you feel are the strengths of (H.S. AP or Dual Enrollment, CC, and Large University) learning environments in terms of teaching and learning general chemistry?

21) What are the limitations of teaching of teaching in H.S. AP or Dual enrollment? CC? Large University?

22) How aware are you of the changes being made to the AP chemistry curriculum?

    A. How might these changes influence the goals you set for your course?

    B. What strengths and/or weaknesses do you foresee as resulting from these curriculum changes?

    C. Do you feel that an AP course will still be equivalent to a CC or Regents course after these curriculum changes?

23) Now that we've discussed these questions, is there anything else you'd like to see addressed in my study or any further comment you'd like to add?

24) A couple of final questions, would you be willing to send me a copy of your course syllabus via email? And, I am having difficulty finding dual-enrollment and community college instructors. Do you know of anyone within the state that might be willing to participate?

Thank you. That concludes the interview. Now that we are finished, do you have any questions for me? Again, thank you for your time and your honest answers to these questions. Your feedback is of great use to this study.

APPENDIX C: QUALITATIVE RESEARCH CODEBOOK

Code ID	Parent Node ID	Title	Description
1		Overarching learning goal	Goal that the participant sets as an overall goal for the course.
2		Student demographics	Any description relating to the type of students who enroll in the course.
3		Department LG	Learning goal that is used by the whole department or for all instructors who teach the course.
4		Desired or future learning goal	A goal that is not currently part of the course, but the instructor desires or plans to include it in the future.
5		Learning goal definition	Participant's definition of a learning goal.
6		Good quotes	A quote that may be useful to showcase data later.
7		Measurement of LG	Any description of how the learning goals are measured.
8		Location	The learning goal is designed to be achieved in a specific setting.
9	8	Both	Learning goal is emphasized in both lab and lecture.
10	8	Lab	The learning goal is designed to be implemented or assessed in the laboratory.
11	8	Lecture	The learning goal is designed to be implemented or assessed in the lecture setting of the course.
12		Inquiry	Any discussion of the term "inquiry" or concepts related to it (i.e. guided inquiry labs).
13	12	Experimental design	Students get to design aspects of their experimental procedure within the realm of guided inquiry.
14		Reasons for learning goals	Participant discusses why he or she chose these learning goals for the course.

15	14	Reasons prompted	Participant was prompted to discuss the reasons why he or she chose these learning goals for the course.
16		Importance of learning goals	Participant describes why he or she thinks these learning goals are important.
17	16	Prompted importance	Participant was asked, or prompted, to discuss the importance of these learning goals for the course.
18		Misconceptions	Description of students' misconceptions, either in general or relating to a specific topic.
19		Student entitlement	Description of students feeling entitled to certain grades, experiences, etc. simply because they are enrolled in the course.
20		Content learning goal	Any knowledge based learning goal related to course content.
21	20	Content application	Any mention of the ability to apply the content to a new or novel situation.
22	20	Unit LG	Any learning goal that the participant describes as being related to content in a certain unit or chapter.
23	20	Specific example of content	a specific example of a problem or goal related to a content topic
24	23	Naming	Being able to properly identify and name compounds or ions is stated as a goal.
25	23	Phases or States	Content knowledge goals include the understanding of phase changes and/or states of matter.
26	23	Stoichiometry	Determining stoichiometry via balanced chemical equations is described as a content knowledge goal.
27	23	Equilibrium	Mention of equilibrium and its related concepts as content learning goals for the course.

28	23	Electronic structure	The electronic structure of matter is described as a content knowledge goal.
29	23	Gases	Content knowledge goals related to properties or equations specifically associated with gases.
30	23	Chemical reaction	Participant describes chemical reactions as a part of content associated with the course.
31	23	Trends	Description of atomic and periodic trends that are part of course content goals.
32	23	Thermochemistry	Description of thermochemistry or any topic related to thermochemistry as it relates to content knowledge goals.
33	23	Atomic structure	Content knowledge goals relating to the structure and function of atoms.
34	23	Molecular shape	Content knowledge goals related to theories or methods of determining molecular shape or concepts related to/dependent upon molecular shape.
35	23	Moles	The concept of or use of moles is important in terms of content knowledge goals. This could include the use of the concept within calculations.
36	23	Bonding	Content knowledge goals related to the concepts of bonding.
37	23	Oxidation-Reduction	Concepts related to oxidation and reduction are stated as content learning goals.
38	23	Electrochemistry	Concepts related to electrochemistry are stated as a learning goal.
39	23	Nuclear chemistry	The concept of nuclear chemistry is discussed as a learning goal.
40	23	Organic chemistry	Concepts related to organic chemistry are learning goals for the general chemistry course.

41	23	Particulate Nature of Matter	Participant describes the particulate nature of matter as a learning goal.
42	23	Kinetic Molecular Theory	Description of kinetic molecular theory's incorporation into content learning goals.
43	23	Equations	Students know and are able to manipulate the proper equation for a given content topic (e.g., knowing the formula for calculating density).
44	23	Measurements	Content knowledge goal related to the concept of chemical measurements and proper use of significant figures.
45	23	Intermolecular Forces	Content knowledge goal relating to any type of intermolecular force and/or its properties.
46	23	Solutions	Content knowledge goals that relate to the specific properties, equations, or calculations associated with solution chemistry. This includes molarity, molality, and dilution.
47	23	Lewis Structures	Participant describes students' ability to draw Lewis Structures as a goal related to content knowledge.
48	23	Interactions of Matter	The interactions of matter is described as a content knowledge goal.
49	23	pH	A content learning goal related to pH is described.
50	23	Acid/Base	Participant describes content knowledge goals related to acid base chemistry.
51	20	Macroscopic vs Microscopic	Description of the microscopic and macroscopic views of chemistry.
52	20	Conceptual understanding	Description of the conceptual nature of chemistry and student goals for understanding.

53		Non-content Learning Goal	Any learning goal that is related to material or skills that are not chemistry specific.
54	53	Appreciation	Mention of the value of chemistry and getting students to appreciate it.
55	53	Communication	Description of communication as a necessary skill.
56	55	Inter-personal communication	This may include examples of students working in groups or giving presentations to improve communication skills.
57	55	Scientific communication	Description of development of scientific communication skills. This can include both verbal and written skills.
58	53	Critical thinking	Any mention of critical thinking as a skill used in the course.
59	53	Problem solving	Any description of problem solving goals in the course.
60	59	Specific examples	Examples of specific problems or situations that require the use of problem solving skills
61	59	Problem solving method	Description or mention of the method involved in solving a problem. This includes strategies or techniques associated with problem solving.
62	61	Information analysis	Student uses knowledge of information given in the problem (units, equations, etc.) and creates a means to solve the problem.
63	61	Factor Label Method	Description of solving problems by using the factor-label method
64	61	Problem identification	Student is able to identify that there is a problem that needs to be solved (they may not yet know how to solve it.)
65	61	Not explicit	Participant states that a specific method to problem solving is not explicitly taught to students.

66	61	Answer interpretation	Description of how the student is able to make meaning out of the answer to a problem and understand what the answer means/is telling them.
67	61	Pattern recognition	Students identify patterns in the problem in order to understand how to solve the problem and how it connects to other chemistry concepts.
68	59	Applying to new situations	Use or application of problem solving skills in new environments, or when faced with different types of problems than they have previously seen.
69	59	Practice problems	Used to define situations where students assigned or encouraged to complete problems for practice at improving and developing problem solving skills.
70	59	Use of equations	Any mention of students use of or ability to manipulate equations when solving problems.
71	59	Unprompted PS	The participant was not asked or prompted to talk about problem solving skills. The participant discusses PS of their own accord.
72	59	Algorithmic vs Conceptual	Description or mention of the algorithmic and conceptual nature of problem solving.
73	59	Prompted PS	The participant was asked, or prompted, to talk about Problem Solving learning goals.
74	53	Analysis of data	Any description of the use of data analysis in the course or laboratory.
75	74	Graphing	Description of skills related to the creation and/or interpretation and/or manipulation of graphs.
76	74	Use of data as evidence	The use of data to support a claim or serve as evidence in a justification of students' reasoning.

77	74	Interpretation	A learning goal related to the students' ability to interpret the data they have analyzed. This includes understanding how and when their data applies to a situation, and also what assumptions they may have made when analyzing the data.
78	53	Self-efficacy (SE)	Any mention of goals related to students' self-confidence in chemistry.
79	78	Prompted SE	Participant was asked, or prompted, to discuss goals relating to students self-efficacy or confidence in the subject of chemistry after taking the course.
80	78	Experiences	Participant describes how students' experiences with chemistry (past or present) influence their confidence.
81	78	Unprompted SE	The participant was not prompted to talk about self-efficacy or students' confidence, but brought it up on his or her own accord.
82	53	Attitude	Description or mention of attitude toward the subject of chemistry
83	82	Prompted attitude	Participant was asked, or prompted, to discuss goals related to students' attitudes toward the subject of chemistry
84	82	Unprompted attitude	The subject was not asked, or prompted, to discuss goals for students' attitudes toward the subject of chemistry.
85	53	Laboratory goal	Description of non-content learning goal that is specific to the laboratory setting.
86	85	Measurement skills	Related to making measurements, significant figures, or using measurement devices.
87	85	Observation	Skills related to making observations in the laboratory setting.

88	85	Technique	Proper laboratory techniques or skills are described as a learning goal.
89	85	Equipment use	Ability to demonstrate proper use of laboratory equipment is described as a goal.
90	85	Safety	Laboratory safety which includes proper handling and disposal of chemicals, PPE, and accident procedures are stated as non-content learning goal.
91	85	Prediction	Being able to predict outcomes based on what students already know and compare the prediction to what happens is a learning goal.
92	53	NOS	Relating to the nature of science
93	92	NOS not explicit	The participant describes ways in which the nature of science is presented to students, but does not state that he or she explicitly teaches the nature of science in the course.
94	53	Life skills	Goals related to skills that are transferrable to everyday life situations, such as skills needed for jobs.
95	94	Organization and planning	Any description or mention of planning or organizational skills
96	94	Effort	Description of effort related to success in the course
97	94	Study skills	Anything related to students' study habits
98	94	Time management	Skills related to time management such as minding deadlines or balancing work, school, and extracurricular activities.
99	94	Responsibility	Any description of responsibility as a skill or goal for the course
100	94	Attention to detail	Non-content goal of observing and paying attention to details.
101	94	Self-directed	Any discussion of students becoming self-directed or self-motivated learners

102	94	Independence	The participant describes a goal for students to gain independence by taking the course.
103	94	Scientific literacy	Goals related to creating scientifically literate citizens that are able to understand basic science presented in the media to make informed decisions.
104	94	Lifelong learning	The participant describes a goal for instilling a love for learning or lifelong learning into students.
105	94	Self-reliance	Learning goal related to students being self-reliant rather than depending on the instructor or others for guidance.
106	94	Citizenship	The participant explicitly states citizenship as a goal for the course.
107	53	Self-analysis	Any mention of students' analyzing their learning habits or skills. This does not include development of metacognitive skills.
108	107	Student expectations of chemistry	Description of what students expect to get out of their chemistry course.
109	53	Metacognition	The participant specifically describes how development of metacognitive skills is important to student development and growth. Participant says the word "metacognition."
110		Student expectations	Description of what students expect the course to be like or what to gain from it.
111	110	Prompted expectations	Participant was asked, or prompted, to discuss student expectations for the course.
112		Student perceptions of chemistry	Any description or discussion of how students' perceive the chemistry course to be.

113		Students' experiences	Any description or explicit discussion of students' experiences with chemistry (past or present). This can relate to classroom, lab, or other experiences involving chemistry that have shaped students' perceptions of chemistry.
114		Pedagogy	Participant mentions their pedagogies and/or how they use them in the classroom.
115		Content LG first	When asked about general learning goals for the course, the participant described content learning goals first.
116		Non-content LG first	When asked to describe learning goals for the course, the participant described non-content goals first.
117		Expressed to Students	The participant describes how the learning goals are expressed to students.
118	117	Textbook	The participant states that the textbook expresses the content learning goals to students. This can include the content of chapters, order of chapters, and/or the listing of key concepts within the text.
119	117	Non-Content NOT explicit	Non-content learning goals are not explicitly stated to students.
120	119	Teacher modeled	The non-content goals or skills are not formally stated to students, but are modeled by the teacher during normal class interactions.
121	117	Verbally	Participant describes verbally expressing learning goals to students, even if they have not been formalized.
122	117	Content formalized	Content goals are formalized and explicitly stated to students.
123	117	Non-content formalized	Non-content goals are formalized and explicitly stated to students.

124	117	Content goals not formalized	The participant does not acknowledge formally creating content learning goals and expressing them to students. Content LG may be expressed to students informally.
125	117	Syllabus/Hand-out	The participant describes using the syllabus or a handout as a means to express the learning goals to students.
126	125	Content LG syllabus/handout	The participant describes content knowledge learning goals are expressed on the syllabus or handout.
127		Students meet LG	The participant describes how well they feel that their students are meeting the learning goals that the instructor sets for the course.
128		LG assessed	The participant describes how the learning goals are assessed (or not assessed) within the course.
129	128	Content assessed	The participant describes methods in which the content learning goals are assessed.
130	129	Examinations	The participant describes examinations as a method of assessing the content learning goals in the course.
131	128	Non-content assessed	The participant describes how the non-content goals are assessed in the course.
132	131	Analysis of data LG assessment	The participant describes how the non-content learning goal related to analysis of data is assessed.
133		Student resistance	Descriptions of students' resistance to instructors incorporating learning goals.

CHAPTER 3: THE ROLE OF NON-CONTENT GOALS IN THE ASSESSMENT OF CHEMISTRY LEARNING

Jessica J. Reed and Thomas A. Holme

A book chapter published in *Innovative Uses for Assessments in Teaching and Research*, K. Murphy and L. Kendhammer, Eds., Washington, DC: ACS Books, 2015.

Abstract

As technology continues to make information and facts readily accessible, the importance of understanding the context of the information and demonstrating how to use it appropriately will provide better indications of learning than factual recall. This chapter examines the manner in which curriculum and assessment reforms are moving toward promotion of student skill development beyond traditional content knowledge recall. A discussion of the current state of non-content skill assessment in chemistry is presented noting in particular that instructor interest in non-content aspects of learning appears to outpace the measurement of them. Additionally, the chapter presents data from a national survey. These data were used to understand the relative importance of non-content goals and skills in the general chemistry classroom. How these data will inform future efforts to create appropriate formative and summative assessments of goals and skills beyond content knowledge is also discussed.

Introduction

In a world where facts are accessible with a click of a button, simple factual recall is no longer the appropriate principle indicator of learning. Rather the context of the

knowledge and the ability to use it appropriately are of greater importance. Official reports that use this premise to call for various education reforms have been prominent components of policy debates (Hess, Kelly, & Meeks, 2011). Not surprisingly, calls for curriculum reform in chemistry often echo these sentiments. One theme for implementation of suggestions such as these notes the need for data-driven and evidence-based curriculum and assessments (Cooper, 2010, 2013; Lloyd & Spencer, 1994; National Research Council, 2012).

Beyond the policy calls, and at least partly in response to them, several efforts to revise science curricula have arisen. Among the most ambitious are the recent changes in both the curriculum and tests associated with Advanced Placement (AP)® courses in several sciences, including chemistry (College Board, 2011a, 2011b, 2014). In this case, developers at College Board have shifted to an evidence-based approach to curriculum design that utilizes Evidence Centered Design (ECD) (Mislevy, Almond, & Lukas, 2003) along with principles of "backwards design" (Brennan, 2010; Huff, Steinberg, & Matts, 2010). In this model for curriculum design, learners are expected to master not only content essential to the understanding of scientific concepts, but additionally meet expectations about what they should be able to do with that knowledge (Mislevy, 2011). In order for ECD to accomplish its goals, assessments need to be carefully constructed in order to measure whether a learner has successfully achieved all of the desired outcomes for the course beyond recall of factual knowledge. The current state of this curriculum development process is described in the re-designed AP chemistry curriculum by College Board (2013). A key component of this approach lies in the definition of learning

objectives (LOs) that were specifically created to integrate "essential knowledge" (content) and "science practices."

The Next Generation Science Standards (NGSS) are designed in similar fashion to the reforms of AP courses at the high school level. The ultimate goal of the NGSS is to aid science education at the K-12 level by describing what all students should know and be able to do by certain grade levels (Achieve, 2013). While there is no standardized curriculum or assessment associated with the NGSS, the interconnectedness of core content, practices, and crosscutting concepts implies that assessments will need to measure all three cohesively.

Regardless of the intended audience of the reform effort, it is evident that attempts to move beyond simple factual recall assessments to develop rich assessments that measure the development of student skills and practices are becoming increasingly commonplace. The effects of such efforts promise to change how chemistry is taught and assessed at the post-secondary level, as future generations of college students may enter the classroom prepared to engage with the content in different ways. With this potential future in mind, the goals of general chemistry instruction and assessment at the collegiate level should be prepared to consider the development of content knowledge and to encompass development of skills and practices that students can transfer to other courses and disciplines.

What such a curriculum and assessment regime might look like in practice is not yet established in the literature. The concept of considering curriculum development in conjunction with assessment reform has been proposed (Holme et al., 2010) where

assessment design is driven by curriculum prerogatives, and assessment data informs changes in curriculum. This is not to say that multiple modes of assessment have not already been developed within chemistry. Nonetheless, evidence suggests that many chemistry faculty members are aware of a relatively small number of assessment methods and instruments (Emenike, Raker, & Holme, 2013; Raker, Emenike, & Holme, 2013; Towns, 2009).

Currently, efforts within the chemistry education research community are seeking to provide means for assessment of student performance beyond content. Assessment instruments used in chemistry education include several that are not directed strictly at content knowledge measurement. For example, an instrument to measure student attitudes about learning chemistry, Attitude toward the Subject of Chemistry Inventory (ASCI), was created by Bauer (2008). The instrument measures students' attitudes by asking students to select the position on a semantic differential that most closely relates to their perceptions of chemistry. Xu and Lewis later refined the instrument to a shorter version which measures fewer constructs than the original (Xu & Lewis, 2011). Other instruments such as CHEMX (Grove & Bretz, 2007) and CLASS (Adams, Wieman, Perkins, & Barbera, 2008) focus on students' expectations and beliefs about chemistry. The CHEMX instrument aims to compare student expectations of the chemistry learning environment to those of faculty within the context of a specific chemistry course, including the laboratory. The CLASS instrument compares student beliefs about chemistry in general to those of experts. While some of the constructs measured by the two instruments overlap, each instrument measures a unique piece of the chemistry experience from the perspective of students. Additionally, the Metacognitive Activities

Inventory (MCAI) measures students' metacognitive awareness and how that awareness influences chemistry problem solving skillfulness (Cooper & Sandi-Urena, 2009; Cooper, Sandi–Urena, & Stevens, 2008). It is important to note that this summary highlights only a small fraction of the number of published instruments available for use in chemistry instruction. While these assessment instruments do not specifically intertwine the measurement of chemistry content knowledge with content independent skills, they are important for use in classroom contexts to understand better the development of specific attitudes and skills by students.

The number and variety of assessment instruments that have been developed illustrates the apparent demand for assessment measures that go beyond content knowledge. To some degree, however, instrument development has tended to result in only modest implementation. In other words, the number of times in which non-content aspects of learning have been explored in a preliminary way via instrument development is growing, but day-to-day usage of such tools has shown a less robust pattern, at least in terms of literature (Arjoon, Xu, & Lewis, 2013). This does not imply an outright lack of interest in the measurement of non-content learning goals. Indeed, usage of assessment tools in classrooms that go unreported in the literature may be common. Nonetheless, from the literature base alone, it is not easy to ascertain the key non-content characteristics chemistry instructors feel are important to measure. Therefore, it is important to 1) understand what skills and practices general chemistry instructors value for students to develop and 2) think about how future assessment designs might incorporate essential content with skill assessment.

Arguing the Importance of Non-Content Assessment in Chemistry

Beyond the impetus from emerging curriculum development and studies within chemistry education research, there are two important aspects of chemistry instruction that suggest the measurement of non-content goals may be important. First, theories of how people learn have repeatedly included key components that are not formally related to content knowledge alone. Second, for many forms of pedagogical improvement, an increase in non-content components of learning may be important. In this sense, the potential importance of measuring non-content goals follows a familiar theory and practice breakdown that can be elaborated further.

Theories of Learning and the Role of Non-Content Assessment

Novak's theory of education, human constructivism, is integral to the design and analysis of this research (Bretz, 2001; Novak, 1977). Novak draws heavily upon the ideas of psychologist and philosopher David Ausubel's assimilation theory which describes the differences between rote and meaningful learning, outlines the conditions necessary for meaningful learning, and suggests that meaningful learning occurs when the learner is afforded experiences in each of the three learning domains (cognitive, affective, and psychomotor) (Ausubel, Novak, & Hanesian, 1968). Meaningful learning is achieved only when all three components are present.

Novak's theory asserts that knowledge is a human construction, and thus it is incumbent upon the educational system to support learners as they construct knowledge (Novak, 1977). Additionally, meaningful learning empowers students to commit and be responsible for learning by integrating thinking, feeling, and acting. Therefore, this

framework provides a lens to analyze the learning goals of general chemistry instructors because it establishes a basis to understand how the learning goals provide an opportunity for meaningful learning in a general chemistry course (Bretz, Fay, Bruck, & Towns, 2013).

It is also important to consider that the general chemistry classroom provides experiences that are unique to the discipline of chemistry. That is to say that the learning that occurs within the general chemistry classroom is situated within the context of a chemistry community. Thus it is useful to consider that activity, concept, and culture found within the chemistry classroom are interdependent. The theory of situated cognition provides an additional lens for understanding the role of activity to develop skill and concept creation within the realm, or culture, of general chemistry (Brown, Collins, & Duguid, 1989). It is posited that even though students acquire tools, or skills, they will not know how to use them appropriately if not given opportunities to use them within the context of the discipline (Brown, et al., 1989). This suggests that even though opportunities for meaningful learning may be presented to students, the knowledge and skills acquired may remain decontextualized, and even inert, unless students are presented with insight about how those concepts and skills are actually used within chemistry and how to transfer them to applicable real-life situations (Roth & McGinn, 1997).

Additionally, the importance of the interconnection of content knowledge and procedural skills in understanding learning is shown within the Unified Learning Model (ULM) (Shell et al., 2009). The ULM provides a model of how people learn, and a resultant model of teaching and instruction, by drawing on the principles of cognitive

science and psychology. In this model, working memory, knowledge, and motivation are central to understanding how all people learn. Knowledge in this case refers not only to concepts or facts (declarative knowledge), but also to the skills, behaviors, and thinking processes that an individual knows (procedural knowledge). Learning is then influenced by the individual's working memory capacity, the concepts and skills he or she already knows (prior knowledge), and the goals that drive him or her to put forth effort. In this model the instructor aids the learner by directing the student's attention (working memory) to the concept or skill to be learned, providing opportunities for the creation of new connections between prior knowledge and the new concept or skill, and creating goals to support the motivation of the student to learn. In this sense the instructor serves as a mere facilitator of individual learning, yet guides the course of the learning experience by influencing the content and skills developed through specific course goals and objectives.

Practical Implications of Measuring Non-Content Learning in the Classroom

There is little question that content knowledge gains represent the main goal of any science course, and chemistry courses are no exception. However, it is also true that understanding just how teaching methods influence learning often hinges on non-content aspects. In particular, the concepts of student engagement, student motivation or student persistence have received considerable attention in research studies regarding how to promote learning success in chemistry (Sadler, Sonnert, Hazari, & Tai, 2012; Seymour, Hewitt, & Friend, 1997; Zusho, Pintrich, & Coppola, 2003). Perhaps just as importantly, the measurement of non-content variables is often measured as a part of formative assessment during attempts at curriculum or pedagogical innovation. Determining

whether or not students "like" a new approach is often reported – but it is arguable that non-content learning can be parsed with significantly more resolution than this construct.

Several teaching methodologies have emerged with an intention to improve content learning and provide non-content gains as well. Within chemistry, Process Oriented Guided Inquiry Learning (POGIL) is perhaps the most prominent example (Chase, Pakhira, & Stains, 2013; Farrell, Moog, & Spencer, 1999; Hein, 2012; Minderhout & Loertscher, 2007). For this teaching method, the process-orientation component is focused on enhancing the development of generalizable process skills that allow students to gain more content knowledge. Other teaching methods such as problem based learning (Overton, Byers, & Seery, 2009), case-based historical development of chemical concepts (Obenland, Munson, & Hutchinson, 2013) and active learning via a "flipped" classroom (Smith, 2013) all include aspects that relate to student engagement and non-content skill development.

While a number of research questions related to the assessment of the non-content components of these emerging methodologies still remain, the methodologies themselves serve to exemplify the practical nature of enhancing student skills in addition to content knowledge.

Before researchers can address creation of assessment materials for measurement of non-content goals and skills, it is necessary to understand what are the goals and skills that chemistry instructors value. The survey and data presented here aim to inform the community about the types of goals and skills that are valued in the general chemistry curriculum.

Methods

Survey Development

Quantitative survey items were developed from themes present in qualitative interviews conducted with chemistry instructors about the learning goals present in introductory general chemistry courses. A detailed discussion of the qualitative research study is provided in the previous chapter. The semi-structured interviews were conducted with 18 general chemistry instructors from high schools, community colleges, and state-funded universities. Participants were asked open-ended questions that progressively became more specific depending on a participant's response, such as "What are the learning goals you have for your general chemistry course?" to "What are the non-content goals you have for students in your course?" The interviews were then transcribed and open-coded using a Grounded Theory approach (Glaser & Strauss, 1967). Additionally, learning goals were labeled according to the primary domain (cognitive, affective, or psychomotor) associated with the goal. Interestingly, participants often discussed incorporating a variety of goals into their courses, but felt that students did not meet the often implicit expectations associated with these goals even though they did not formally assess their non-content goals. In order to obtain more generalizable results about the status of non-content learning goals, the ten most frequently discussed non-content goals from the interviews were transformed into survey items. The survey items were part of a national online survey from the ACS Examinations Institute about conceptual understanding in general chemistry.

The major non-content goals surveyed were: appreciation of chemistry in everyday life, development of communication skills, laboratory skills, graphing of data, interpreting and drawing conclusions from data, life skills (e.g., study skills, responsibility, time management), problem solving skills, nature of science (i.e., how science works and has developed), critical thinking, and conceptual understanding of traditionally algorithmic problems.

Survey Items

Three questions on the survey related to non-content goals and each question evaluated all ten non-content skills identified as common themes amongst qualitative interview participants. These questions are found in the Appendix immediately following this chapter.

The first question related to learning goals asked participants to indicate how often they intentionally and explicitly incorporated the learning objectives into their course. Response choices ranged from "I do not incorporate this" to "Every class period," with options of "Once or twice per semester," "Once per month," and "Once per week" in between. Participants were only able to select one response choice per learning goal.

The second question in the set related to how the learning goals were assessed in the course. Participants were asked to select all modes of assessment that applied to each learning goal. Methods of assessment surveyed were clickers (student response systems), exams, homework, laboratory reports, and quizzes. Additionally, response options were available for participants that did not assess or did not incorporate a goal in the course.

The final question related to learning goals asked participants to describe, on average, how well they felt that students met their expectations for these learning goals. Respondents were allowed to choose one response from five choices ranging from "Below my expectations" to "Exceeds my expectations."

Sample

The sample consisted of chemistry instructors and faculty at community colleges, four-year colleges, and universities in the United States who had taught a general chemistry course within the past five years. Institutional classifications were based upon the self-reported highest degree offered in chemistry at the participant's institution. The sample excluded instructors of special topics courses and General, Organic, and Biochemistry (GOB) courses. For analysis purposes, only participants who completed all questions relating to learning goals were considered as part of the sample (N = 1,075). Table 1 shows participant distribution by institution type. General chemistry teaching experience of participants ranged from one year to 40 years of experience, with an average of approximately 15 years. Additionally, 84% of the sample had taught a full-year (two-semester) general chemistry course and 75% were responsible for teaching both a lecture and laboratory component of the course.

Results and Discussion

Quantitative Survey Results and Discussion

Results from the survey provided insight into chemistry instructors' values of non-content goals and skills.

Responses to the first question about frequency of intentional incorporation of non-content learning goals were as expected. Skills traditionally associated with chemistry courses, such as conceptual understanding, critical thinking, and problem solving, were reported to be incorporated into every class period by a majority of instructors. Figure 1 displays the frequency of incorporation of the non-content goals and skills as self-reported by instructors. Problem solving appeared to have the highest frequency of incorporation. Approximately 74% of instructors reported incorporating problem solving into every class period, and an additional 22% reported incorporating it on a weekly basis. Less than 1% (0.28%) of instructors reported not incorporating development of problem solving skills as a goal of their general chemistry course. Critical thinking and conceptual understanding also had a majority of respondents indicate that they incorporate those skills into every class period with 58% and 56%, respectively. Additionally, nearly 70% of instructors reported incorporating laboratory skills on a weekly basis. This is consistent with the typical general chemistry course design, which includes a weekly laboratory section. Other goals, such as development of communication skills, showed a broader range of reported incorporation.

While these statistics are not surprising due to the nature of general chemistry coursework, it is important to note that these data are self-report so we cannot ascertain for certain whether instructors are actually incorporating these goals in the manner in which they claim. For example, while over 95% of instructors claim to incorporate problem solving into their course at minimum on a weekly basis, it is unclear as to whether participants in this survey were differentiating the nature of problem solving, such as how the course activities might be compared with students performing learning

exercises (Bodner, 1987). Such distinctions are not wholly necessary for this study because these data were not meant to assert sweeping observations about the condition of the collegiate general chemistry classroom. Rather, the objective is to understand the types of goals and skills that are valued by general chemistry instructors in an effort to understand better the types of non-content skills that future formative and summative assessments could be designed to measure. In this context, it is considered that an instructor who makes an effort to incorporate a goal or skill more frequently likely values that skill more and desires to develop it in students more so than goals that are incorporated on a less frequent basis.

The frequency with which instructors reportedly incorporate non-content goals and skills in their courses provides an indication of the types of skills they hope to develop in their students. Yet, incorporation of a goal into a curriculum does not imply that students successfully develop that skill. Assessment plays a key role in understanding and rating student skill development. In order to understand better how future assessments might be designed to measure content independent learning goals, it was important to elicit how instructors assess non-content goals within their general chemistry courses. Again, these are self-reported data intended for use to understand how instructors perceive these learning goals to be assessed. Respondents were allowed to select multiple modes of assessment for a single learning goal. The modes of assessment were selected from the most frequent responses collected in qualitative interviews, and included clickers, exams, homework, lab reports, and quizzes. Respondents were also allowed to indicate that a particular learning goal was not assessed in their course.

Instructors' responses regarding modes of assessment used can be seen in Figure 2. For ease of interpretation, responses have been combined to reflect summative assessments (exams and quizzes), formative assessments (homework and clickers), laboratory reports, and responses indicating a goal was not assessed. It is of interest to note that laboratory reports were the most frequent response for assessment of communication skills, laboratory skills, graphing of data, and drawing conclusions from data, whereas problem solving skills, critical thinking about concepts or problems, and conceptual understanding of problems traditionally solved algorithmically are reported as most commonly assessed by exams and quizzes.

Other methods of assessment were not selected as frequently. For example, clickers make up a smaller fraction of the formative assessment category compared to homework. Clickers had minimal use in assessment of the non-content goals except for problem solving. This result may not be surprising in light of previous research about clicker usage among chemistry instructors (Emenike & Holme, 2012). Goals related to development of an appreciation of the subject of chemistry, understanding of the nature of science (NOS), and life skills were reported as most frequently not assessed in any fashion.

Instructors reported use of assessments gives insight into how opportunities for meaningful learning are being evaluated in the classroom. Skills that lie predominantly in the cognitive domain (problem solving, conceptual understanding, and critical thinking) are reported as most frequently assessed by exams, whereas skills that lie predominantly within the psychomotor domain, with some overlap of the cognitive domain, such as laboratory skills, communication skills, and graphing are measured by laboratory reports.

Affective goals such as appreciation of chemistry and life skills are reported as not assessed at all. While it is not surprising that there is disconnect between the methods of assessment (or lack thereof) for each domain, it is indicative of the challenge faced by assessment designers to incorporate more than one domain within a single format of assessment.

Regardless of how the learning goals are purportedly assessed, there appears to be room for improvement in student performance. Instructors were asked to evaluate how students met expectations regarding successful development of these learning goals, and their responses can be seen in Figure 3. Although the percentage of students meeting the expectations of their instructors for development of these non-content goals was generally over half, a sizable fraction of students appear to have fallen short in the estimation of the participants in this survey. Indeed, more instructors rated student performance as "Does not meet expectations" than "Exceeds expectations," suggesting that there is room for improvement in student performance in non-content aspects of learning. It is important to remember, however, that assessment methods that instructors have indicated are used for non-content goals tend to be more informal. As such, the impressions they form (which presumably inform their answers to this survey item) may lack quantitative rigor. Thus, the expectations reported here, while informative about future challenges related to assessment of non-content learning, should not be considered a rigorous judgement of student non-content learning.

Conclusions

Although it may not be routinely articulated by chemistry instructors, the development of skills beyond the scope of content knowledge in chemistry courses is important and most instructors view it as such. Curriculum reform efforts often influence non-content learning outcomes but without a more rigorous effort to enhance assessment it may be argued that these changes essentially resort to a "hope for the best" approach. The survey research presented here provides evidence that non-content learning goals are valued by the chemistry education community. As such, assessments are needed to measure the development of students' skills beyond typical content exams.

Calls for changes in the chemistry curriculum focus on the need for evidence-centered and data-driven reform efforts (Cooper, 2010, 2013; Lloyd & Spencer, 1994; National Research Council, 2012), beyond measuring whether students "like" an activity. Instruments have been developed to measure student skills beyond the domain of chemistry content knowledge; however, these instruments appear to be underutilized by the traditional chemistry community, perhaps due to a lack of awareness of these instruments. Additionally, these instruments tend to be quite specific and measure only specified constructs. This means that to gain a whole picture of the classroom environment, an instructor would likely need to devote significant effort to administering and analyzing survey instruments. This level of effort may not be practical in the typical general chemistry classroom.

Ultimately, the most attractive trajectory for addressing the need for non-content assessment may lie in finding ways to incorporate it more closely within traditional

content assessments. Efforts to devise such assessment are part of the high profile developments in AP Chemistry (Brennan, 2010; College Board, 2011a, 2011b, 2014; Mislevy, 2011; Mislevy, et al., 2003) and the Next Generations Science Standards project (Achieve, 2013). In order to guide such development the current work suggests an iterative process may be particularly helpful to determine what non-content skills are most important to assess in this way. Instructors appear to be interested in gaining better information about student learning, but it seems reasonable to expect that initial attempts to measure non-content aspects may require refinement. Thus, the collaboration between curriculum reform efforts and assessment development efforts (Holme, et al., 2010) will take on ever more importance as chemistry education moves forward over the next few years.

References

Achieve. (2013). Next generation science standards: The National Academies Press.

Adams, W. K., Wieman, C. E., Perkins, K. K., & Barbera, J. (2008). Modifying and validating the Colorado Learning Attitudes about Science Survey for use in chemistry. *Journal of Chemical Education, 85*(10), 1435.

Arjoon, J. A., Xu, X., & Lewis, J. E. (2013). Understanding the state of the art for measurement in chemistry education research: Examining the psychometric evidence. *Journal of Chemical Education, 90*(5), 536-545.

Ausubel, D. P., Novak, J. D., & Hanesian, H. (1968). *Educational psychology: A cognitive view*. New York, NY: Holt, Rinehart, and Winston.

Bauer, C. F. (2008). Attitude toward chemistry: a semantic differential instrument for assessing curriculum impacts. *Journal of Chemical Education, 85*(10), 1440.

Bodner, G. M. (1987). The role of algorithms in teaching problem solving. *Journal of Chemical Education, 64*(6), 513.

Brennan, R. L. (2010). Evidence-Centered Assessment Design and the Advanced Placement Program®: A Psychometrician's Perspective. *Applied Measurement in Education, 23*(4), 392-400.

Bretz, S. L. (2001). Novak's theory of education: Human constructivism and meaningful learning. *Journal of Chemical Education, 78*(8), 1107.

Bretz, S. L., Fay, M., Bruck, L. B., & Towns, M. H. (2013). What Faculty Interviews Reveal about Meaningful Learning in the Undergraduate Chemistry Laboratory. *Journal of Chemical Education, 90*(3), 281-288.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational researcher, 18*(1), 32-42.

Chase, A., Pakhira, D., & Stains, M. (2013). Implementing process-oriented, guided-inquiry learning for the first time: Adaptations and short-term impacts on students' attitude and performance. *Journal of Chemical Education, 90*(4), 409-416.

College Board. (2011a). The AP biology curriculum framework. New York: The College Board.

College Board. (2011b). The AP chemistry curriculum framework. New York: The College Board.

College Board. (2013). AP® Chemistry: Course and Exam Description. New York, NY: College Board.

College Board. (2014). The AP physics curriculum framework. New York: The College Board.

Cooper, M. M. (2010). The case for reform of the undergraduate general chemistry curriculum. *Journal of Chemical Education, 87*(3), 231-232.

Cooper, M. M. (2013). Chemistry and the next generation science standards. *Journal of Chemical Education, 90*(6), 679-680.

Cooper, M. M., & Sandi-Urena, S. (2009). Design and validation of an instrument to assess metacognitive skillfulness in chemistry problem solving. *Journal of Chemical Education, 86*(2), 240.

Cooper, M. M., Sandi–Urena, S., & Stevens, R. (2008). Reliable multi method assessment of metacognition use in chemistry problem solving. *Chemistry Education Research and Practice, 9*(1), 18-24.

Emenike, M., & Holme, T. A. (2012). Classroom response systems have not "crossed the chasm": Estimating numbers of chemistry faculty who use clickers. *Journal of Chemical Education, 89*(4), 465-469.

Emenike, M., Raker, J. R., & Holme, T. (2013). Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *Journal of Chemical Education, 90*(9), 1130-1136.

Farrell, J. J., Moog, R. S., & Spencer, J. N. (1999). A guided-inquiry general chemistry course. *Journal of Chemical Education, 76*(4), 570.

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory; strategies for qualitative research*. Chicago,: Aldine Pub. Co.

Grove, N., & Bretz, S. L. (2007). CHEMX: An instrument to assess students' cognitive expectations for learning chemistry. *Journal of Chemical Education, 84*(9), 1524.

Hein, S. M. (2012). Positive impacts using POGIL in organic chemistry. *Journal of Chemical Education, 89*(7), 860-864.

Hess, F. M., Kelly, A. P., & Meeks, O. (2011). The case for being bold: A new agenda for business in improving STEM education. *Institute for a Competitive Workforce, Washington, DC*.

Holme, T., Bretz, S. L., Cooper, M. M., Lewis, J., Paek, P., Pienta, N., et al. (2010). Enhancing the role of assessment in curriculum reform in chemistry. *Chemistry Education Research and Practice, 11*(2), 92-97.

Huff, K., Steinberg, L., & Matts, T. (2010). The promises and challenges of implementing evidence-centered design in large-scale assessment. *Applied Measurement in Education, 23*(4), 310-324.

Lloyd, B. W., & Spencer, J. N. (1994). The forum: New directions for general chemistry: Recommendations of the task force on the general chemistry curriculum. *Journal of chemical education, 71*(3), 206.

Minderhout, V., & Loertscher, J. (2007). Lecture-free biochemistry. *Biochemistry and Molecular Biology Education, 35*(3), 172-180.

Mislevy, R. J. (2011). Evidence-Centered Design for Simulation-Based Assessment. CRESST Report 800. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). A brief introduction to evidence-centered design. *ETS Research Report Series, 2003*(1), i-29.

National Research Council. (2012). *Discipline-based education research : understanding and improving learning in undergraduate science and engineering*. Washington, D.C.: The National Academies Press.

Novak, J. D. (1977). *A theory of education*. Ithaca, NY: Cornell University Press.

Obenland, C. A., Munson, A. H., & Hutchinson, J. S. (2013). Silent and vocal students in a large active learning chemistry classroom: Comparison of performance and motivational factors. *Chemistry Education Research and Practice, 14*(1), 73-80.

Overton, T. L., Byers, B., & Seery, M. K. (2009). Context-and Problem-based Learning in Higher Level Chemistry Education. *Innovative Methods of Teaching and Learning Chemistry in Higher Education*, 43-59.

Raker, J. R., Emenike, M., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education, 90*(8), 981-987.

Roth, W.-M., & McGinn, M. K. (1997). Deinstitutionalising school science: Implications of a strong view of situated cognition. *Research in Science Education, 27*(4), 497-513.

Sadler, P. M., Sonnert, G., Hazari, Z., & Tai, R. (2012). Stability and volatility of STEM career interest in high school: A gender study. *Science Education, 96*(3), 411-427.

Seymour, E., Hewitt, N. M., & Friend, C. M. (1997). *Talking about leaving: Why undergraduates leave the sciences* (Vol. 12): Westview Press Boulder, CO.

Shell, D. F., Brooks, D. W., Trainin, G., Wilson, K. M., Kauffman, D. F., & Herr, L. M. (2009). *The unified learning model: How motivational, cognitive, and neurobiological sciences inform best teaching practices*: Springer Science & Business Media.

Smith, J. D. (2013). Student attitudes toward flipping the general chemistry classroom. *Chemistry Education Research and Practice, 14*(4), 607-614.

Towns, M. H. (2009). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education, 87*(1), 91-96.

Xu, X., & Lewis, J. E. (2011). Refinement of a chemistry attitude measure for college students. *Journal of Chemical Education, 88*(5), 561-568.

Zusho, A., Pintrich, P. R., & Coppola, B. (2003). Skill and will: The role of motivation and cognition in the learning of college chemistry. *International Journal of Science Education, 25*(9), 1081-1094.

Table 1: A description of quantitative survey participants by institution type.

Survey Participant Demographics

Institution Type	Participants	Percent of Sample
Community College	170	15.8
Bachelor's Institution	513	47.7
Graduate Institution	392	36.5
Total	1,075	100

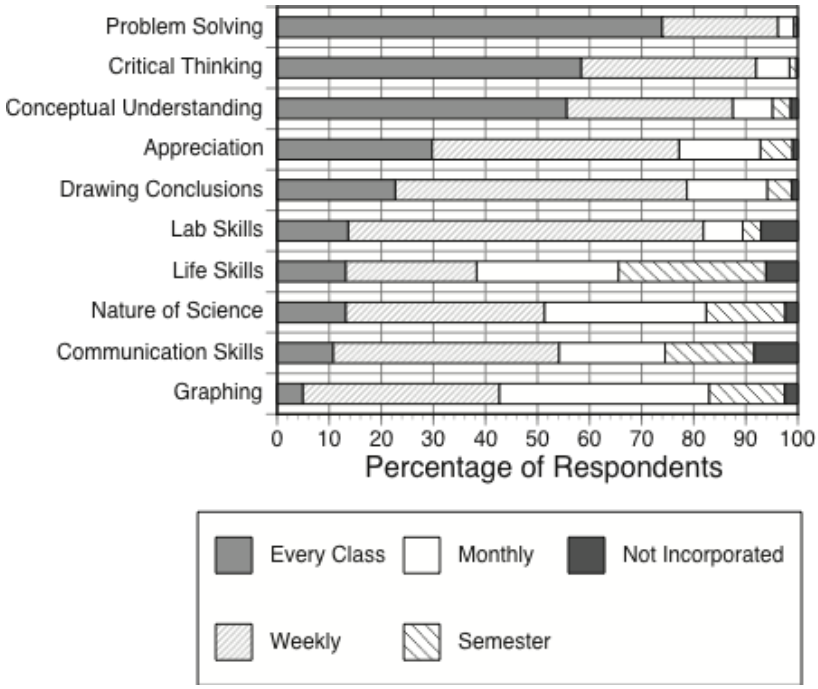


Figure 1. General chemistry instructors' self-reported incorporation of non-content goals and skills. Incorporation ranges from every class meeting to not incorporated at all.

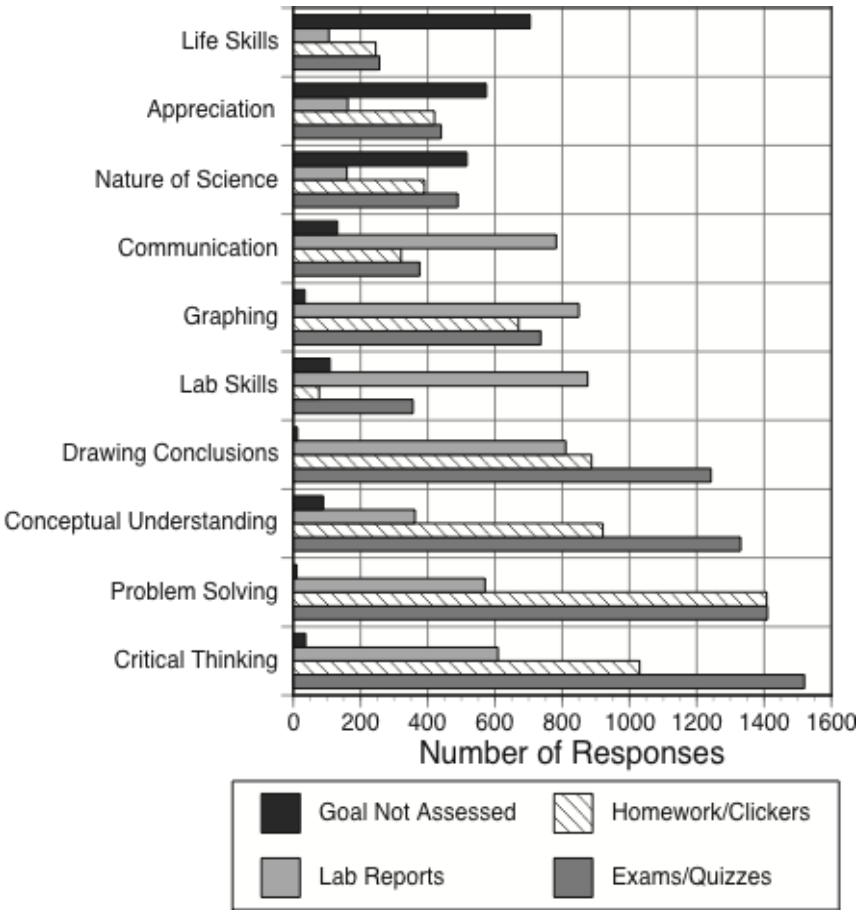


Figure 2. Instructors' self-reported methods of assessment of content independent goals and skills in general chemistry courses.

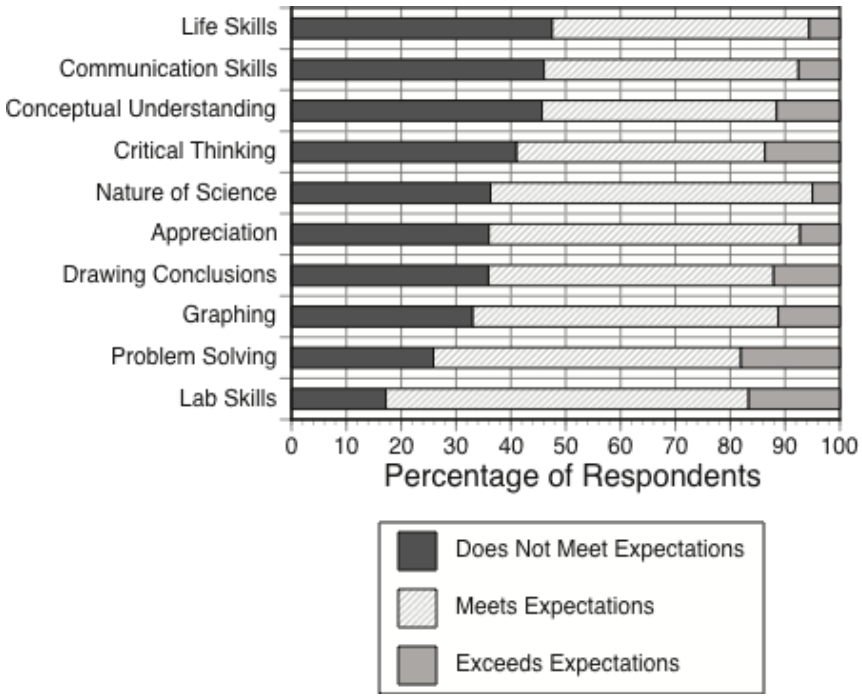


Figure 3. Instructors' evaluation of student performance on achievement of non-content learning goals.

APPENDIX: QUANTITATIVE SURVEY ITEMS

The following three items were the last items on a survey related to conceptual understanding in general chemistry administered by the ACS Examinations Institute during 2013. Radio buttons or check boxes were provided for response selection. The survey was administered electronically through SurveyMonkey.

25. How often do you intentionally and explicitly incorporate the following objectives/goals into your course?

	I do not incorporate	Once or twice per semester	Once per month	Once per week	Every class period
Appreciation of the subject of chemistry in everyday life					
Development of communication skills (e.g., scientific writing, collaboration)					
Development of laboratory skills (e.g., technique, safety habits)					
Graphing data					
Interpreting and drawing conclusions from data					
Life skills (e.g., study skills, responsibility, time management)					
Problem solving skills					
Nature of science (e.g., how science works and has developed)					
Thinking critically about concepts or problems					
Understanding concepts behind problems traditionally solved algorithmically					

26. Please select all methods you use to assess each of the objectives/goals.

	Clickers	Exams	Homework	Laboratory Reports	Quizzes	I do not assess this	I do not incorporate
Appreciation of the subject of chemistry in everyday life							
Development of communication skills (e.g., scientific writing, collaboration)							
Development of laboratory skills (e.g., technique, safety habits)							
Graphing data							
Interpreting and drawing conclusions from data							
Life skills							
Problem solving skills							
Nature of science (e.g., how science works and has developed)							
Thinking critically about concepts or problems							
Understanding concepts behind problems traditionally solved algorithmically							

27. On average, how well do you feel your students meet your expectations for these objectives/goals?

	Below my expectations	Slightly Below	Meets my expectations	Slightly above	Exceeds my expectations
Appreciation of the subject of chemistry in everyday life					
Development of communication skills					
Development of laboratory skills					
Graphing data					
Interpreting and drawing conclusions from data					
Life skills					
Problem solving skills					
Nature of science (e.g., how science works and has developed)					
Thinking critically about concepts or problems					
Understanding concepts behind problems traditionally solved algorithmically					

CHAPTER 4: MODIFICATION AND USE OF A NOVEL RUBRIC TO ANALYZE STANDARDIZED CHEMISTRY EXAM ITEMS FOR INCORPORATION OF SCIENCE PRACTICES

Jessica J. Reed, Alexandra R. Brandriet, and Thomas A. Holme

A paper to be submitted to the *Journal of Chemical Education*

Abstract

Recent reforms in science education have generated interest in the importance of measuring skills beyond content proficiency in assessments. More specifically, ability to engage in science practices as defined by *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (National Research Council, 2012) is of interest because it provides evidence of what students should be able to do with their content knowledge. The research herein analyzed chemistry assessments for the presence of science practices through the modification and use of a newly developed rubric designed to aid practitioners and researchers in the evaluation and creation of assessment materials that incorporate measures of science practices. By analyzing standardized American Chemical Society exams for incorporation of science practices, the research reports how large-scale chemistry assessments are currently making use of science practices and suggests how future assessment development may make use of these findings to create more explicit measures of science practices.

Introduction

The current climate of assessment, including within science education, is based upon an action and reaction model. Actions to change practices of assessment provide evidence to spur curricular reform efforts, which then generate reactive reforms in the

realm of assessment (Holme et al., 2010; Murphy, Picione, & Holme, 2010). This cyclic model to assessment reform has played out extensively as instructors seek alternative measures to large-scale traditional forms of assessment. The traditional, well-established forms of assessment, such as standardized tests, are often ineffective at detecting individual changes in curriculum or pedagogy (Koretz, 2002). Thus, instructors are often left to their own devices to create assessments to measure the effectiveness of their reform efforts, often resulting in evidence that students "like" a particular activity or method of teaching, but often lacking in evidence of validity and reliability (Holme, 2011). These assessments, while informative to the individual instructor, provide little evidence that students have learned the content or skills desired (Bodner, MacIsaac, & White, 1999; Holme, et al., 2010). Additionally, instructor created assessments may not meet the psychometric objectives of traditional, large-scale assessments because they are designed to meet the needs of the individual instructor's efforts. This is not to say that these forms of assessment are inherently bad, just that they are not effective measures of large-scale reform efforts, because they engage a bottom-up assessment approach which becomes too convoluted for sense-making by the time it progresses across curricula and disciplines. A more meaningful approach would be a top-down assessment design in which new assessments are specifically designed to be sensitive to nuances of student learning affected by the change in pedagogy or curriculum. In this manner, the cyclic nature of reform still occurs, but in a fashion that is controlled more by research and evidence, as the curricular reform drives the assessment development and the assessment results meaningfully validate the curricular change. So how do these considerations relate to science education reforms, specifically within chemistry education?

Current reform efforts in science education, particularly within K-12 curricula, support the need for reconsideration of the goals and aims of large-scale assessment. Considering historic reform efforts in science education demonstrates the longstanding desire to bridge the gap between curriculum and assessment to provide evidence of student performance beyond basic content knowledge (Achieve, 2013; American Association for the Advancement of Science, 1989; National Research Council, 1996). A particularly important recent effort is embodied in a report entitled *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas*, and referred to herein as the *Framework* (National Research Council, 2012). One of the biggest outcomes from the *Framework* is the creation of the Next Generation Science Standards (NGSS) (Achieve, 2013). The NGSS, if implemented, will change the face of science education and assessment by providing specific outcomes of K-12 science education. By outlining what students should know and be able to do with that knowledge, the NGSS support the efforts of assessment reforms by providing targets for measurements beyond content proficiency. Additional reforms such as those to the Advanced Placement® science curriculum (College Board, 2011a, 2011b, 2014), and even changes to the Medical College Admissions Test® (Association of American Medical Colleges, 2014; Kirch, et al., 2013), add to the clear and consistent message suggesting the need for substantial changes in the expectations of what students should know and be able to do with that knowledge in science. Of particular interest is how these expectations will be assessed, especially within realm of traditional forms and modes of assessments, such as multiple choice items. Failure to effectively assess both the content and skills associated with these reforms could diminish the possibility for coherence across the K-16 science

education landscape (Pellegrino, 2012). Efforts to define how these new forms of assessment will look in relation to the NGSS are taking shape (Pellegrino, et al., 2014), however, the ability of current assessments to measure components of the NGSS has yet to be established in the literature.

Chemistry has certainly not been excluded from the calls to curricular and assessment reform efforts. Despite numerous attempts and calls to action, little change was observed in introductory chemistry courses at the postsecondary level (Lloyd, 1992) until recent curriculum development projects (American Chemical Society, 2005; Cooper & Klymkowsky, 2013; Talanquer & Pollard, 2010) within chemistry education sought to provide evidence based reform and restructuring of the general chemistry curriculum. While implementation of these curricula in introductory chemistry courses is currently modest, they represent a transition away from the emphasis of content breadth and isolated concept application to a curricular approach in which depth of knowledge and application of skills associated with chemistry are emphasized. Such curricula also support the need for redesigned assessments to measure beyond the broad strokes of content recall and algorithmic application of chemical principles. In order for future assessments to meet the challenges of assessing skills and practices beyond content, it is important to understand the current status of assessments in relation to skills measurement, particularly within chemistry.

Traditional assessments in chemistry, such as exams, quizzes, and homework, have focused almost exclusively on content. This is not to say that assessment materials to measure skills beyond content knowledge do not exist in chemistry, merely that they are incorporated infrequently relative to traditional forms of assessment often because

instructors are unaware such assessments exist or how to use them appropriately (Emenike, Raker, & Holme, 2013; Raker, Emenike, & Holme, 2013; Towns, 2009). Highlights and examples of such assessment instruments can be found in the previous chapter, or in Holme, et al (2010). Additionally, research within the chemistry education community highlights the need for quality assessments that are economical in terms of the time required for administration and interpretation of performance results by instructors (Emenike, Schroeder, Murphy, & Holme, 2013). Thus, practicality and necessity dictate that future assessments consider the combination of content and skill measures within one assessment design. In order to design such assessments in the future, understanding the current status of the presence of measures of skills beyond content knowledge in chemistry examinations is important. Standardized chemistry examinations from the American Chemical Society Examinations Institute (ACS-EI) are ideal for such an evaluation because they are widely used within the chemistry community, and are highly regarded in terms of item construction and overall exam quality (Holme, 2003). The exams are large-scale assessments within the chemistry domain, and consist of multiple-choice items developed by committees of chemistry practitioners (Holme, 2003). Analysis of ACS exams for the incorporation of practices associated with the development of scientific skills is the predominant focus of this research. Knowing how these exams incorporate measures of scientific practices provides two-fold benefits. First, it provides evidence in support of current reform efforts to incorporate such skills more explicitly into performance expectations, as in the NGSS, and second, it informs future assessment designers how such practices may be embedded with multiple-choice items

on large-scale assessments. The analysis of the exams is detailed in the Methods section of this chapter.

In order for curriculum and assessment reform efforts to move beyond a tangential relationship, dissemination of quality assessment materials is necessary. Consideration of the role of large-scale assessment in this movement is prudent, as these assessments have the potential to reach the broadest audience and often set the standard for what is to be taught. By and large, however, these assessments are typically limited to multiple-choice items which have seen their fair share of complaints about measuring learning in what some deem a limited fashion (Archbald & Newmann, 1988; Linn, 2001; Sacks, 2000). Yet, others assert the value of multiple-choice assessments when they are constructed properly and highlight their potential (Burton, 2005; Haladyna, 2012; Haladyna, Downing, & Rodriguez, 2002; Little, Bjork, Bjork, & Angello, 2012).

Chudowsky and Pellegrino (2003) outline the considerations necessary for large-scale assessment to support learning and inform curriculum reforms. Additionally, Pellegrino (2014) suggests that measurements of a broad range of competencies, such as science practices, are "essentially untapped by current assessments" (p. 71). These suggestions support the idea that large-scale assessment, if reformed, will be able to accommodate measures beyond content proficiency. Rather than dismiss multiple-choice exams based upon their perceived limitations, the research herein focuses on the merits of multiple-choice items for measuring dimensions of learning beyond traditional content knowledge in order to support science education reform.

Research Framework

Three-Dimensional Learning

The *Framework* describes three-dimensional learning to mean the blending of Scientific Practices (SP), Crosscutting Concepts (CC), and Disciplinary Core Ideas (DCI) (National Research Council, 2012). It is important to note that three-dimensional learning is not meant to be used as a learning theory to describe how students learn, rather it is used to describe how a learning environment can be modeled to address three dimensions of content learning. The Next Generation Science Standards (NGSS) were constructed from these concepts of three-dimensional learning, and provide detailed descriptions of how three-dimensional learning should progress in the K-12 science curriculum (Achieve, 2013). It is important to note that three-dimensional learning and the NGSS do not mandate the use of a particular pedagogy, rather they provide a means for developing curriculum and assessment materials that challenge students to learn meaningfully by intertwining each dimension. It should be considered that these reform efforts are fairly recent, so it is not expected that current endeavors in chemistry instruction and assessment would encompass all of the dimensions of this framework. Of the three dimensions of the NGSS, however, science practices are perhaps the most important to consider in an established assessment program, such as the American Chemical Society examinations analyzed herein, as they are most likely to have some connection to current assessment practice.

Crosscutting concepts are ideas that transcend the boundaries of science, mathematics, engineering, and technology. These concepts bridge disciplinary boundaries and provide an organizational framework to connect discipline specific content

knowledge to broader scientific views and understandings. Often students are left to construct these connections on their own as they transition through disciplinary science courses. By incorporating crosscutting concepts as a dimension of three-dimensional learning, the authors of the *Framework* aim to signify the importance of explicit instructional supports needed while constructing interdisciplinary connections (National Research Council, 2012).

The seven crosscutting concepts are outlined in Table 1, and more detailed descriptions of each, along with progressions for their development in K-12 science curriculum, are provided in the *Framework* (National Research Council, 2012). The research herein did not focus on incorporation of crosscutting concepts in assessment due to the complex nature of the concepts themselves and the difficulty isolating definitions of each concept suitable for a rubric evaluation tool. Additionally, the primary component of this research focuses on the analysis of exam items that have not been developed with the interdisciplinary connections in mind. This task will likely become easier in the future as more empirical research is conducted on the status of crosscutting concepts and their disciplinary definitions.

Another dimension of three-dimensional learning is disciplinary core ideas. These ideas were deemed to be fundamental to the understanding of science content. The *Framework* provides rich detail about the specific core ideas for the disciplines of physical science, life science, earth and space science, and engineering and technology (National Research Council, 2012). The four disciplinary core ideas for physical science are outlined in Table 2. Each core idea also has additional component ideas that narrow the focus of the DCI.

While the DCI for physical sciences do include fundamental concepts within the discipline of chemistry, they are very broad and, by design, function at a K-12 level. For the purpose of this research exploring the status of three-dimensional learning in college chemistry assessments they were not used, rather a set of chemistry specific core ideas was used instead. A more detailed discussion of the chemistry specific DCI is found in the Methods section.

Science practices link knowledge and skills to articulate what a student should be able to do with science content knowledge. The *Framework* defines eight science practices, outlined in Table 3, and delineates their application in science and engineering (National Research Council, 2012). The research herein focuses only on the practices as they relate to science education, particularly within the discipline of chemistry.

Special care is taken by the authors of the *Framework* to denote that the term "practices," rather than "skills," is used "to stress that engaging in scientific inquiry requires coordination both of knowledge and skill simultaneously" (National Research Council, 2012, p. 41). Additionally, the progression of development of the practices across the K-12 science curriculum is described, and goals for what a student should be able to do by the end of grade 12 in relation to each practice are defined.

Three-dimensional learning provides a basis for understanding the future of curriculum and assessment reform, particularly in K-12 science education. A view of competence in science is provided through the overlap of each of the three dimensions. Vague terms such as "know" or "understand" are eliminated in favor of terms that encompass the practices of science with core content, such as analyze, argue, explain,

predict, and represent. Thus, three-dimensional learning provides a powerful lens for evaluating the status of measurement beyond content in current assessment efforts.

Research Questions

This study aimed to answer the following primary research questions:

1. Have current assessment efforts in chemistry incorporated science practices?
2. What science practices are most frequently incorporated in chemistry assessments?

As occurs in most research, additional questions that are related to these two main research questions arose as analysis was carried out. These additional questions will also be noted in describing this project.

Methods

Assessment Items

The primary source of data for this research project came from standardized chemistry exams developed by the American Chemical Society Examinations Institute (ACS EI). ACS exams carry secure exam copyrights, so exam security is of utmost importance. Therefore, ACS exam items cannot be shown herein. Instead, mock questions that are similar in content and construct have been created to serve as examples for discussion purposes. ACS exams are developed by committees of chemical practitioners and are not governed by item specifications from the ACS EI, other than the content must be appropriate for the level of chemistry the exam is to be associated (Holme, 2003). Therefore, exam items assess the content that the exam committee deemed important to measure, and provide a unique snapshot of what the chemistry

community values in assessment (Holme, 2003; Luxford & Holme, 2015; Luxford et al., 2015)

A variety of ACS exam items were used in this study with the intent to investigate how science practices are being incorporated primarily across the general chemistry curriculum. Before exams are released, they go through trial testing and items that do not perform well are not included in the final version of the exam. Analysis of trial test items that were not released for science practices was also conducted on some exams in order to gauge how science practices are incorporated into the exam design process as a whole. Trial test items for all of the released general chemistry conceptual exams, in addition to the trial test items for not yet completed for release 2015 general chemistry conceptual exam, were analyzed with the rubric described in the following section. Additionally, trial test items for the 2013 general chemistry exam were analyzed. No other trial test items for full-year general chemistry exams were analyzed because it was found that the items on the trial tests were routinely algorithmic in nature and nearly identical to the released exam items. In total, 12 released ACS exams (N = 735 items) and 12 trial exam forms (N = 401 items) were analyzed for incorporation of science practices and phenomena. Table 4 displays descriptive information about the types of exams used in this analysis.

Exams were selected for analysis for various reasons. The GCC exam series was selected due to the conceptual manner in which chemistry content is assessed on the exams, because it was hypothesized that conceptual type items might incorporate science practices more frequently than traditional items. This is also why the PQ exams were included in the analysis. The manner in which general chemistry content is assessed on

the GCC and PQ exams is in contrast to the types of questions on the GC exams. The GC exam series is the most frequently released, so the number of exams in the series was deemed too large to analyze. Thus, a subset of the series whose release dates corresponded closely to those of the GCC exams was analyzed. The GCF exam was analyzed because it is being used in other ACS EI projects in which knowledge of the science practices incorporated on the exam may be useful. The DUCK and CIC exams represent other facets of chemistry knowledge assessment and were included to broaden the scope of chemistry content analyzed. The LAB exam was included in the analysis to understand whether certain science practices may lend themselves more readily to laboratory assessments.

Data Collection

Two chemical education researchers worked together to rate chemistry assessment items for incorporation of science practices and phenomena. Raters reviewed assessment items independently and then convened weekly to discuss their ratings. In cases where the raters disagreed, a discussion was had until the raters reached 100 percent agreement. Occasionally, the raters would need to consult with a third rater, an assessment expert, in order to resolve a rating dispute.

In order to understand the status of incorporation of science practices in chemistry examinations, it was necessary for the raters to have a rubric for how each of the science practices outlined by the *Framework* is to be incorporated into assessment. Collaboration with researchers at Michigan State University allowed for the use and revision of an assessment protocol designed to evaluate the three dimensions of learning discussed by

the *Framework.* Use of the Three Dimensional Learning Assessment Protocol (3D-LAP) allowed the two raters to efficiently and effectively classify chemistry items containing science practices. The creators of the 3D-LAP describe its purpose as two-fold: 1) the 3D-LAP is designed to characterize the alignment between three-dimensional learning and formative and summative assessments, and 2) to guide creation and redesign of assessments to provide explicit evidence of student understanding (Cooper, 2014b; Underwood, Cooper, Krajcik, Cabellero, & Ebert-May, 2014).

The rubric was designed for use across all science disciplines, and as such, needed minor modifications in order to provide specific criteria for classification of chemistry-specific items. In particular, the raters found it necessary to elaborate on the definitions of science practices associated with developing and using models (SP2), analyzing and interpreting data (SP4), using mathematics and computational thinking (SP5), constructing explanations (SP6), engaging in argument from evidence (SP7), and obtaining, evaluating, and communicating information (SP8). Modifications primarily focused on adding descriptions for how to determine if a specific science practice was present, particularly in a multiple choice chemistry exam item. Additionally, a new category was added to the rubric under the practice of planning and carrying out investigations (SP3). The new category (3d) relates to the understanding of how to use scientific equipment and techniques appropriately. Analysis of "Crosscutting Concepts" was beyond the scope of this study, therefore the research herein focused only on the operationalization of "Science Practices." The rubric used by the rating team is found in the Appendix to this chapter.

Classification of Assessment Items

The 3D-LAP was used to analyze each of the items on the ACS exams selected for science practices and phenomena. The raters agreed on a pattern of classifications in order to keep ratings consistent. For example, on occasion where a graph was present in an item, the researchers considered how a student would have to interact with the graph in order to answer the question when determining a science practice classification for the item. If the graph was a plot of theoretical values in which the student had to explain or predict an event or observation, the item was considered to contain the practice of "Developing and Using Models." If the graph represented data collected through an experiment in which the student had to conduct and/or interpret some form of the analysis, then the item was considered to contain the practice of "Analyzing and Interpreting Data." These clarifications were added to the 3D-LAP to ensure consistency in use. Items were not limited in the number of science practices that could be assigned, and occasionally multiple science practices were associated with an item. The majority of items, however, had only one science practice present, if any. The maximum number of science practices in one item was three.

If an item had a science practice, the next step was to determine whether the practice was "explicit" or "implicit" within the item. The goal is to have items that explicitly measure science practices in addition to content knowledge. A practice was considered to be present explicitly if all criteria for the practice contained in the 3D-LAP rubric were met. In the case where one criterion was to "provide the reasoning link" between pieces of information or representations, it was determined that the criterion was met when selection of the correct choice (or a distractor) clearly indicated a student's

understanding of the chemistry content. Denotation of the explicitness of a practice aids in informing the status of current use of science practices within ACS exam items. Since the exam items were all multiple-choice and already incorporated into a released exam, it was not possible to change the items to have the science practice be explicitly demonstrated. In the future, the 3D-LAP could be used in a different scenario to design assessment items, and in those instances revisions to make the science practice explicit could be made.

In addition to analyzing the items for science practices, it was also important to determine whether the item had a phenomenon present. Phenomena aid students in relating the content of the item to events or observations at the macroscopic scale. This helps students to create connections between chemistry content and everyday experiences, allowing them to learn more meaningfully. Determination of whether the item contained a phenomenon was predicated on whether the item contained an event, experiment, or data observable at the macroscopic scale. For example, a thermochemistry item discussing the use of a Styrofoam coffee cup calorimeter in an experiment would exemplify a phenomenon, whereas a similar item that simply listed numerical values associated with the calculation of heat transfer would not. Presence of a phenomenon in an item was classified on a binary scale, and was independent of the presence of a science practice within the item.

Disciplinary Core Ideas (DCIs) were not discussed by the two raters because many of the exam items had already been content aligned to the ACS Anchoring Concepts Content Map (ACCM) in a prior study (Luxford, et al., 2015). Due to the nature of the development of the ACCM (Holme & Murphy, 2012; Murphy, Holme, Zenisky,

Caruthers, & Knaus, 2012), it was presumed that the Big Ideas of the ACCM would correspond to the DCI of chemistry. The ten Big Ideas of the ACCM are: Atoms, Bonding, Structure and Function, Intermolecular Forces, Chemical Reactions, Thermodynamics, Kinetics, Equilibrium, Experiments, and Visualization.

Mock items to demonstrate the use of the 3D-LAP are shown in Figure 1 and Figure 2. The sample items are mock ACS items, and were developed for use on a national survey conducted by the ACS EI (Brandriet & Holme, 2015). Mock Item 1 (Figure 1) represents a more conceptual type of item. The item contains a particulate nature of matter (PNOM) diagram, graphs, and a symbolic equation of the reaction, and the student must use all three cohesively to answer the question. First, the raters would discuss whether a science practice was present. In this item, a student would need to interpret the PNOM diagram and relate it to the graphs. This corresponds to the science practice "Developing and Using Models," and more specifically, "2b" of the 3D-LAP. The use of the models is implicit, however. The item does not meet all three of the criteria iterated in "2b," because it does not ask students to provide the reasoning link between the representation and their prediction (response choice). Additionally, the item contains the science practice "Obtaining, Evaluating, and Communicating Information," or "8b" on the 3D-LAP. In this case, in order to answer the question, a student must translate between the symbolic chemical equation and the PNOM diagram, followed by a translation of the chemical information into a graphical representation. This translation is representative of practice "8b," but it is an implicit practice because the student is not required to justify or explain the need for the translation. After the science practices have been identified, the presence of a phenomenon would be determined. Mock Item 1 does

not include a phenomenon, because there are no macroscopic details or descriptions in the item stem or response choices.

Mock Item 2 (Figure 2) represents a more traditional algorithmic item. The item presents a scenario in which a student must calculate the new volume of a balloon filled with gas after an increase in temperature. Upon review, the item does not contain a science practice because students can, and most likely do, solve the problem via an algorithmic use of mathematics. The item is situated within the context of a gas filling a macroscopic balloon, so the item would be classified as having a phenomenon. The cartoon balloon image associated with the item stem is not relevant for solving the problem, but does add to the macroscopic context of the item.

The two mock items provide a means for explaining how the 3D-LAP was used when reviewing ACS exam items for science practices and phenomena. The variety of items found on ACS examinations cannot be fully described in such a small number of examples, but the two mock ACS items are representative of the typical styles of items found on ACS examinations.

Data Analysis

All data were compiled in a master data file in Microsoft Excel. Statistical analyses were conducted with STATA® version 13 (StataCorp, 2013). Additional data related to item content alignment with the ACCM, discrimination, and difficulty was obtained where available. Due to the nature of exam development and analysis, additional psychometric data were not available for every exam analyzed by the 3D-LAP. For reference, Table 5 classifies the exams based on the item statistics available.

One of the primary analyses conducted was to determine what science practices are being incorporated into ACS exam items. Basic statistics such as frequency counts and percentages were most useful for this analysis. Additional analysis examined item difficulty, ρ, which is the proportion of students who answered the item correctly (Alagumalai & Curtis, 2005; Ding & Beichner, 2009). For reference, items with a difficulty index below 0.30 are traditionally considered difficult, because below this cutoff random guessing could give a similar result (*e.g.* 0.25 for a four response item), while items 0.80 are considered easy (Ding & Beichner, 2009). Other statistical tests and results are discussed further in the results and discussion section.

Results and Discussion

Presence of Science Practices

The presence of science practices was examined across a variety of ACS exams. In this section, analyses presented pertain to exams in the following categories: CIC, DUCK, GC, GCC, GCF, PQF, and PQS. Table 4 provides more information about each type of exam. The LAB exam and trial tests for released exams will be discussed in later sections. In total, the present analysis spans 11 released exams and 695 unique items.

Of the 695 items analyzed, 287 (41.3%) contained at least one science practice. Of the 287 items that contained a science practice, 254 (88.5%) items contained only one science practice, while 30 (10.5%) items contained two science practices, and 3 (1.0%) items contained three different science practices. In total, 323 unique occurrences of a science practice were observed in the 287 items containing science practices. The number

times each science practice occurred and the distribution across 3D-LAP classifications are tabulated in Table 6.

From this distribution it is easy to see that the science practice of Developing and Using Models (SP2) occurs most frequently, at least five times more often than any other science practice. The inclusion of models and representations is frequent within ACS exams in the form of Lewis Structures, graphs, and PNOM diagrams, so it is not unexpected that a science practice involving their use would inherently be incorporated as well. Another finding was also expected, but still of interest. Examples of all of the science practices except Asking Questions were found within the items analyzed. It is not surprising that Asking Questions was not found as a practice within these items due to how chemistry content is traditionally assessed, especially within a multiple-choice context, however, it provides evidence for a practice that is underrepresented in these assessments.

So, how well are chemistry assessments doing at incorporating science practices? While it may seem modest, the fact that roughly 40% of items analyzed contained some form of a science practice supports the idea that these practices are of value to the chemistry community, even though exam writers were not explicitly aware of, nor trying to include science practices, within the exams.

In terms of assessment reform, understanding of how students perform on items containing science practices is critical. Good assessment practice requires that tests are aligned with the cognitive domain to be assessed. If instructors do not emphasize, and therefore students do not expect tests to include explicit measures of science practices,

there would be a mismatch between the assessment domain and the teaching domain that would inherently introduce measurement error in the test. An analysis of item performance was conducted to investigate how items with science practices compare to those without in terms of item difficulty, in part because ACS items are currently not designed to overtly incorporate science practices, and in part because it is reasonable to expect that including science practices in test items may introduce additional challenges for students. The average difficulty of items with science practices is approximately ρ = 0.55. ACS exams are constructed and norm referenced such that average performance is commonly between 50-60%, thus it makes sense for average item difficulty values for items that contain science practices to be near this range as well. Items with and without a science practice were compared with an independent samples *t*-test, and the difference in average difficulty between items with a science practice (M = 0.556, SD = 0.163) and without (M = 0.551, SD = 0.165) was not significant, t (638) = -0.358, $p > 0.05$. This suggests that the current methods of incorporating science practices, even inadvertently, into ACS items have little to no impact on average student performance on these items, which is an important factor to consider when designing future assessments to explicitly contain science practices.

The majority of science practices were present implicitly (81%) rather than explicitly (19%). Determination of how explicit a practice was within an item is detailed in the methods section, but it stems from how criteria in the 3D-LAP rubric are met within an item. Many items did not meet all of the criteria because they did not ask the student to provide reasoning links between components of the item and answer selection. This means that even though a science practice is present, it is not explicitly known how

a student uses the practice to engage with the content of the item. Since the items are all multiple choice there are limits on the amount of information that can be contained within the response choices, so these findings are consistent with the nature of the exam. Nonetheless, future exam development committees may consider how to include more explicit evidence of how a student uses a science practice embedded within an item.

It was important to determine how phenomena were incorporated into exam items, as they provide a bridge between the particulate level chemistry content and macroscopic level understandings of everyday events. The presence of a phenomenon in an item may help students relate the chemistry content to personal schema about macroscopic observations and experiences, and as such, aid in the practice of meaningful learning. However, care should be taken to ensure that the incorporation of a phenomenon does not substantiate misconceptions of chemical knowledge. Overall, approximately 46% of the items in this analysis contained a phenomenon, regardless of whether a science practice was present. About 17% of items contained both a science practice and a phenomenon. Analysis of frequency of a phenomenon relative to a specific science practice did not provide any unique insights as presence of a phenomenon was approximately equally likely across all practices. Inclusion of phenomena in items is more likely due to goals that exam development committees often have to improve the connection between chemistry and "the real world" than to a specific reason relating to content or practices.

Science Practices across Exam Types

Incorporation of science practices varied by the type of exam analyzed. Figure 3 shows the percent of items with and without science practices by type of exam analyzed. The LAB, GCC, and DUCK exams had the greatest relative percentage of items containing a science practice, with 84%, 76%, and 68%, respectively. The GC and GCF exams had the lowest relative percentage of items with a science practice, 25% and 21%, respectively. These differences are likely due to the nature of the exams themselves. The GCC and DUCK are designed to assess a more conceptual understanding of chemistry content, while the GC and GCF tend to rely more heavily on measures of traditional, or algorithmic, problem solving methods. In order to understand these trends, it is important to consider the history of exam development in these areas.

The GCC exam series was started in 1996 in response to the literature supporting a difference between conceptual and algorithmic learning, particularly the findings of Nurrenbern and Pickering (1987; Pickering, 1990). At this time, much of conceptual learning in chemistry was thought to relate to visual representations and models, which explains the large percentage of items with that science practice on the GCC 1996 exam. Prior to this time, the GC exam series was the only exam series to assess general chemistry content, and as such, contained conceptual and algorithmic items. It is speculated that as the conceptual exam series became more established, the GC series no longer needed to provide as many conceptual items because there were now exam products designed to assess conceptual understanding of general chemistry content. Additional studies have supported the idea that students often rely on heuristics, rather than conceptual understanding of chemistry topics to make reasoning judgements when

solving chemistry problems (Maeyer & Talanquer, 2010; McClary & Talanquer, 2011; Talanquer, 2006, 2014). All of these studies point to a difference in conceptual versus algorithmic problem-solving methods in chemistry that are evident in the types of items that appear within ACS general chemistry exams. Additionally, exam committees tend to use the distribution of item types on the most recently released exam to build upon for development of a new exam, so if at one point in time an exam did not include many conceptual items, the trend may be perpetuated until committee members champion the inclusion of conceptual items again. Thus, the number of items assessing a science practice decreased over time in the GC exam series.

Further investigation of the items that contained science practices revealed how the science practices were distributed throughout ACS exams. Figure 4 shows the distribution of the percentage of exam items that contain a specific science practice, as defined by 3D-LAP, compared to the type of ACS exam analyzed. SP2b was present in the highest relative percentage of exam items on the DUCK and GCC exams, 40% and 36% respectively. Due to the design of the DUCK exam in which items are embedded within scenarios that often include visual representations or models of data or molecules, this finding is consistent with the design parameters for this exam. All other science practices appeared much less frequently. The practices of constructing explanations (SP6a) and engaging in argument from evidence (SP7a) each occurred in less than 10% of items, and occurred most frequently in the conceptual items of the GCC and PQ exams. The practice of using mathematics and computational thinking, specifically 3D-LAP "5b" (SP5b), occurred most frequently on the paired questions exams. The unique design of these exams allows for the pairing of traditional computation items with

conceptual items that require explanation of the reasoning used to solve the computation. The 3D-LAP specific practice "8b" (SP8b), related to obtaining, evaluating, and communicating information, focuses on the translation between types of representations, and was most commonly found on the GCC exams, which frequently contain items related to translating between PNOM diagrams and chemical equations. Practices related to planning and carrying out investigations, and analysis and interpretation of data, were present infrequently. Perhaps this is due to the nature of the exams in this analysis, since many of them do not contain content related to the laboratory or experimentation, which is most likely where these practices are being developed and assessed. Considering that these exams were not developed to intentionally incorporate science practices into the assessment items, the distribution and variety of science practices across exam types, while modest, is nonetheless impressive.

The analysis of science practices across types of exam raised the question of how incorporation of science practices varies across time within a particular exam series. The GC and GCC exams were the only series in which multiple exams were analyzed, so they were used in this comparison. Figure 5 shows the distribution of the percentage of GCC items across science practices. The distribution appears to be fairly stable over time. Trends show that the practice of using and interpreting models decreased over time, but constructing explanations increased. These trends are represented by fairly small numbers of items, and are more likely to be considered "noise" than important differences. All of the GCC incorporated each of the most common seven science practices in some fashion. Trends related to the incorporation of science practices in GC exams are shown in Figure 6. The trends of incorporation of science practices in the GC exams appear to represent

trends beyond the realm of "noise." A steady decline in practices related to constructing explanations and engaging in argument from evidence, but an increase in the use of models over time is observed. In order to understand these trends, it is important to consider the history of these exam development efforts as described previously in this section. Knowledge about conceptual versus algorithmic learning (Nurrenbern & Pickering, 1987; Pickering, 1990) promoted changes in exam construction which may explain the trends observed over time. This analysis provides a snapshot of the incorporation of science practices over time, and for the GC exam series represents only a fraction of the total number of GC exams released.

Overall, the analysis of a variety of ACS exams for the incorporation of science practices showed that science practices were incorporated, at least to some extent, into all of the exam types analyzed. The GCC exams and the DUCK exam had the greatest number and variety of science practices incorporated, while traditional content exams, GC and GCF, contained the fewest number of science practices. The importance of this analysis stems from the status of incorporation of science practices in ACS exam items. The realization being that science practice concepts are so much a part of the discipline of chemistry and chemistry education in higher education that even without explicitly trying to do so, items on ACS exams implicitly include science practices. These observations may be of use to ACS exam development committees should they choose to incorporate more explicit measures of science practices into exam items.

Science Practices across ACCM Big Ideas

The ten Big Ideas of the ACCM can be argued to represent the DCI of chemistry. Therefore, it was important to understand how one dimension of three-dimensional learning, science practices, intertwines with another dimension, disciplinary core ideas. The items included in this analysis had been aligned to the ACCM in a previous study (Luxford, et al., 2015). The previous study created a historical database of the most commonly used general chemistry exams, thus content alignment data were only available for the GC, GCC, and GCF exams, a total of 465 unique items. Nuances in how items are assigned to content domains on the ACCM and science practices on the 3D-LAP necessitated delineation of items with multiple practices or Big Ideas. In the current analysis, items that had more than one science practice were included in the dataset multiple times so that each science practice was represented individually, bringing the total number of items to 488. Additionally, if an item was aligned to more than one location on the ACCM, only the primary location was included in this analysis. The primary location indicates the major content theme within the item, so it allows for the most accurate comparison between science practices and content. The number of items containing a science practice compared to the total number of items aligned to a Big Idea is shown in Table 7. It is important to note that some areas of the ACCM tend to be represented more than others in exam items on general chemistry exams (Luxford & Holme, 2015), so the relative percentage of items with a science practice within a Big Idea is also included within the table.

Items containing a science practice represented 189 (38.7%) items in this analysis. A distribution of specific science practices across the ten Big Ideas is depicted in

Figure 7. In general, multiple science practices were found within each Big Idea, with the exception of Bonding (II). SP2b was the most prevalent science practice present, particularly in content areas related to Big Ideas of Bonding (II), Structure and Function (III), and Visualization (X), as these content areas readily lend themselves to the use of representations and models. SP6a, related to construction of explanations, was frequently associated with Big Ideas I and IV, related to Atoms and Intermolecular Forces, respectively, but was not found in other content areas where constructing explanations could also be of great benefit, such as Structure and Function (III) and Equilibrium (VIII). The same could also be said for other science practices. SP6a was used as an example because the practice of constructing explanations provides opportunity for more explicit understanding of how students' connect pieces of content knowledge, so it is a bit disappointing that it is not more widely distributed across Big Ideas. Nonetheless, the distribution of science practices across the disciplinary core ideas of chemistry is promising, especially when considering that the analysis of incorporation of science practices was conducted post hoc to item construction and release.

The distribution of science practices across content domains led to questions about how performance on items with a science practice compared to items without a science practice but with similar content. Additional analysis was conducted to investigate how items with and without science practices aligned to the same content Big Idea compared on the basis of average item difficulty. Comparison of items with and without science practices found within a particular Big Idea was conducted through the use of independent samples *t*-test. The use of an independent samples *t*-test with unequal variances showed a statistically significant difference in average item difficulty for items

with and without science practices in Big Ideas I, II, IV, and VII, as displayed in Table 8. In the instances of significance, items without a science practice were significantly easier in Big Ideas I, II, and VII. This could suggest that the science practice is indeed being assessed in addition to the chemistry concept in these content domains. Thus, the cognitive complexity of the item increases, making it more difficult for students. In the case of Big Idea IV, items without a science practice were significantly more difficult than items with a science practice. Perhaps this is due to the presence of a model in many of the items with a science practice within this Big Idea providing a scaffold for relating information about molecular structure and intermolecular forces. If this speculation is correct, it suggests that it is possible to incorporate the use of models to a great enough extent in general chemistry that students actually will use them to improve their reasoning about chemistry topics. Thus, calls for reformed chemistry curricula that are focused on depth rather than breadth of content may have a basis to support their argument that depth of content learning does indeed matter.

Science Practices by Item Type (ACR)

In addition to content alignment, a subset of the items had been previously classified based upon whether they were algorithmic (A), conceptual (C), or recall (R) questions (Luxford, et al., 2015). When comparing items with and without science practices, an important distinction can be made based upon item typology as shown in Figure 8. For items without a science practice (N = 283), the number of conceptual items is approximately equal to the number of algorithmic items at 45%, respectively. Whereas items with a science practice (N = 205) were more likely to be deemed conceptual (79%). This difference is important due to the nature of algorithmic and conceptual learning.

Conceptual learning was first described to be different from problem solving by Nurrenbern and Pickering (1987), and has since been further expanded upon in the literature (Nakhleh, 1993; Nakhleh & Mitchell, 1993; Pickering, 1990). Algorithmic problems may be solved by relying on mathematical tools without fully understanding the concepts behind the math, whereas conceptual problems require a deeper understanding of the ideas behind the chemical constructs. Since science practices are a combination of content knowledge with skill, it makes sense that the majority of items found to have a science practice would also be classified as conceptual items that require a higher order of thinking. Items with science practices are intended to demonstrate what a student knows and can do with that knowledge, and as such, are likely to apply to situations beyond the realm of memorization and algorithms. This is not to say that algorithmic or recall items are insufficient as assessment items, rather that the knowledge about student performance and understanding gleaned from these items is less likely to indicate what a student truly understands and can do with that knowledge.

Table 9 displays how science practices are divided across ACR items. The science practices of "Constructing Explanations" and "Engaging in Argument from Evidence" were found only within items considered to be conceptual in design. Many of the other science practices were primarily found in conceptual items, however, "Developing and Using Models" and "Analyzing and Interpreting Data" also showed up in algorithmic items fairly frequently. Of the items containing the practice of "Developing and Using Models," 19% were algorithmic, and 42% were algorithmic for the practice of "Analyzing and Interpreting Data." Students are capable of creating a variety of algorithms for solving problems, so even if an item requires the use of a science practice,

it may only be part of a larger algorithm students have created for solving the problem. This relates back to the discussion of implicit and explicit presence of science practices that occurred earlier in this chapter. Items that contained a science practice and were classified as recall items were infrequent, and constituted less than 5% of the items with a science practice.

Science Practices in Trial Tests

Due to the nature of ACS exam development, and the standard of quality released exams represent, not all items that are developed make it onto a released exam. Trial exams are used to test items developed by the exam committee in order to determine how they will perform with students. Items with very low difficulty index values (too hard) or very high difficulty index values (too easy) are often omitted from the final version of the exam, because items with difficulty ranges in the mid-range are most likely to spread the distribution of student scores over the entire exam, and thereby make the scores students achieve on these norm-referenced exams more discerning. For the purpose of understanding how science practices are incorporated into ACS exam items, these omitted items provide a unique dataset for analysis. Questions as to whether items that contained a science practice were more likely to be omitted from the released exam form, or how the difficulty indices of items with a science practice that were not incorporated into the released exam compares to those of released science practice items, were investigated. The trial test dataset was weighted more heavily toward GCC trial items than GC items. This stems from the fact that in the development of some forms of trial exams, and GC exams often fit into this category, the trial forms are designed to have fairly similar content but minor differences in the construct or specific chemical examples

incorporated. As a result, items from the trial tests that are not on released exams are nonetheless rather similar to those that are on the released exam. Thus, the dataset is weighted toward GCC items because it would have been redundant to review GC trial items that are so similar to the released items already reviewed. Only items that were not incorporated into a released exam from the trial exams were analyzed, because successful trial items that made it onto the released exams were already analyzed in prior phases of the study.

A total of 401 trial items were analyzed, of which 182 items (45.4%) contained at least one science practice. Twenty-three (12.6%) of the 182 items with a science practice contained two science practices, but none of the items contained more than two science practices. The trial items have limited data about them available. Particularly, items from a number of years ago are prone to having limited student performance data available in the Exams Institute archives. Thus analysis was frequently limited to 3D-LAP classification only. More information about the data available from each exam, including trial exams, is shown in Table 5. A comparison of the trial exams analyzed and how they compared to the released exams in terms of incorporation of science practices is shown in Table 10. The percentage of released and unreleased items with a science practice is approximately the same for each exam analyzed, and there does not appear to be a pattern with one type of item being more likely to contain a science practice.

Science practices were found within trial test items with approximately the same distribution as science practices within released items. Figure 9 shows the distribution of incorporation of science practices across trial exams, with the first and second term GCC 2015 trial exams combined for clarity. The distribution of science practices in unreleased

trial items is approximately the same as the distribution of science practices within released items, suggesting that there should not be concern about items being omitted based on the presence of a specific science practice. The unreleased items from the GCC items contain a greater percentage of science practices than the GC trial items. As time progresses, the percentage of GCC trial exam items containing a science practice diminishes. This is consistent with the findings from the analysis of the released exam items and the types of exams involved in the analysis. SP3d and SP8a did not occur at all within the unreleased exam items, but did manage to appear, albeit infrequently, on released exam items. In this regard, the lack of certain science practices in trial exam items is inconsequential since those science practices are being measured within released exam items.

Since the trial exam items analyzed were not released, the question arose as to whether the items with a science practice were omitted because they were more difficult than items without a science practice. Of the 401 trial items analyzed, item difficulty indices were available for only 126 of the items, 45 with a science practice and 81 without. An independent samples *t*-test was conducted on these items, and no significant difference was found between items with a science practice ($M = 0.438$, $SD = 0.201$) and without a science practice ($M = 0.450$, $SD = 0.198$) in terms of average item difficulty, $t(124) = 0.3229$, $p > 0.05$. In this sense, it is not the incorporation of a science practice that makes an item too difficult to be included on the released form of the exam. Speculation suggests that omission was due to content intricacies or item complexity, rather than presence of a science practice. The data available did not allow for statistical analyses between specific science practices individually. Regardless of the reason these items

were not released, cause should not be associated with the presence of a science practice, and future assessment designers need not worry about negative performance arising exclusively from the inclusion of science practices in content measurements.

Science Practices in the Online Laboratory Exam

The design and content of the laboratory exam is very different from the content of the other exams analyzed, so its analysis is presented separately from the other exams. The laboratory exam is conducted electronically via a secure online platform. This allows for the design of the exam items to include pictures, videos, and other interactive features, but items are still selected response rather than free-response. General chemistry content is embedded within the context of laboratory scenarios. The exam does not have item numbers and often questions may have multiple parts or tasks. The raters determined that they had reviewed 42 unique questions. Of the 42 items analyzed, 36 (86%) contained a science practice, and one of the 36 items contained two science practices. The science practice occurred explicitly in 19 (52.8%) of the 36 items, while the remaining 17 (47.2%) items had the science practice present implicitly. Of the 19 explicit occurrences of science practices, 11 were the science practice "3d" (SP3d) that relates to understanding of appropriate use of scientific equipment and techniques. This practice was added to the 3D-LAP in this project in order to encompass an additional component of the practice of planning and carrying out investigations. The criteria for SP3d make it such that if the practice is present, it is present explicitly. Since items were embedded within laboratory scenarios, all but one item contained a phenomenon.

A distribution of science practices found in the laboratory exam can be seen in Figure 10. The science practice of planning and carrying out investigations (SP3) is most prevalent on this exam, which aligns with the objective of the exam to assess laboratory skills and content. More specifically, SP3c and SP3d, related to prediction of experimental observations and demonstrated knowledge of scientific techniques, are incorporated on more than 20% and 25% of the exam items, respectively. Interestingly, the LAB exam contains practices that were incorporated on the other more traditional exams infrequently, and omits practices that were common on the other exams. For example, the LAB exam contains a much larger percentage of items incorporating science practices relating to planning and carrying out investigations (SP3) and analyzing and interpreting data (SP4), but does not include the practices of developing and using models (SP2) or obtaining, evaluating, and communicating information (SP8). These differences in incorporation of science practices can likely be attributed to differences in the types of content assessed by the exams. The differences are important to note, however, because they highlight how laboratory content provides opportunities for incorporation of science practices relative to traditional content. Overall, the unique design and content of the laboratory exam allowed for incorporation of science practices that were not frequently observed in the other exams analyzed.

Conclusions

As assessment efforts shift to meet the demands of reformed curriculum, it is important to consider how the current incorporation of science practices, particularly within standardized chemistry assessments, can be a stepping stone to understand best practices for the inclusion of science practices within multiple-choice items. Future

assessment designers should take confidence in the current status of incorporation of science practices, but be aware of how this distribution is skewed. Currently, the science practice of Developing and Using Models (SP2) is most prevalent within the standardized chemistry exams analyzed. While this is not necessarily problematic, especially given the content domain, consideration of inclusion of the other science practices more frequently may be useful as assessment designers look for options to measure beyond content knowledge retention.

Additional results suggest that the presence of science practices within the multiple-choice test items analyzed had no apparent effect on the average item difficulty index compared to items without a science practice. This observation is of importance because there are undoubtedly concerns about how the addition of science practices may be to the detriment of student performance. Traditionally, multiple-choice item development has been recommended to be done with narrowly defined topics to avoid measurement error (Haladyna, 2012; Haladyna & Downing, 2004). Even though these guidelines are reasonable, from the analysis presented herein it does not appear that multiple-choice items that incorporate science practices are inherently more prone to measurement error because of their complexity. Items that were deemed conceptual in nature were more likely to contain science practices than algorithmic or recall items; useful information for selecting the most viable item type for use with science practices. All ten Big Ideas of the ACCM contained items with science practices. Even though their distribution throughout the content was slightly skewed, the science practices appeared to be relevant across all ten content domains. Science practices were also found across all of the exams analyzed, but particularly on the DUCK, GCC, and LAB exams. The content

and design of these exams made them ideal for the types of items that are most conducive to science practices. Further observations add support to the idea of science practices as viable components of large-scale, multiple-choice assessment.

Limitations of This Study

While the findings suggest that incorporation of science practices within large-scale chemistry multiple-choice assessments is entirely possible without unavoidable detriment to test takers, the limitations of the findings and the study should also be acknowledged. The study was limited to standardized multiple-choice exams in the domain of chemistry, particularly general chemistry, and as such it would be inappropriate to generalize the results to other item types (e.g. free response) or other disciplines without further investigation. Additionally, because the analysis was conducted on previously released exam items across a 20 year time period, some performance data were not readily available in ACS EI archives. The committee structure of ACS exam development results in high quality test items, even if they do not make it to the released exam. Therefore, if the 3D-LAP rubric were to be used to analyze items with less rigorous development processes, different results may be observed. The conclusions from this study relate primarily to the incorporation of science practices, with some implications for how disciplinary core ideas intertwine with science practices. Crosscutting concepts were not within the realm of this study. Future work within the field to provide empirical evidence for how crosscutting concepts are to be incorporated and evaluated within the chemistry curriculum will likely remedy this current limitation.

Implications and Future Research

The study provides evidence for the presence of science practices within ACS general chemistry exams and across key content domains of chemistry. Since these exams were not intentionally designed to incorporate science practices into content assessments, one implication of this analysis is that these practices are integral to the understanding of chemistry and a valuable component of assessment. Additionally, the fact that these assessments are multiple-choice and reach a large user group within the chemistry community suggests that the use of science practices to assess content knowledge is feasible within the constraints of large-scale assessment designs.

Further research is needed on the design and performance of items that contain were explicitly constructed to measure science practices. The current efforts herein support the idea that science practices can be readily incorporated into multiple-choice assessment items, but additional studies are needed to corroborate and validate evidence that the measurement corresponds to student skill development beyond mastery of algorithms. Additionally, future studies should consider how to garner evidence to support the measure of science practices independent of content. Further research is also needed on the efficacy of incorporating multiple science practices within a single item since the current project did not contain enough items to make strong conclusions about how multiple practices impact item performance.

Overall, the reform efforts in chemistry curriculum necessitate reevaluation of assessment construction. Successful evaluation of large-scale reforms, such as the NGSS, requires robust assessments capable of intertwining multiple dimensions of learning.

Tools, such as the 3D-LAP rubric, are being developed to aid assessment designers challenged with the task of measuring student learning beyond the traditional realm of content knowledge. The use of the 3D-LAP to gauge the status of three-dimensional learning within ACS examinations provides information fundamental to the chemistry community and assessment reform efforts.

References

Achieve. (2013). Next generation science standards. Washington, DC: National Academies Press.

Alagumalai, S., & Curtis, D. D. (2005). *Classical test theory*: Springer.

American Association for the Advancement of Science. (1989). *Science for all Americans: A Project 2061 report on literacy goals in science, mathematics, and technology* (Vol. 1): AAAS.

American Chemical Society. (2005). *Chemistry*. New York, NY: W.H. Freeman.

Archbald, D. A., & Newmann, F. M. (1988). Beyond Standardized Testing: Assessing Authentic Academic Achievement in the Secondary School.

Association of American Medical Colleges. (2014). What's on the MCAT 2015 Exam? Retrieved March 26, 2015, from https://www.aamc.org/students/download/377882/data/mcat2015-content.pdf

Bodner, G., MacIsaac, D., & White, S. (1999). Action Research: Overcoming the sports mentality approach to assessment/evaluation. *University Chemistry, 3*(1).

Brandriet, A. R., & Holme, T. A. (2015). Unpublished work.

Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72.

Chudowsky, N., & Pellegrino, J. W. (2003). Large-scale assessments that support learning: what will it take? *Theory into practice, 42*(1), 75-83.

College Board. (2011a). The AP biology curriculum framework. New York: The College Board.

College Board. (2011b). The AP chemistry curriculum framework. New York: The College Board.

College Board. (2014). The AP physics curriculum framework. New York: The College Board.

Cooper, M. M. (2014). Personal Communication.

Cooper, M. M., & Klymkowsky, M. (2013). Chemistry, life, the universe, and everything: a new approach to general chemistry, and a model for curriculum reform. *Journal of Chemical Education, 90*(9), 1116-1122.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*(2), 020103.

Emenike, M., Raker, J. R., & Holme, T. (2013). Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *Journal of Chemical Education, 90*(9), 1130-1136.

Emenike, M., Schroeder, J., Murphy, K., & Holme, T. (2013). Results from a national needs assessment survey: A view of assessment efforts within chemistry departments. *Journal of Chemical Education, 90*(5), 561-567.

Haladyna, T. M. (2012). *Developing and validating multiple-choice test items*: Routledge.

Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23*(1), 17-27.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education, 15*(3), 309-333.

Holme, T. (2003). Assessment and quality control in chemistry education. *Journal of Chemical Education, 80*(6), 594.

Holme, T. (2011). Assessment Data and Decision Making in Teaching. *Journal of Chemical Education, 88*(8), 1017-1017.

Holme, T., Bretz, S. L., Cooper, M., Lewis, J., Paek, P., Pienta, N., et al. (2010). Enhancing the role of assessment in curriculum reform in chemistry. *Chemistry Education Research and Practice, 11*(2), 92-97.

Holme, T., & Murphy, K. (2012). The ACS Exams Institute undergraduate chemistry anchoring concepts content map I: General Chemistry. *Journal of chemical education, 89*(6), 721-723.

Kirch, D. G., Mitchell, K., & Ast, C. (2013). The new 2015 MCAT: testing competencies. *JAMA, 310*(21), 2243-2244.

Koretz, D. M. (2002). Limitations in the use of achievement tests as measures of educators' productivity. *Journal of human resources*, 752-777.

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment, 7*(1), 29-38.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-Choice Tests Exonerated, at Least of Some Charges Fostering Test-Induced Learning and Avoiding Test-Induced Forgetting. *Psychological Science, 23*(11), 1337-1344.

Lloyd, B. W. (1992). A review of curricular changes in the general chemistry course during the twentieth century. *Journal of Chemical Education, 69*(8), 633.

Luxford, C. J., & Holme, T. A. (2015). What Do Conceptual Holes in Assessment Say about the Topics We Teach in General Chemistry? *Journal of Chemical Education*. DOI: 10.1021/ed500889j

Luxford, C. J., Linenberger, K. J., Raker, J. R., Baluyut, J. Y., Reed, J. J., De Silva, C., et al. (2015). Building a Database for the Historical Analysis of the General Chemistry Curriculum Using ACS General Chemistry Exams as Artifacts. *Journal of Chemical Education*, *92*(2), 230-236.

Maeyer, J., & Talanquer, V. (2010). The role of intuitive heuristics in students' thinking: Ranking chemical substances. *Science Education, 94*(6), 963-984.

McClary, L., & Talanquer, V. (2011). Heuristic reasoning in chemistry: Making decisions about acid strength. *International Journal of Science Education, 33*(10), 1433-1454.

Murphy, K., Holme, T., Zenisky, A., Caruthers, H., & Knaus, K. (2012). Building the ACS Exams anchoring concept content map for undergraduate chemistry. *Journal of Chemical Education, 89*(6), 715-720.

Murphy, K., Picione, J., & Holme, T. A. (2010). Data-Driven Implementation and Adaptation of New Teaching Methodologies. *Journal of College Science Teaching*, *40*(2), 80-86.

Nakhleh, M. B. (1993). Are our students conceptual thinkers or algorithmic problem solvers? Identifying conceptual students in general chemistry. *Journal of Chemical Education, 70*(1), 52.

Nakhleh, M. B., & Mitchell, R. C. (1993). Concept learning versus problem solving: There is a difference. *Journal of Chemical Education, 70*(3), 190.

National Research Council. (1996). *National science education standards*. Washtington, DC: National Academies Press.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

Nurrenbern, S. C., & Pickering, M. (1987). Concept learning versus problem solving: Is there a difference? *Journal of chemical Education, 64*(6), 508.

Pellegrino, J. W. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831-841.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65-77.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*: National Academies Press.

Pickering, M. (1990). Further studies on concept learning versus problem solving. *Journal of Chemical Education, 67*(3), 254.

Raker, J. R., Emenike, M., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education, 90*(8), 981-987.

Sacks, P. (2000). *Standardized minds: The high price of America's testing culture and what we can do to change it*: Da Capo Press.

StataCorp. (2013). Stata Statistical Software: Release 13. College Station, TX: StataCorp, LP.

Talanquer, V. (2006). Commonsense chemistry: a model for understanding students' alternative conceptions. *Journal of Chemical Education, 83*(5), 811.

Talanquer, V. (2014). Chemistry education: Ten heuristics to tame. *Journal of Chemical Education, 91*(8), 1091-1097.

Talanquer, V., & Pollard, J. (2010). Let's teach how we think instead of what we know. *Chemistry Education Research and Practice, 11*(2), 74-83.

Towns, M. H. (2009). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education, 87*(1), 91-96.

Underwood, S. M., Cooper, M. M., Krajcik, J., Cabellero, D., & Ebert-May, D. (2014). *Designing a rubric to characterize assessments*. Paper presented at the 248th National Meeting of the American Chemical Society.

Table 1. Crosscutting concepts associated with the NGSS.

Crosscutting Concepts
1. Patterns
2. Cause and effect: Mechanism and explanation
3. Scale, proportion, and quantity
4. Systems and system models
5. Energy and matter: Flows, cycles, and conservation
6. Structure and function
7. Stability and change

Table 2. The disciplinary core ideas in the physical sciences as outlined by the NGSS.

Disciplinary Core Ideas in the Physical Sciences
1. Matter and Its Interactions
2. Motion and Stability: Forces and Interactions
3. Energy
4. Waves and Their Applications in Technologies for Information Transfer

Table 3. Science practices associated with the NGSS.

Practices for K-12 Science Education
1. Asking questions (for science) and defining problems (for engineering)
2. Developing and using models
3. Planning and carrying out investigations
4. Analyzing and interpreting data
5. Using mathematics and computational thinking
6. Constructing explanations (for science) and designing solutions (for engineering)
7. Engaging in argument from evidence
8. Obtaining, evaluating, and communicating information

Table 4. A description of the ACS exams used in the analysis.

Examination	Exam ID	Released Exams Analyzed (Year)	Number of Items per Exam	Content
General Chemistry (Full Year)	GC	1995, 2001, 2013	70*	Associated with a year-long general chemistry course.
General Chemistry First-Term	GCF	2012	70	Associated with the first term of a general chemistry sequence.
General Chemistry Conceptual	GCC	1996, 2001, 2008	60	General chemistry content associated with a year-long course assessed in a conceptual manner.
Paired Questions First-Term	PQF	2005	40	Pairs of questions, conceptual and traditional (algorithmic), associated with the first semester of general chemistry.
Paired Questions Second-Term	PQS	2007	40	Pairs of questions, conceptual and traditional (algorithmic), associated with the second semester of general chemistry.

Table 4. *(continued)*

Examination	Exam ID	Released Exams Analyzed (Year)	Number of Items per Exam	Content
Laboratory (Online)	LAB	2013	Approx. 40	Content associated with general chemistry laboratory experiments, equipment, and procedures. Conducted via an online computer interface.
Chemistry in Context	CIC	2009	90	Content associated with chemistry as it relates to real-world applications and contexts.
Diagnostic of Undergraduate Chemistry Knowledge	DUCK	2008	60	Scenarios to assess content across the undergraduate chemistry curriculum.

* The 1995 GC exam contained 75 items.

Table 5. ACS exams listed by the type of data available.

3D-LAP Classification Only	3D-LAP and Difficulty/Discrimination	3D-LAP, Difficulty/Discrimination, and ACCM
LAB (2013)	PQF (2005)	GCC (1996, 2001, 2008)
CIC (2009)	PQS (2007)	GC (1995, 2003, 2013)
GCC (1996 & 2015 trial tests)	GCC (2001 & 2008 trial tests)	GCF (2012)
	GC (2013 trial tests)	
	DUCK (2008)	

Table 6. Distribution of Science Practices in ACS Exam Items (*N* = 695).

Science Practice		Number of Occurrences
1. Asking Questions		**0**
2. Developing and Using Models		**176**
	2b.	176
3. Planning and Carrying Out Investigations		**19**
	3a.	4
	3c.	11
	3d.	4
4. Analyzing and Interpreting Data		**21**
	4a.	6
	4b.	15
5. Mathematical and Computational Thinking		**26**
	5a.	2
	5b.	24
6. Constructing Explanations		**35**
	6a.	35
7. Engaging in Argument from Evidence		**23**
	7a.	23
8. Obtaining, Evaluating, and Communicating Information		**23**
	8a.	2
	8b.	21
TOTAL		**323**

Table 7. Distribution of items with science practices compared to total number of items within an ACCM Big Idea.

ACCM Big Idea	Number of SP Items	Total Number of Items	Items with SP Relative to Big Idea Total Items (%)
I. Atoms	17	74	22.97
II. Bonding	10	18	55.56
III. Structure and Function	22	37	59.46
IV. Intermolecular Forces	40	84	47.62
V. Reactions	20	73	27.40
VI. Thermodynamics	24	66	36.36
VII. Kinetics	15	33	45.45
VIII. Equilibrium	22	69	31.88
IX. Experimental	14	26	53.85
X. Visualization	7	8	87.50
TOTAL	189	488	--

Table 8. Statistical results of comparison between items with and without science practices aligned to the same content Big Idea on the ACCM.

Big Idea	No. of Items with SP	Item Difficulty with SP Mean (*SD*)	Item Difficulty without SP Mean (*SD*)	*t*-value	Cohen's *d*
I. Atoms (*N*=74)	17	0.487 (0.180)	0.594 (0.161)	2.1953*	0.644***
II. Bonding (*N*=18)	10	0.564 (0.135)	0.672 (0.102)	1.9341*	0.888****
III. Structure and Function (*N*=37)	27	0.578 (0.170)	0.607 (0.154)	0.5375	--
IV. Intermolecular Forces (*N*=84)	43	0.609 (0.175)	0.521 (0.151)	-2.4709**	-0.537***
V. Chemical Reactions (*N*=73)	20	0.551 (0.182)	0.540 (0.165)	-0.3257	--
VI. Thermodynamics (*N*=66)	24	0.485 (0.137)	0.551 (0.183)	1.6679	--
VII. Kinetics (*N*=33)	17	0.465 (0.123)	0.595 (0.137)	2.8530**	0.997****
VIII. Equilibrium (*N*=69)	26	0.516 (0.170)	0.488 (0.159)	-0.6705	--
IX. Experiments (*N*=26)	14	0.616 (0.124)	0.603 (0.235)	-0.1715	--
X. Visualization (*N*=8)	7	0.481 (0.180)	0.700 (n/a)	--	--

*Denotes significance at $\alpha = 0.05$ level.
**Denotes significance at $\alpha = 0.01$ level.
***Denotes moderate effect size.
****Denotes large effect size.

Table 9. Science practices across ACS exam items classified as Algorithmic, Conceptual, and Recall.

Science Practice	Algorithmic	Conceptual	Recall	Total
2. Developing and Using Models	21	82	7	110
3. Planning and Carrying Out Investigations	1	11	1	13
4. Analyzing and Interpreting Data	5	7	0	12
5. Using Mathematics and Computational Thinking	3	12	1	16
6. Constructing Explanations	0	21	0	21
7. Engaging in Argument from Evidence	0	17	0	17
8. Obtaining, Evaluating, and Communicating Information	3	13	0	16
TOTAL	33	163	9	205

Table 10. Information about each of the trial examinations analyzed, including comparison of released and not released items with science practices.

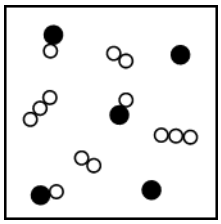
Trial Exam Name	Number of Items Analyzed	Unreleased Items with a Science Practice (%)	Released Items with a Science Practice (%)	Item Difficulty Data Available
GCC 1996	51	72.5	73.3	No
GCC 2001	64	68.8	56.7	No
GCC 2008	60	56.7	61.7	Yes
GCC 2015 FT	80	45.0	n/a	No
GCC 2015 ST	80	46.3	n/a	No
GC 2013	66	25.8	21.4	Yes

Mock Item 1.

Two chemicals, A and B_3, are placed in a container and react according to the equation:

$$A + B_3 \rightleftharpoons AB + B_2.$$

The reaction reaches equilibrium and the chemicals present at equilibrium are diagrammed below.



Which graph best describes the reaction progress of the above reaction?

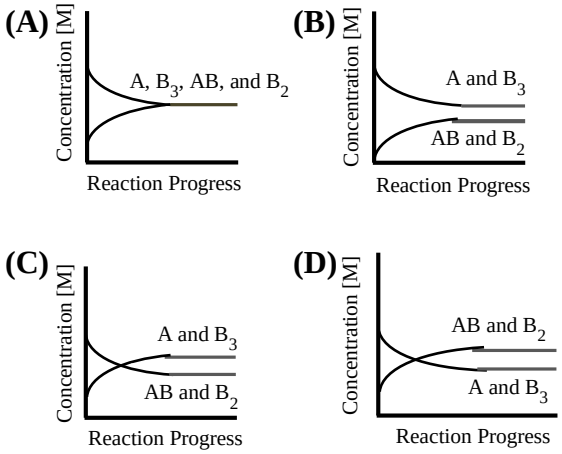


Figure 1. Mock ACS exam item with science practices.

Mock Item 2.

A balloon is filled with 2.0 liters of gas at 22°C. If the temperature increases from 22°C to 30°C, what would the new volume of the balloon be at constant pressure?

(A) 2.7 L

(B) 0.37 L

(C) 2.1 L

(D) 3.0 L

Figure 2. Mock ACS exam item without a science practice.

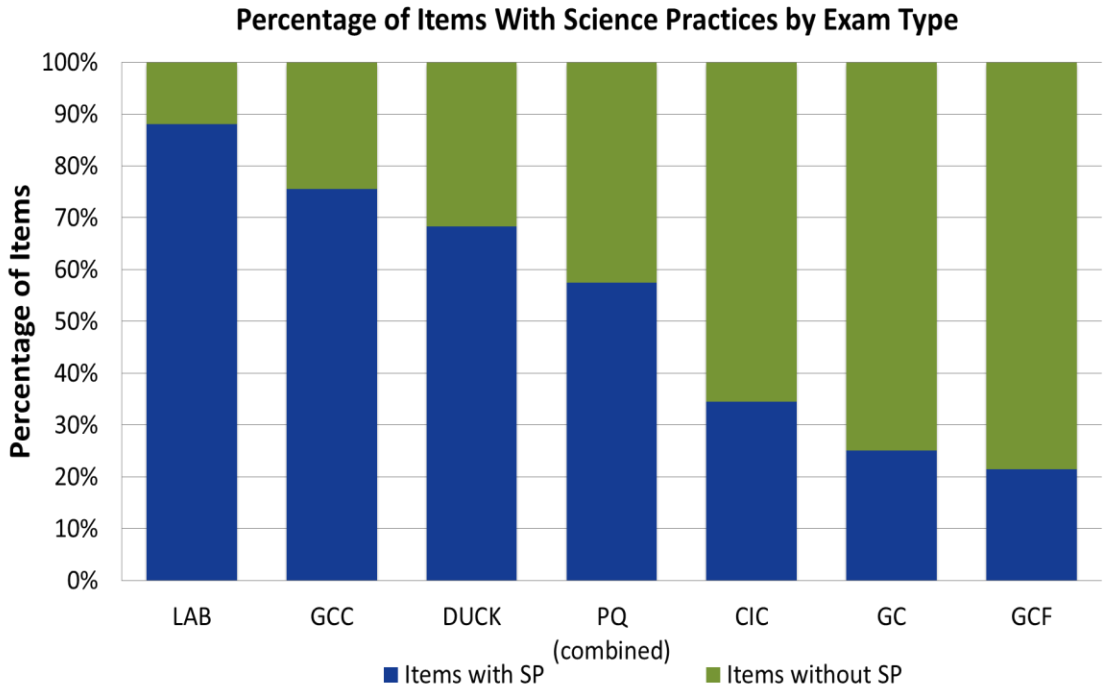


Figure 3. Percentage of items with and without science practices by type of ACS exam analyzed.

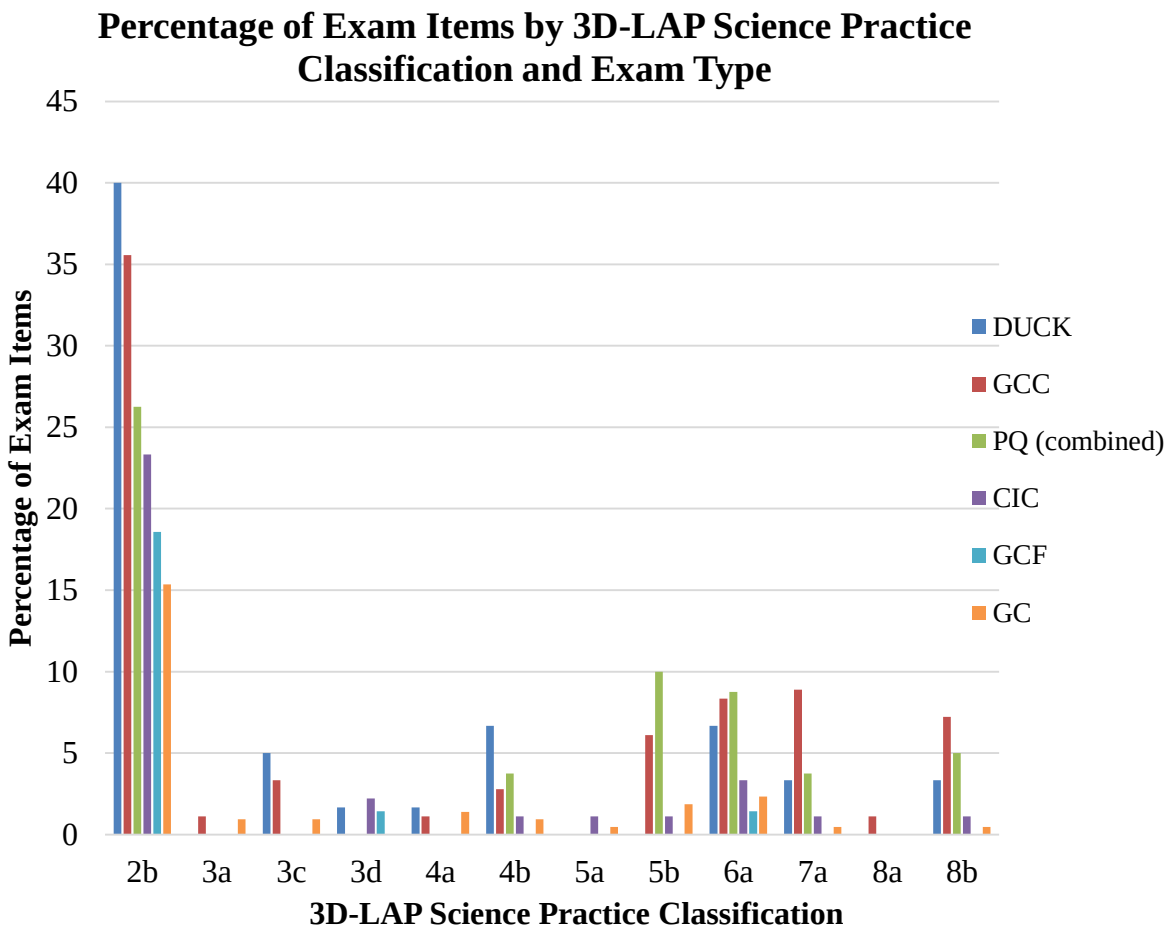


Figure 4. Percentage of items containing a science practice by 3D-LAP rubric classification and exam type.

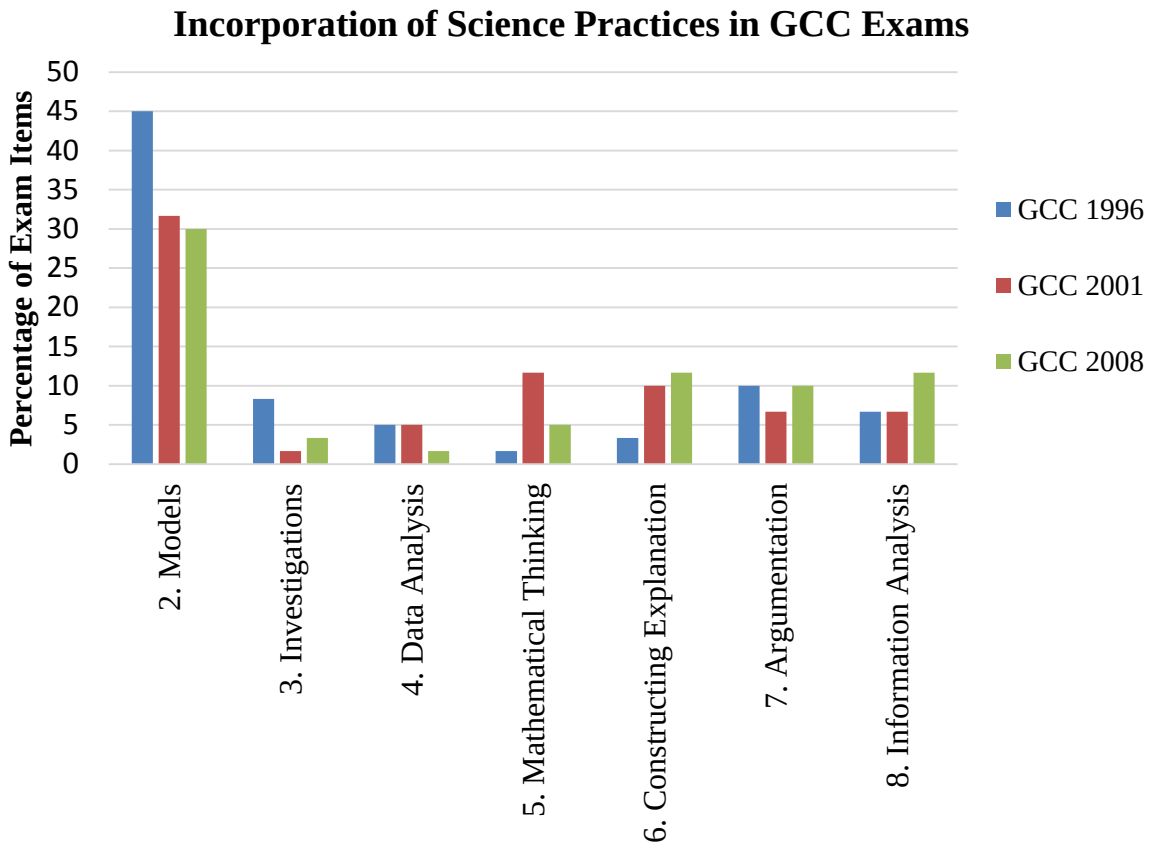


Figure 5. Analysis of the incorporation of science practices in GCC exams over time.

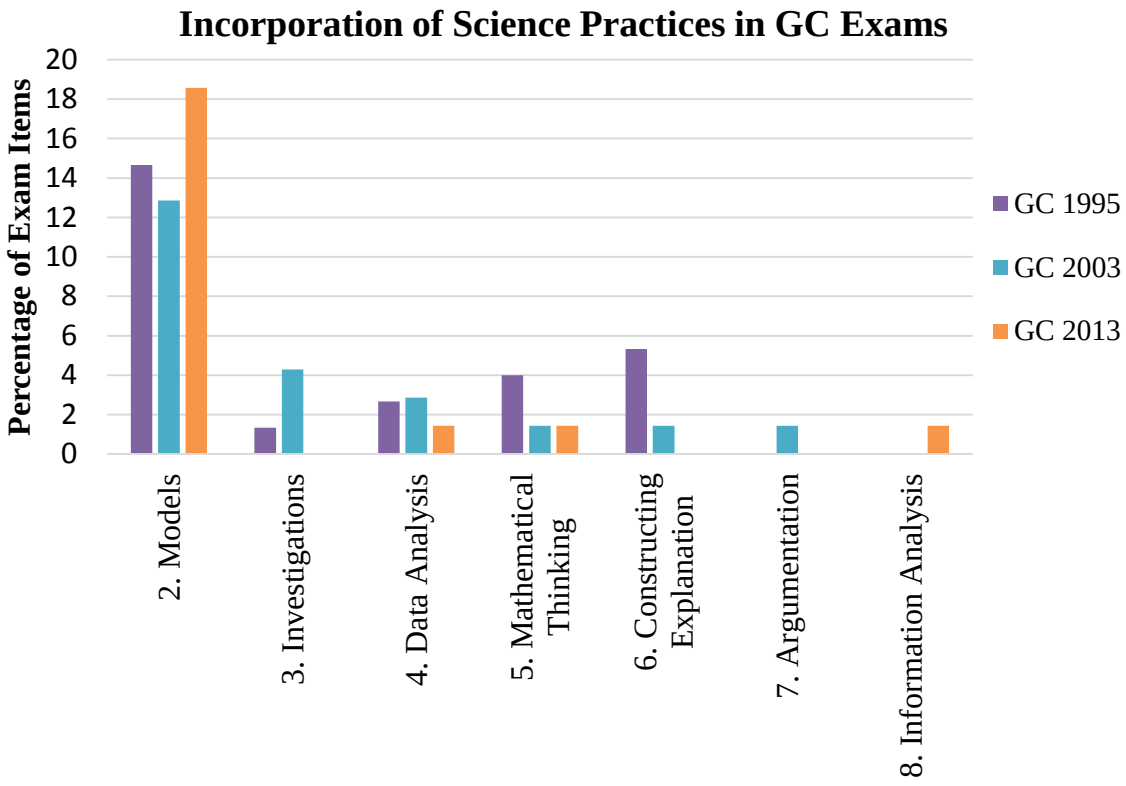


Figure 6. Analysis of incorporation of science practices within GC exams over time.

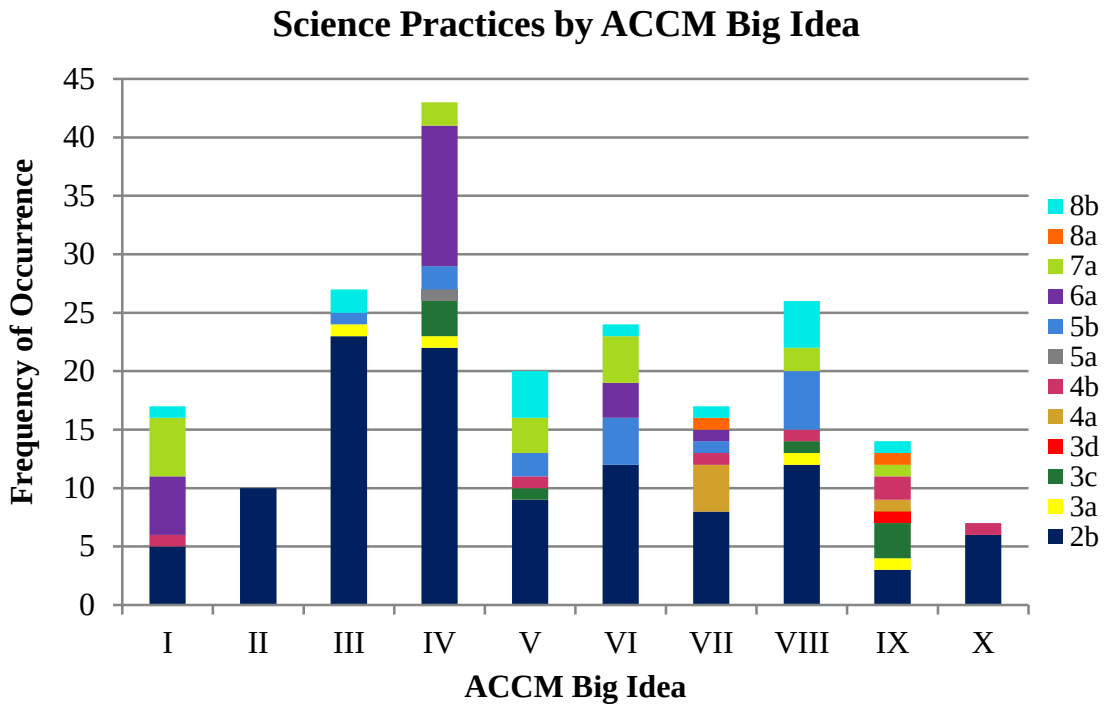


Figure 7. Distribution of items with science practices across the ten Big Ideas of the ACCM by 3D-LAP classification.

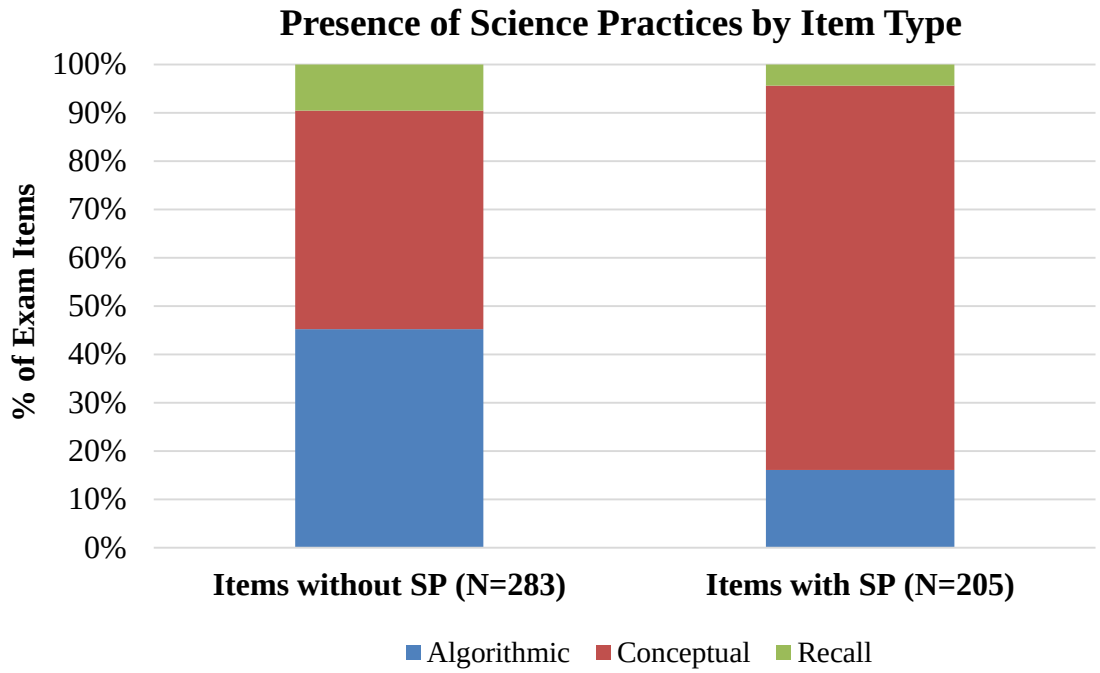


Figure 8. Percentage of items with and without science practices classified as algorithmic, conceptual, or recall.

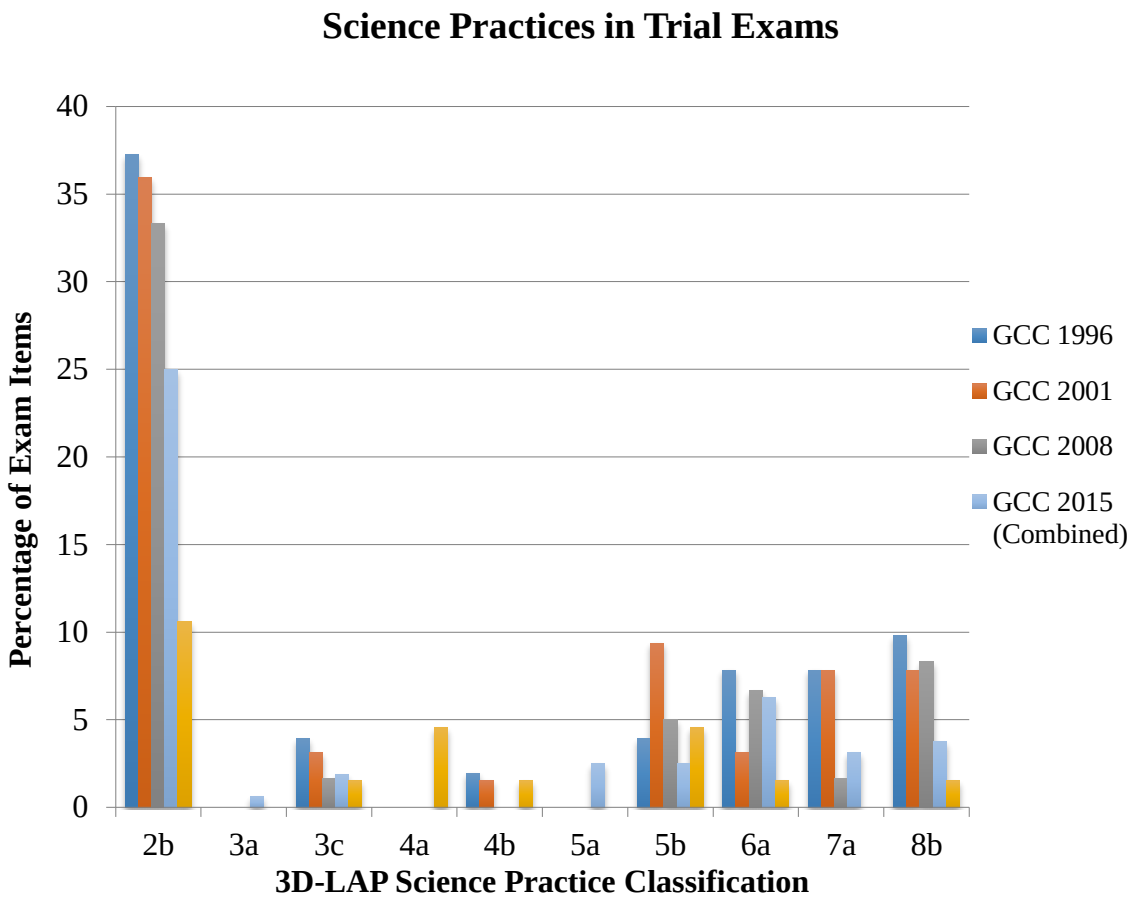


Figure 9. Distribution of science practices in trial items that were not released by 3D-LAP classification.

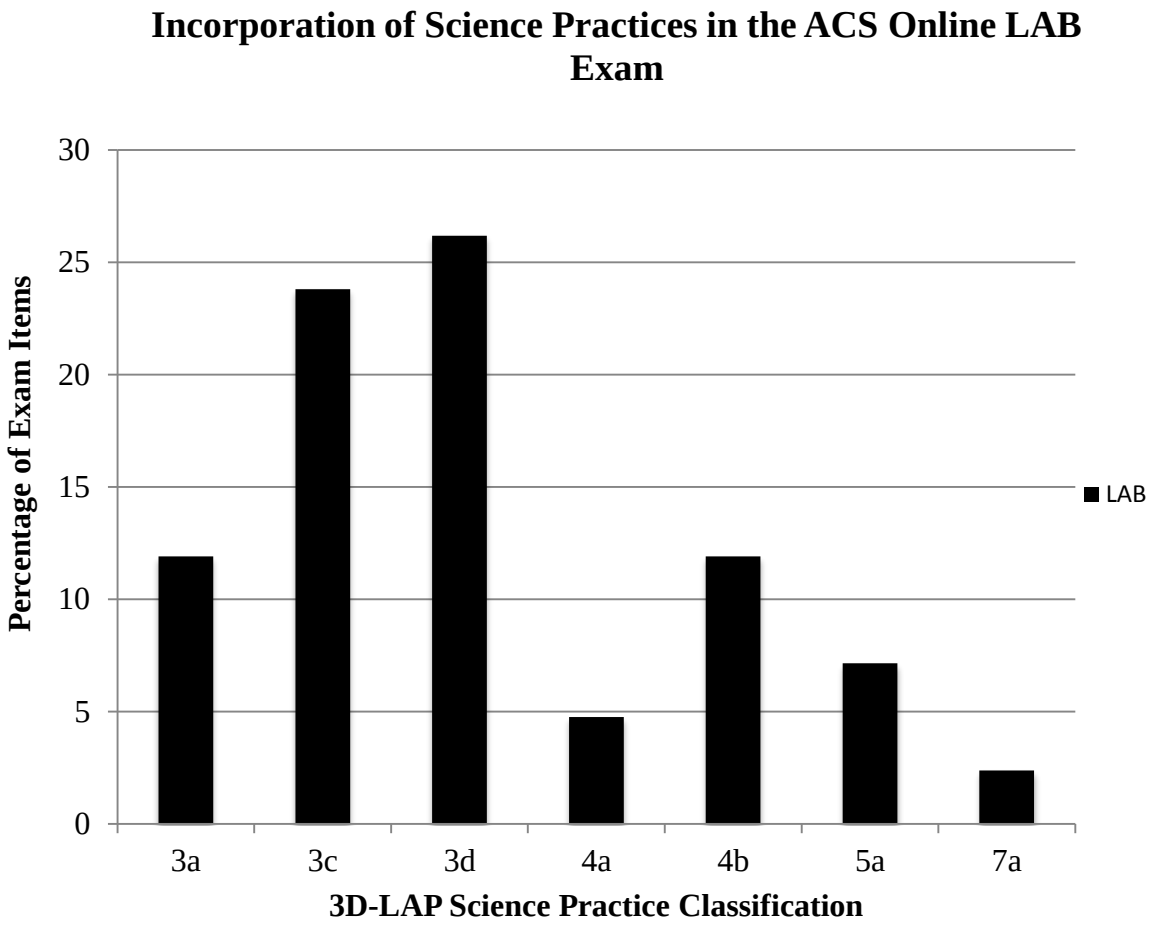


Figure 10. Percentage of items on the LAB exam containing a science practice by 3D-LAP classification.

APPENDIX

*Changes made by the Iowa State University research team are in blue text.
Three Dimensional Learning Assessment Protocol (3D-LAP)

Three-Dimensional Learning

For the purposes of this document, we define "Three-Dimensional Learning" to mean the blending of Scientific Practices, Crosscutting Concepts, and Disciplinary Core Ideas (as defined by *A Framework for K-12 Science Education* and the *Next Generation Science Standards*).

The 3D-LAP

The Three-Dimensional Learning Assessment Protocol is being designed for two purposes: (1) to characterize the extent to which formative and summative assessments are aligned with three-dimensional learning and (2) to guide the redesign of current assessment questions to provide explicit evidence of student understanding.

Part 1 - Characterizing assessments

1. Format of the question: Is the question multiple-choice or free-response?
2. Scientific practices (SP)
    a. Does the item contain a scientific practice?
    b. If there is a practice, which practice is assessed?
    c. If there is a practice, is the practice explicit/implicit?
3. Crosscutting concepts (CC)
    a. Does the item contain a crosscutting concept?
    b. If there is a crosscutting concept, which crosscutting concept is assessed?
    c. If there is a crosscutting concept, is the crosscutting concept explicit/implicit?
4. Disciplinary core idea (DCI)
    a. Does the item contain a disciplinary core idea?
    b. If there is a disciplinary core idea, which disciplinary core idea is assessed?
    c. If there is a disciplinary core idea, is the disciplinary core idea explicit/implicit?
5. Phenomenon: Is the question situated in an observation, event, or phenomenon?

Part 2 - Rewriting assessment questions

1. Explicit evidence: Does the item elicit explicit evidence of student learning that is aligned with the intent of the question?
2. Intent
    a. What is your interpretation for the intent or goal of the question?
    b. Do you think the intent of the question is - clear, mostly clear, or unclear?
3. Can students answer this question by relying on heuristics - yes/no?

4. For the author - what was your intent for this question?
5. Learning goal: Does the question address an explicit learning goal?
6. Question construction: Does the question meet acceptable practices for valid item construction (e.g. appropriate level of math and reading literacy, reasonable number of choices of similar length)?
7. Recommendation for question: What is your recommendation for this item - discard and try again, asks major revision, asks minor revision, revision with addition of more questions, use as is with additional questions, use as is?

Operationalization of the 3D-LAP

This protocol takes individual questions to be the unit of analysis. If a question has sub-parts (such as 3a, 3b, 3c), each of those should be analyzed separately (in the example, treated as 3 questions) but it should be noted that they are part of a series.

Definition of a Phenomenon: *A question is considered to represent a phenomenon if it contains an event, experiment, or data observable at the macroscopic scale. Additionally, if data are given, they should represent something a student would likely be able to collect in a laboratory experiment to answer the question (e.g. masses and temperature changes in a calorimetry experiment would be considered part of a phenomenon whereas freezing point data could be obtained from a table or handbook of such values, so analysis of these data would not constitute a phenomenon in and of itself.)*

Operationalization of the Scientific Practices:

1. Asking Questions
    a. Question asks student to propose a scientific question about an event, observation, or phenomenon.
        i. Question gives an event, observation, or phenomenon
        ii. Question asks student to propose a question that can be answered by using the other scientific practices
2. Developing and Using Models
    a. Question asks student to construct a (mathematical, graphical, diagrammatic) representation and use it to explain or predict an event, observation, or phenomenon.
        i. Question gives an event, observation, or phenomenon for the student to explain or predict
        ii. Question asks student to construct a representation
        iii. Question asks student to generate an explanation or make a prediction based on the representation
    b. Question gives a representation and the student is asked to use it to explain or predict an event, observation, or phenomenon. Interpretation of the representation must be the only way to answer the question.
        i. Question gives a representation
        ii. Question asks student to use the given representation to explain or predict an event, observation, or phenomenon

iii. *Question asks student to provide the reasoning link between the representation and their explanation and/or prediction*

c. Question gives a representation of a model and the student is asked to evaluate a model
   i. Question gives a representation
   ii. Question gives a model that uses the given representation
   iii. Question asks student to identify the merits or limitations of the model
   iv. Question asks student to provide justifications for those merits/limitations

3. Planning and Carrying Out Investigations
   a. Question asks student to describe the procedure and experimental conditions they would use to answer a question, refute or support a claim/hypothesis, or solve a problem, with explanation of why each method and condition is used.
      i. Question gives a question, claim/hypothesis, or a problem
      ii. Question asks student to determine what measurements need to be made to answer a question, refute or support a claim/hypothesis, or solve a problem
      iii. Question asks student to determine equipment necessary to run experiment
      iv. Question asks student to explain why each measurement is important to answering a question, refuting or supporting a claim/hypothesis, or solving a problem
      v. Question asks student to explain why each apparatus is important to make the necessary measurements
   b. Question asks student to execute an experiment and report on the resulting measurements or observations
      i. Question gives student an experiment to run or an observation to make
      ii. Question asks student to report the measurements or observations that were made
   c. Question asks student to predict measurements or observations for a given experiment/observation
      i. Question describes an experiment
      ii. Question asks student to predict the resulting measurements or observations of the experiment
      iii. Question asks student to explain how the measurements/observations relate to each other

d. Question asks student to demonstrate a knowledge of how to use scientific equipment or techniques. This includes the understanding of appropriate equipment or technique to accomplish a specific task or achieve a specified result. (e.g. The student may be provided with a series of images of a balance being used and in order to answer the question they have to understand which measurements are needed for the calculation, etc. or if given a picture of

various pieces of equipment, identify what is necessary to perform a specific task, such as a filtration)

i. Question describes a scientific task, experiment, or observation.

ii. Question asks student to demonstrate knowledge of how to use or identify appropriate equipment and/or technique to achieve the task, experiment, or observation.

4. Analyzing and Interpreting Data

**Data are to be represented numerically or graphically. Observations are not considered data unless they are tabulated and include a numerical value (e.g. classifying a substance based on its properties would not be considered an analysis of data unless the student had to use/interpret/analyze data from a chart, table, or graph to answer the question).*

***A note about distinguishing 2b (models) from 4b (data analysis): An item containing a graph/table, etc in which the data are not specifically identified (e.g. the axes are not fully numerically specified; the compounds used are not identified, etc) is considered to be more theoretical in nature and is considered as SP "2b." Items that contain graphs/tables, etc in which all data are specifically identified and would be possible for a student to collect in an experiment are considered SP "4b" if the student is asked to interpret the analysis of the data. For example, an item asking for interpretation or prediction using a plot of generic trends of pH changes as water is added to acid would be considered "2b" even though a student could do such a thing in a lab. Whereas an item that presents specific pH readings for the same scenario, identifies the acid and volumes of water, and asks for an interpretation/prediction would be considered "4b".*

a. Question gives data to the student and asks student to analyze that data
   i. Question asks student to select or develop an analysis method for the given data
   ii. Question asks student to conduct the analysis on the data *(e.g. the student has to organize the data, conduct any mathematical, graphical, or statistical tests on the data, and present a final product of analysis (table, numerical result, graph, etc.)*

b. Question gives student an analysis of data (e.g. graph) collected from an experiment (not a plot of theoretical values or relationships) and asks student to interpret what it means (e.g. The student is provided with the graph of data collected during a titration and asked to interpret the types of species involved (strong vs. weak acid/base), end point, etc would be an analysis of data, whereas a graph of how potential energy changes as distance changes would be considered to represent a model.)
   i. Question gives an analysis of data (graph/table of numerical values)
   ii. Question asks student to interpret the results of the analysis

5. Using Mathematics and Computational Thinking

a. Question asks student to choose a mathematical tool (a set of mathematical operations), construct a problem or sub-problem for which the math can be used (self-manufacture scaffolding to do the math),

execute the relevant mathematical procedures, and then reflect on the solution (including how they believe their solution is accurate/correct). *The question asks the student to describe why they believe their solution is accurate/correct. Therefore, traditional algorithmic problems that only require a student to execute the math without reflecting or reasoning through the mathematical process and/or answer are not considered to contain this practice.*

i. Question asks student to construct a problem or sub-problem for which a mathematical tool can be used (self-manufacture scaffolding to do the math)
ii. Question asks student to execute the relevant mathematical procedures

b. *Question asks students to use computational reasoning to solve a problem. This means that the student would use reasoning, such as proportional reasoning, to infer a relationship to solve the problem. However, the student may or may not need to perform a mathematical calculation. This could mean that a student would need to interpret a relationship from variables traditionally present in a mathematical equation or use conceptual reasoning about a numeric situation.*

i. Question asks students to use a computation reasoning skill to solve a problem
ii. Question asks students to infer a relationship to solve the problem
iii. Question asks student to describe why they believe their solution is accurate

6. Constructing Explanations
    a. Question asks student to explain a phenomenon, event, or observation (may be hypothetical.)
        i. Question asks student to reference scientific principles and/or data
        ii. Question asks student to provide reasoning linking scientific principles and/or data to phenomenon, event, or observation
7. Engaging in Argument from Evidence
    a. Question asks student to provide evidence and reasoning to support, refute, or critique a claim. **In MC items, the claim is typically embedded in the response choices, so having a claim alone (i) does not make the item implicitly contain this SP. Item must have (i) and (ii) or (iii) to be marked as implicit use of this SP.*
        i. Question gives a claim (In MC, student may be asked to select one choice over another, and then justify their choice).
        ii. Question asks student to provide evidence
        iii. Question asks student to provide reasoning linking evidence to claim

8. Obtaining, Evaluating, and Communicating Information
    a. Question asks student to read/view *information* on scientific topics and describe it in their own words, evaluate its legitimacy, or critique it. *This could include evaluating claims in popular media, scientific journals, or other reputable sources.*
        i. Question gives *information* on a scientific topic (or identifies how to obtain it)
        ii. Question asks student to describe the *information* in their own words, evaluate its reliability, or critique it
    b. Question asks student to translate from a visual or mathematical representation to a written or oral representation. *This could relate to translating between different symbolic representations in chemistry (e.g. Given a chemical formula, identify the correct Lewis structure). *Written representations mean symbolic alpha/numeric representations such as chemical formulas or equations, not paragraphs of words.*
        i. Question gives a visual, mathematical, graphical, or written representation
        ii. Question asks student to translate from the given representation into another representation
        iii. *Question asks student to justify or explain the need for the translation between representation types*

CHAPTER 5: DESIGN AND USE OF ITEMS TO MEASURE SCIENCE PRACTICES IN A GENERAL CHEMISTRY COURSE

Jessica J. Reed and Thomas A. Holme

A paper to be submitted to the Journal of Chemical Education

Abstract

New visions of science education call for reformed curricula and assessments that intertwine science practices with core content and crosscutting concepts. These reforms provide an impetus to examine how post-secondary assessments provide measures of not only what students know, but also what they can do with that knowledge. The research in this chapter focuses on the use of a rubric to create multiple-choice assessment items that explicitly incorporate science practices. The items created were then implemented in course exams in a large-enrollment general chemistry course. Analysis of student performance revealed that performance was lower on items with science practices compared to items without science practices, suggesting that adding science practices may increase the cognitive demand of the item. Thus, course instruction that emphasizes the development and use of science practices may be useful to bridge the gap in performance scores.

Introduction

As has been discussed in previous chapters, new visions of science education call for reformed curricula and assessments that intertwine science practices with core content and crosscutting concepts. The core ideas behind these reforms are found in the report entitled *A Framework for K-12 Science Education: Practices, Crosscutting Concepts and Core Ideas* (National Research Council, 2012) and the Next Generation Science

Standards (NGSS) (Achieve, 2013) derived from it. While these documents are directly aimed at the K-12 science classroom, they have implications at the post-secondary level. First, they provide college-level educators with a means to ground their teaching, second they suggest that future students may enter the college science classroom prepared to engage with content in new ways, and third, they provide an impetus for reexamining how science content is assessed.

Envisioning and redesigning formative and summative assessments to meet the demands of students who have experienced these new standards certainly provides a challenge for the science education community. Just as importantly, assessments that include measures of science practices offer a positive approach to build assessment systems (Pellegrino, 2014). The goals and aims of the NGSS suggest that there is potential to create assessment systems that are continuous in design, meaning that they are able to measure student progress over time (Pellegrino, 2014). Additionally, such systems may reduce the amount of conflict between achievement goals and results that frustrate students and educators alike. Yet, creating assessment materials to align with the NGSS may pose a challenge until more empirical research is conducted to support how core content can be intertwined with science practices and crosscutting concepts in traditional forms of assessment. The National Research Council's report entitled *Developing Assessments for the Next Generation Science Standards* (Pellegrino, et al., 2014) serves as a supplement to the NGSS and offers guidance aimed to help practitioners create innovative assessment materials that align with the three-dimensions of the standards. While this material is beneficial to educational researchers, assessment developers, and K-12 practitioners, it is arguably less likely that college science faculty

will take note of such information. Moreover, the types of assessments proposed may not be feasible in large-enrollment general education science courses. Thus, there is an impetus to embed measures of science practices into traditional forms of assessment, such as multiple-choice exam items, for a practical, value-added approach to summative assessment. The research herein focuses on the development and use of multiple-choice items to measure both chemistry content and science practices in a large-enrollment general chemistry course.

Educational assessment experts assert that student learning would improve if assessment, curriculum, and instruction were connected more intrinsically (Bransford, Brown, & Cocking, 1999; Pellegrino, et al., 2014). Yet, instructors' knowledge of assessment practices may be limited (Emenike, Raker, & Holme, 2013; Raker, Emenike, and Holme, 2013; Stiggins, 1991). One component of this research project was to determine how a rubric, designed to evaluate assessment material for the three dimensions of learning outlined within the *Framework*, could be used to aid in the creation of multiple-choice items that explicitly incorporate measures of science practices. By creating the multiple-choice items through the use of this rubric, it becomes easier to understand how a rubric can be used as a tool to support and guide instructors who choose to create assessment materials to measure science practices.

Research Questions

The following research questions guided the work of this chapter:

1. How can the 3D-LAP rubric be used to construct items for explicit incorporation of science practices?

2. How does student performance on items with science practices incorporated compare to items without science practices?

Methods

Human Subjects Procedures

This study was conducted under full compliance with the Iowa State University's Institutional Review Board (IRB) policies regarding human subjects research. The study was approved under exempt status (IRB ID 12-424). Since the study did not involve participation beyond regular classroom activities, only analysis of examination data, there was no need to solicit participants and gain consent.

Participants and Instructor Descriptions

Participants in this study were students in the first semester of a full-year general chemistry course at a large mid-western research university during the fall semester of 2014. The laboratory and lecture function as separate courses in this paradigm, thus this research focuses only on lecture component of the course. The course is designed primarily for life and physical science majors, including some forms of engineering. The majority of students who enroll in the course are classified as "freshman," and are in their first year at the university. Multiple instructors, two of whom agreed to participate in this study, teach the course. For the remainder of this study, they will be referred to as Instructor A and Instructor B.

Both instructors had taught this course on multiple occasions prior to the semester the study took place. Instructor A taught one section of the course, whereas Instructor B taught two sections, including the section containing all of the honors students. The self-reported pedagogies of the instructors were slightly different, but were generally

consistent with a traditional large-enrollment lecture course. Instructor A reported using a more conceptual approach when engaging with course content and problem solving, while Instructor B frequently incorporated and demonstrated problem solving exercises. These choices led to Instructor B covering slightly more material than Instructor A. Neither instructor reported discussing the development of science practices explicitly with students. Both instructors used student response systems (clickers) in their lectures as formative assessments. Students were also enrolled in a weekly recitation section led by a teaching assistant.

In addition to slight differences in teaching pedagogies, the summative assessments used by the instructors also varied slightly. Each instructor used four instructor created hour exams distributed throughout the semester. Instructor A designed exams with approximately 15 multiple-choice questions, followed by two or three free response questions. Instructor B designed exams that were solely multiple choice, and contained 30 items per exam. The instructors often administered two forms of the same exam differing only in item order or answer order. More specific information about exam content can be found in Table 1. The final exam consisted of a standardized general chemistry 70-item multiple-choice exam developed by the American Chemical Society Examinations Institute to assess content associated with the first semester of a two-semester general chemistry course.

Instructor A had a total of 300 students and Instructor B had 625 students at the end of the semester. A distribution of final letter grades by course instructor is shown in Figure 1. Due to attrition and other factors, not all students took every exam.

Item Construction

Items were constructed by a team of chemistry education researchers in conjunction with Instructors A and B, and were designed to incorporate science practices as defined by the *Framework* (National Research Council, 2012) and operationalized by the Three-Dimensional Learning Assessment Protocol (3D-LAP) (Cooper, 2014; Underwood, et al., 2014). Additionally, item content was situated within a macroscopic scale scenario, or phenomenon, within the item. A more thorough discussion of the 3D-LAP and its modification for use in relation to chemistry assessments can be found in the previous chapter.

The instructors provided the item writers with topics to be covered on each exam. The item writers drafted several items, and worked with the instructors to make revisions as necessary. The instructors selected which items they would like to use on their exams. In general, a small number of items, including the 3D-LAP constructed items, were the same across instructors' exams. All items designed to contain science practices were multiple-choice, and also contained a phenomenon, as defined by the previously discussed rubric. The 3D-LAP was consulted as items were drafted to ensure that the criteria for incorporation of a specific science practice were being met, and whether the practice was implicitly or explicitly incorporated into the item. The instructors initially agreed to include two 3D-LAP items on each of the four course exams, but were so interested in these items by the end of the semester that they agreed to include four items on the fourth exam. The Appendix following this chapter contains each of the items crafted by the item writing team.

Data Analysis

Psychometric methods

Classical test theory was used as the primary methodological framework for analysis of student performance on exam items. In this theory of assessment, a linear model relates the test score (*X*) to the true score (*T*) and the error score (*E*) (Alagumalai & Curtis, 2005):

$$X = T + E$$

The concepts of true score (*T*) and error score (*E*) are latent constructs in this model, meaning they cannot be directly measured. Thus, a student's observed score on an assessment is comprised of the student's true, or deserved, score (*T*) and some amount of error (*E*). The concept of an error score within an assessment is not the primary concern, because all assessments have some degree of error. Care should be taken, however, to minimize this error by creating assessments that are both valid and reliable. Classical test theory assumes that the true score and the error score are uncorrelated, the average error score is 0 within the population assessed, and that parallel tests have uncorrelated error scores. A limitation of this theory is that the item statistics are sample dependent, yet it is still considered an advantageous theory in the fact that its assumptions are easy to meet and the item statistics involved are well-known and frequently used in assessment.

The two most common statistics used in the classical test theory model are difficulty (*p*) and discrimination (*D*). Item difficulty is somewhat counterintuitive as it represents the proportion of students who respond correctly to an item (Alagumalai & Curtis, 2005; Ding & Beichner, 2009), and thus items with a larger difficulty index value are considered easier. In this study, difficulty values below 0.20 are considered difficult

items whereas items above 0.8 are considered easy. Since the constructed items contain five response choices, item difficulty values at, or below, 0.20 are consistent with performance at a level of random guessing. Discrimination represents the difference in the proportion of high performing and low performing students who respond to an item correctly (Alagumalai & Curtis, 2005; Ding & Beichner, 2009). It is generally accepted practice to calculate item discrimination values based upon groups representing the top 27% and bottom 27% of student performance scores on an exam (Feldt, 1961). An item with a discrimination value of 1 would mean that all students in the high achieving group answered the item correctly, while all of the students in the low achieving group answered the item incorrectly. This scenario is highly unlikely to occur, thus it is generally accepted that items with discrimination values above 0.4 are considered to have high discrimination, while items below 0.2 have low discrimination. Plots of difficulty versus discrimination provide a visual representation of student performance on individual items, and give a general idea as to the overall easiness of an exam. Plots of this nature were used extensively during the analysis of exam performance to examine how items designed to contain science practices compared to other instructor-developed items.

Item Response Curves (IRCs) provide additional measures of discrimination by relating response choices for an item to the percentage of students at each possible total score (Morris et al., 2006). In an IRC, the y-axis constitutes the percentage of students who selected a particular response choice while the x-axis represents students' total scores. There are no set criteria for interpreting an IRC to determine how well an individual response choice discriminates, so interpretation is highly arbitrary. Despite the

arbitrariness of interpretation, attention to the slope of the curve provides insight as to how high and low scoring students interacted with an individual response choice. Positive slopes indicate that the response choice tended to be selected more often by high scoring students than by low scoring students. The opposite is true for negative slopes. A response choice with a fairly flat slope does not discriminate well because a similar percent of high and low scoring students chose the response. Since the distribution of total scores is often somewhat bell-shaped, it is important to consider that there are relatively few students at the extremes of the score range when reviewing an IRC.

Statistical tests

Basic statistical tests, primarily independent samples *t*-tests, were used to detect differences in performance on items with and without science practices after determining that the samples had homogeneity of variances. Effect sizes were measured with Cohen's *d* such that 0.2 was considered small, 0.5 medium, and 0.8 was considered a large effect size (Cohen, 1992). It is important to note that in instances where there were multiple forms of an exam, analyses were conducted on each form independently. This is because even though the items on the two forms of the exam were inherently the same, by altering their order and or the order of response choices, a seemingly benign alteration, the psychometric properties of the item have also likely been altered. Due to a possibility of the presence of item order or answer order effects, it would not be considered appropriate to aggregate data across multiple forms of an exam (Schroeder, Murphy, & Holme, 2012). Additionally, frequency counts were used to provide information about the number of times specific science practices were incorporated.

Results and Discussion

Review of Instructor Created Items for Science Practices

In addition to the items created by the researchers to incorporate specific science practices, it was of interest to analyze how the remaining instructor created multiple-choice items may incorporate the use of science practices. The rating duo described in the previous chapter utilized the 3D-LAP to classify all of the multiple choice items constructed by Instructors A and B for their respective exams by following the protocol detailed in the previous chapter. In short, the raters reviewed the instructor created exams independently and then convened to discuss classification of items based upon incorporation of science practices and phenomena.

Of the 65 total multiple-choice items used by Instructor A, 14 items (21.5%) contained a science practice. On exams in the course taught by Instructor A, five (35.7%) out of the 14 items with a science practice had been created by the instructor rather than the research team. Similarly, Instructor B used 120 total multiple-choice items, and 15 items out of 24 (20%) that contained a science practice had been created by the instructor. It is important to clarify that the instructor created items containing a science practice were not created with specific reference to the 3D-LAP. Rather the practices were present in these items because they were inherently coupled to the chemistry content assessed, such as the presence of a Lewis structure in an item invokes the use of the practice related to developing and using models. Even though these items were not constructed under explicit direction from the 3D-LAP, the presence of science practices within the items suggests that chemistry, as a discipline of study, tends to lend itself to the incorporation

of science practices in assessment measures, and that instructors value these practices even when they are not aware of the specific educational underpinnings of the practices.

The value of incorporating science practices into assessment items, as perceived by the instructors, can be seen in Figure 2. Science practices were scarce on Exam 1 but became more commonplace as the semester progressed, and by Exam 4 were present in over one-quarter of each instructor's items. Again, some of the incorporation stemmed from the nature of the content assessed, but it was also influenced by the instructors' comfort level with such items as the semester progressed. Satisfactory student performance on items containing science practices likely assisted these endeavors.

A distribution of items across the eight science practices can be seen in Table 2. While there were markedly fewer items in this analysis as compared to the analysis of ACS exam items in the previous chapter, similar trends still existed. The practice of developing and using models was still the most predominant practice incorporated into exam items. Constructing explanations had less of a presence in the instructors' exams as compared to the ACS exams, but the practice of engaging in argument from evidence was incorporated frequently and was the second most common practice found within the instructors' exams. The laboratory functions as a separate course, so it is not unexpected that no items related the practice of planning and carrying out investigations were found on these tests. Additionally, consideration of the content of the course provides some explanation for the distribution of science practices. The course does not include topics such as kinetics and equilibrium, which incorporated practices related to SP 4 and SP 5 frequently on ACS exam items.

Incorporation of phenomena was also analyzed because phenomena provide a macroscopic, relevant connection between the chemistry content and students' everyday lives. These connections may aid in students' constructions of their own knowledge schema and provide opportunities for meaningful learning to occur. Chemical phenomena were incorporated in 8 (12.3%) of Instructor A's items and 25 (20.8%) of Instructor B's items. The presence of phenomena in instructor created exam items was not a pronounced trend, but it is unclear as to why this trend appears in testing because in conversations with the instructors, both provided anecdotal evidence of incorporation of phenomena in discussions of content. In all likelihood, this aspect of item creation was neglected because it is not commonly a component of traditional item construction.

Student Performance on Items Containing Science Practices

Comparison of student performance on items with and without science practices supports the hypothesis that adding science practices does not alter their psychometric properties unfavorably. Plots of item difficulty versus discrimination provide a visual representation of how exam items perform relative to one another. These plots were beneficial when examining the performance of items with and without science practices. Ideally, the majority of items on an exam should fall near the center of the plot, indicating acceptable difficulty and discrimination indices. Plots of item difficulty versus discrimination for all items constructed to incorporate science practices through the use of the 3D-LAP can be found in the appendix to this chapter.

Analysis of difficulty and discrimination plots indicates that items with science practices function similarly to the other exam items. For example, the difficulty and

discrimination plot shown in Figure 3 represents the performance of 337 students from Instructor B's course on Form 1 of Exam 2. In this figure, items 5 and 6 were created by the research team, and items 18, 25, 28, and 29 were created by the instructor but also incorporate a science practice. Items 5, 6, 25, and 29 all function in the ideal ranges of difficulty and discrimination, whereas items 18 and 28 are considered too "easy." The items with science practices tended to function similar to the other exam items on all of the exams throughout the semester. Instructor A had fewer multiple-choice items on each exam, and also tended to vary item order or response order more so than Instructor B, so it is not surprising there is more spread within the plots representing Instructor A's exams. Since the research design is of a practical, empirical nature, it was not possible to trial-test items before the exam to evaluate their psychometric properties, however, by following best practices for constructing multiple-choice items, the items with science practices tended to have psychometric properties in line with the remaining exam items.

Additionally, comparison of average item difficulty for items with and without science practices revealed that items containing science practices were significantly more difficult than items without science practices regardless of the instructor. Comparison of average difficulty on items with and without science practices on Instructor A's exams using an independent samples *t*-test revealed a significant difference between items with science practices ($N = 14$, $M = 0.546$, $SD = 0.15$) and items without science practices ($N = 51$, $M = 0.703$, $SD = 0.19$; $t\,(63) = 2.805$, $p = 0.0033$). The magnitude of the difference of the means was large (Cohen's $d = 0.85$) indicating that the significance detected was not a fluke, and represents a genuine difference in performance on these two item types. Instructor B's items yielded similar results with an independent samples *t*-test as

Instructor A's items. Items with a science practice (N = 24, M = 0.612, SD = 0.16) were significantly more difficult than items without a science practice (N = 96, M = 0.705, SD = 0.15; t (118) = 2.6079, p = 0.0051). The effect of this comparison was moderate (Cohen's d = 0.60), indicating, again, that students' performance on items with and without science practices was legitimately different. While student performance on items with science practices is significantly lower than items without science practices, the items with science practices have respectable psychometric evidence of their validity, as seen in the difficulty and discrimination plots. The difference in performance likely stems from the type of instruction and formative assessments used within the course. Students had not been routinely expected to explicitly engage in the practices in formative assessments during lecture or recitation, so it is not surprising that performance on items requiring the use of science practices was not as strong. This result provides a good example to support the idea that instruction and assessment should be aligned, and suggests that implementing measures of science practices into assessment items may make these items more challenging when the course instruction does not emphasize these practices routinely.

Student Performance by Achievement Level

Comparison of high versus low achieving students on the correctness of items with science practices was of little value because students within the high achieving group are inherently more likely to answer an item correctly compared those in the low achieving group. A better approach to understanding how the items with science practices discriminated across performance groups was to construct item response curves (IRCs). By comparing total score to the percentage of students who selected a particular response

choice, it was easier to determine how the items with science practices performed across various performance levels as compared to the psychometric data of item performance. IRCs for all 3D-LAP created items can be seen in the Appendix following this chapter.

The IRCs provide information about how the distractors discriminate across performance groups. The desired outcome is to create distractors that are highly discriminating, so the IRCs provide insight into how future items may consider revising distractors. The IRCs for 3D-ITEM 2 (Figure 4 and Figure 5) reveal interesting information about students' conceptions of stoichiometry. Students scoring below approximately 80% on Instructor A's exam and 68% on Instructor B's exam were more likely to select distractor "C," thus indicating that they were not able to recognize the conceptual stoichiometric aspect of the item, and instead relied on false assumptions of base strength to support their claim. The IRCs for 3D-ITEM 8 (Figure 6 and Figure 7) revealed some instructor effects. Low performing students in Instructor A's course were more likely to use incorrect evidence to support their argument, even though they selected the correct graph, than the low performing students in Instructor B's course. Instructor A shared anecdotally that Coulomb's Law was emphasized frequently during this unit, yet no other items on this exam were about Coulomb's Law. Perhaps these students were attempting to use test-savvy rather than reasoning to answer this item, because they expected Coulomb's Law to be tested since it was emphasized during instruction. 3D-ITEM 11 was only incorporated on Instructor B's exam, but it still revealed that some students have misconceptions when constructing explanations about bond energy as shown in Figure 8. High performing students, those who scored above 25 out of 30 points, were able to construct an explanation based upon the correct principle

that bonds forming releases energy, but those scoring below 25 points were more likely to select that bonds breaking releases energy and is the reason why the reaction is exothermic. IRCs for other items revealed that they were relatively non-discriminating. For example, 3D-ITEM 7 on an exam from Instructor B performed well across all performance levels, indicating that revision of the distractors could be considered if they item was to be used again on future assessments. Additionally, the IRC for 3D-ITEM 7 compared to 3D-ITEM 11 suggests that the content of the items is more likely the cause for differing student performances than incorporation of the science practice of argumentation. This aligns with the findings of the previous chapter in which differences in student performance were noted across content ideas compared to incorporation of individual practices.

Conclusions

This study was an empirical investigation on the design and implementation of multiple-choice items that incorporate science practices. The creation of items that explicitly incorporate science practices was aided by the use of the 3D-LAP, but still remained a fairly complex task. This suggests that while the 3D-LAP is a valuable tool to aid creation and evaluation of assessment items, in order to be of use to the traditional practitioner, guidance from chemistry education researchers may be necessary in the form of additional research on the development and performance of items that incorporate science practices. Yet, the incorporation of science practices within multiple-choice assessment items is feasible and, based upon student performance, should be encouraged when meshed with appropriate instruction.

The instructors did not alter their pedagogies to explicitly incorporate science practices into instruction. In this sense, instruction and assessment were misaligned and this likely explains the difference in student performance on items that contained a science practice compared to items without a science practice. Even though items incorporating science practices were more difficult, the psychometric performance data of these items were, in general, quite reasonable, suggesting that multiple-choice items to assess science practices and chemistry content are viable options for instructor created assessments. It is posited that, in future endeavors, the gap in performance on items with and without science practices can be bridged through instruction.

Limitations of this study relate to the comparisons that can be made across items with and without science practices. Due to the time constraints of exam administration and the necessary measures of content proficiency desired by the instructors, it was not feasible to create multiple items assessing the same content with and without science practices. Future work could examine these comparisons within the context of instructor created examinations. Additionally, comparisons of performance on multiple-choice versus open-ended items containing science practices were not within the realm of this study, but such comparisons would likely be highly informative to the design of assessment materials to measure science practices.

Ultimately, incorporation of science practices into multiple-choice exams at the classroom level is feasible and does not appear to negatively impact overall exam performance. The research herein supports the efforts to surmount the challenge of successfully incorporating science practices into traditional modes of assessment. Yet, an additional challenge appears to be aligning instruction to teach the science practices that

are assessed. Osborne offers suggestions for teaching science practices within the new paradigm of the NGSS (2014). New directions for research in the chemistry education community ought to consider how to best support chemistry practitioners as they prepare to engage in the teaching and assessment of science practices.

References

Achieve. (2013). Next generation science standards: Washington, DC: National Academies Press.

Alagumalai, S., & Curtis, D. D. (2005). *Classical test theory*: Springer.

Cohen, J. (1992). A power primer. *Psychological bulletin, 112*(1), 155.

Cooper, M. M. (2014). Personal Communication.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics-Physics Education Research, 5*(2), 020103.

Emenike, M., Raker, J. R., & Holme, T. (2013). Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *Journal of Chemical Education, 90*(9), 1130-1136.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika, 26*(3), 307-316.

Morris, G. A., Branum-Martin, L., Harshman, N., Baker, S. D., Mazur, E., Dutta, S., et al. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics, 74*(5), 449-453.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

Osborne, J. (2014). Teaching scientific practices: Meeting the challenge of change. *Journal of Science Teacher Education, 25*(2), 177-196.

Pellegrino, J. W. (2014). Assessment as a positive influence on 21st century teaching and learning: A systems approach to progress. *Psicología Educativa, 20*(2), 65-77.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*: National Academies Press.

Raker, J. R., Emenike, M. E., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education, 90*(8), 981-987.

Schroeder, J., Murphy, K., & Holme, T. A. (2012). Investigating factors that influence item performance on ACS exams. *Journal of Chemical Education, 89*(3), 346-350.

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan, 72*(7), 534-539.

Underwood, S. M., Cooper, M. M., Krajcik, J., Cabellero, D., & Ebert-May, D. (2014). *Designing a rubric to characterize assessments*. Paper presented at the 248th National Meeting of the American Chemical Society.

Table 1. Descriptions of the content coverage of exams given by each instructor.

	Instructor A	Instructor B
Exam 1	Balancing chemical equations, chemical formulas and naming, density calculations, isotopic abundance, historical experiments, mass spectroscopy measurement and error, phases and properties of matter, reactivity, unit analysis	Balancing chemical equations, chemical formulas and naming, conversions involving the concept of moles, density calculations, isotopic abundance, historical experiments, mass spectroscopy measurement and error, phases and properties of matter, reactivity, unit analysis
Exam 2	Net ionic equations, mass percent and percentage yield calculations, conversions involving the concepts of moles, solution chemistry, stoichiometry	Net ionic equations, mass percent and percentage yield calculations, solution chemistry, stoichiometry, redox chemistry
Exam 3	Dilutions, enthalpy, Hess's Law calculations, thermodynamics, thermochemistry and calorimetry, redox chemistry, wave nature of light	Electronic transitions, electron configurations, enthalpy, Hess's Law calculations, thermodynamics, thermochemistry and calorimetry, wave nature of light
Exam 4	Chemical bonding, electron configurations, electronic transitions, Lewis Structures, periodic trends, photoelectric effect	Chemical bonding, Lewis Structures, periodic trends, reactivity trends

Table 2. Distribution of items by science practice and instructor.

Science Practice	Frequency of Occurrence	
	Instructor A	**Instructor B**
1. Asking questions	1	0
2. Developing and using models	6	14
3. Planning and carrying out investigations	0	0
4. Analyzing and interpreting data	0	0
5. Using mathematics and computational thinking	1	1
6. Constructing explanations	0	1
7. Engaging in argument from evidence	7	10
8. Obtaining, evaluating, and communicating information	2	1

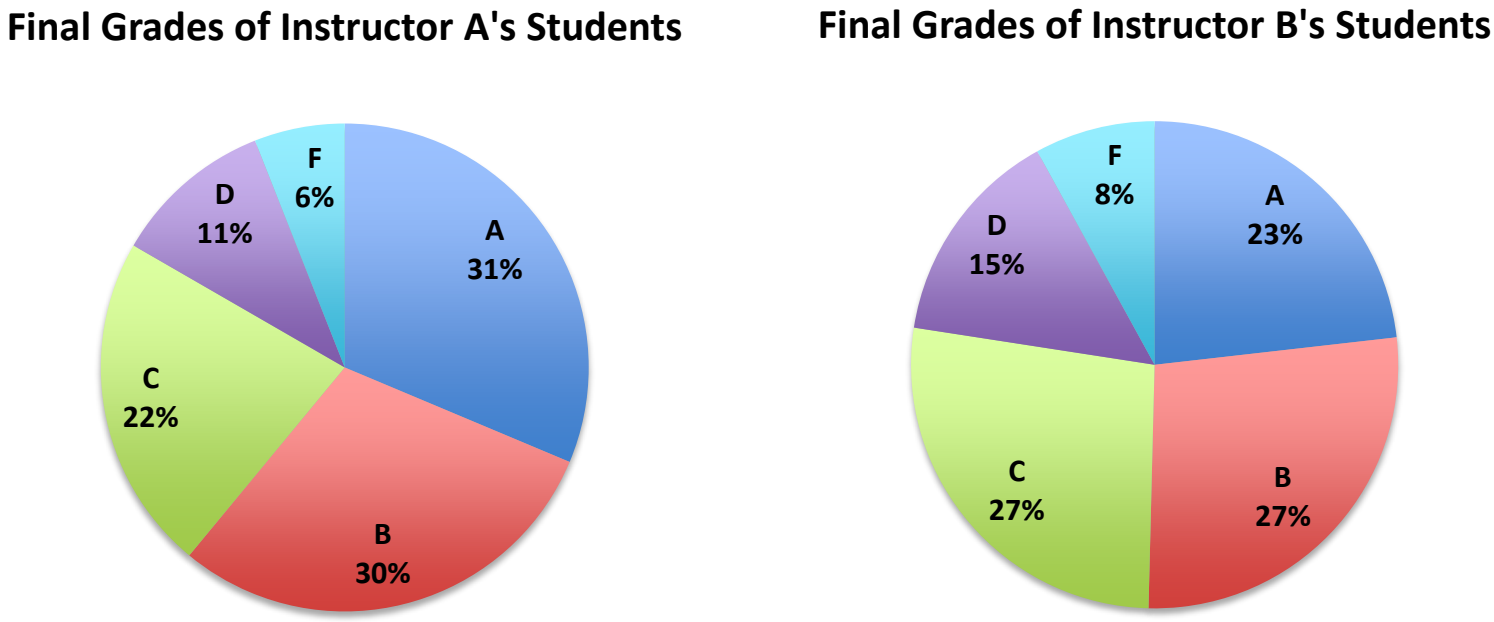


Figure 1. Final grades of students in Instructor A's (N = 300) and Instructor B's (N = 625) general chemistry courses.

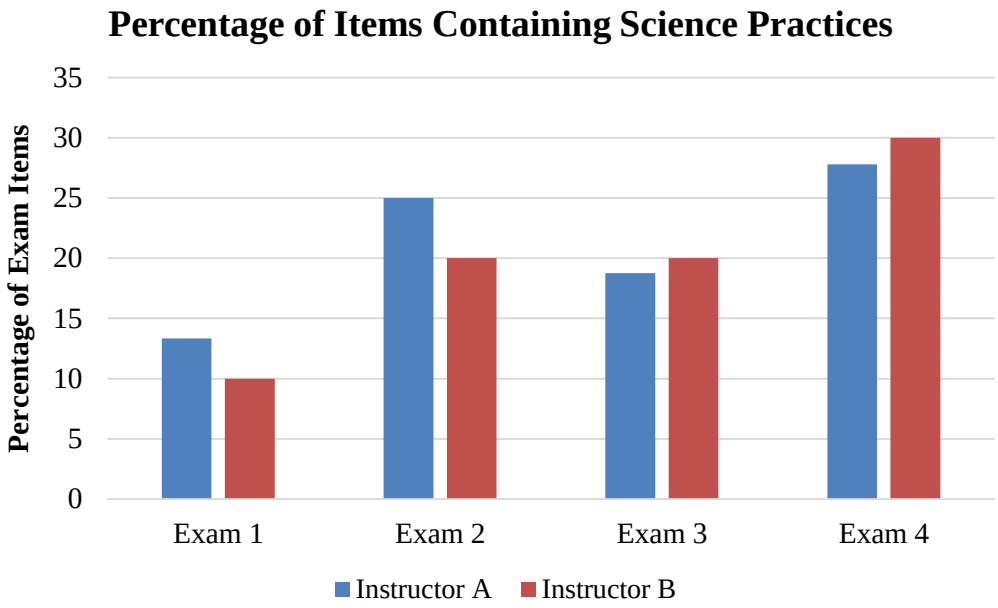


Figure 2. Relative percentage of exam items containing science practices compared across instructors.

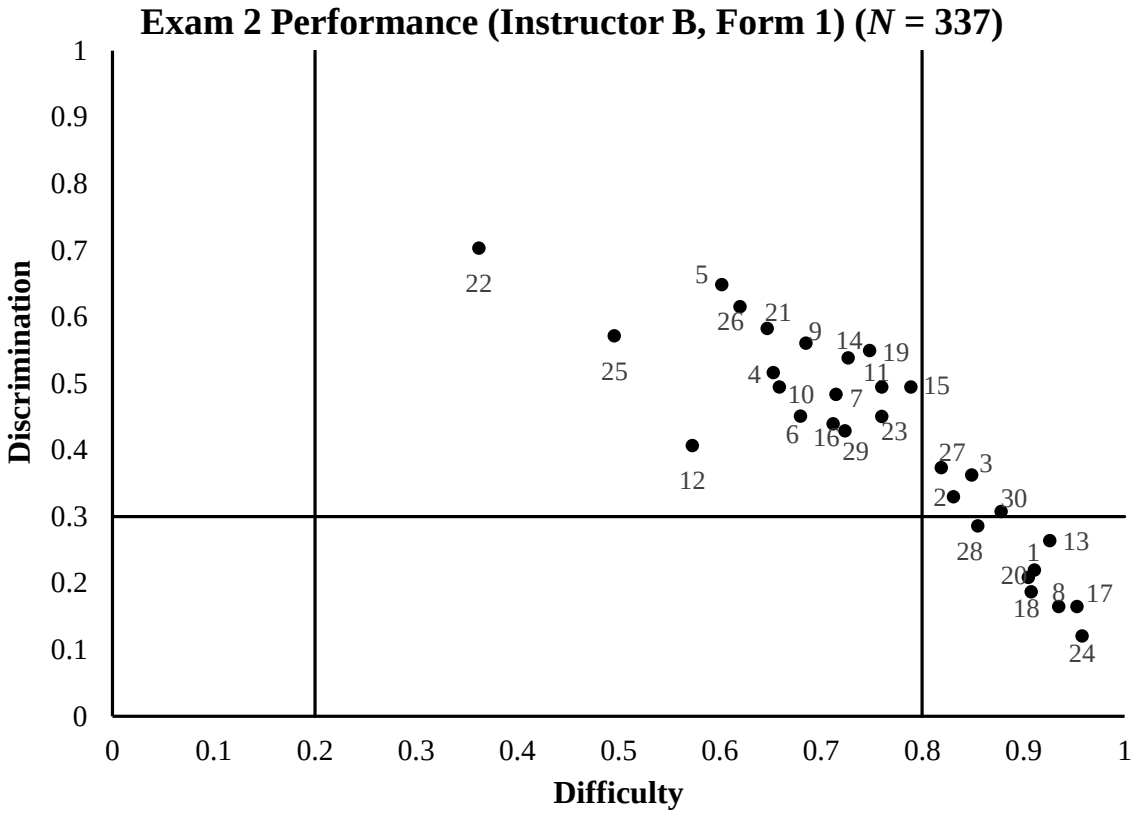


Figure 3. Exam 2 performance on Form 1 given by Instructor B. Items 5 and 6 were created for research purposes to incorporate a science practice. Items 18, 25, 28 and 29 were created by the instructor and also include a science practice.

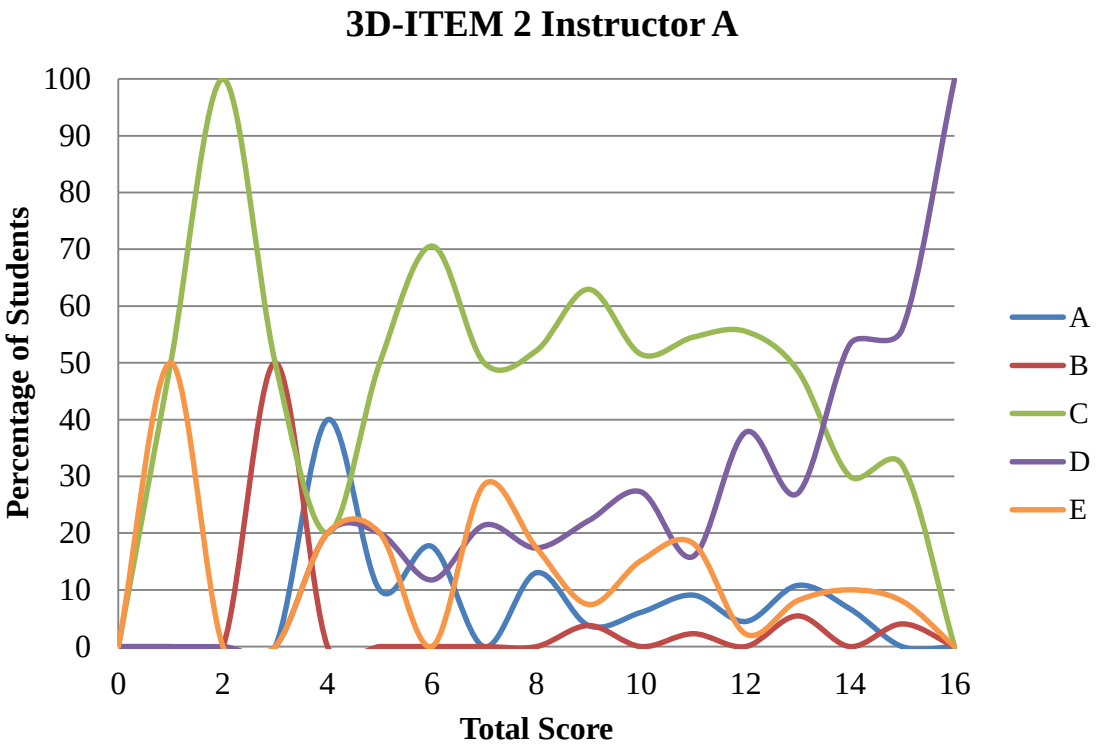


Figure 4. *IRC for 3D-ITEM 2 given by Instructor A on Exam 2 (*N = *331).*

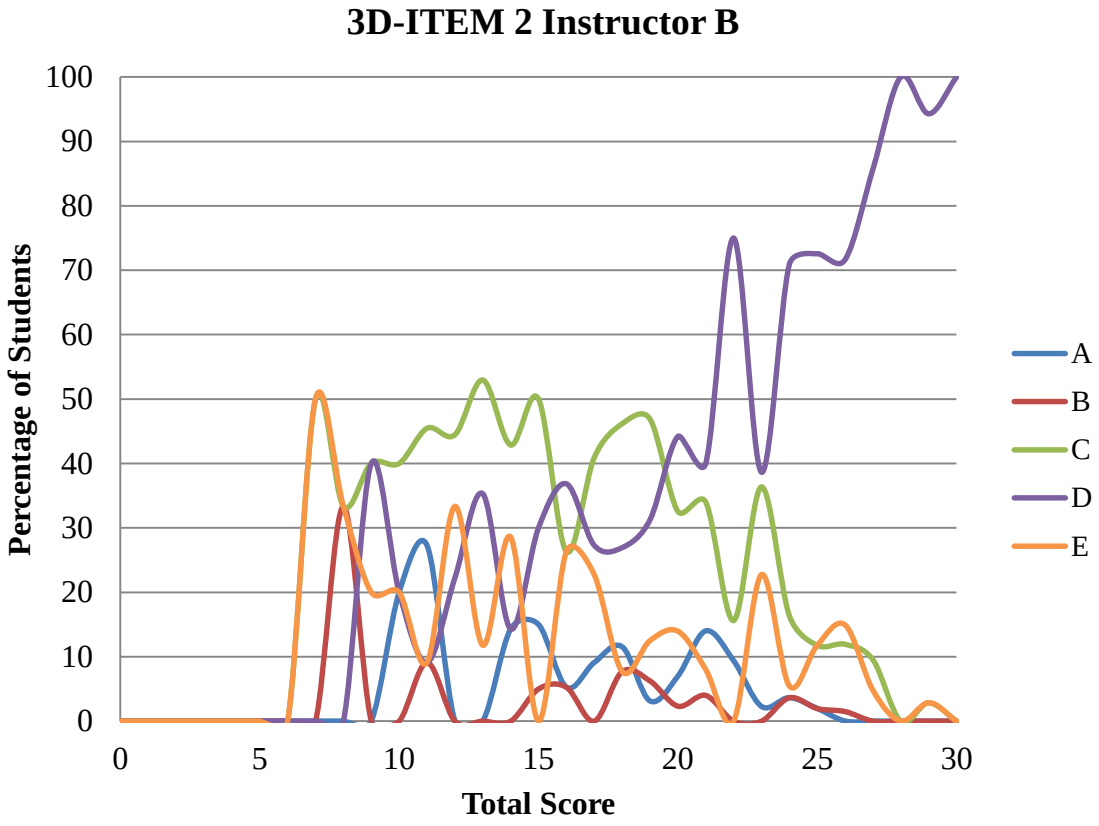


Figure 5. *IRC for 3D-ITEM 2 given by Instructor B on Exam 2 (*N = *661).*

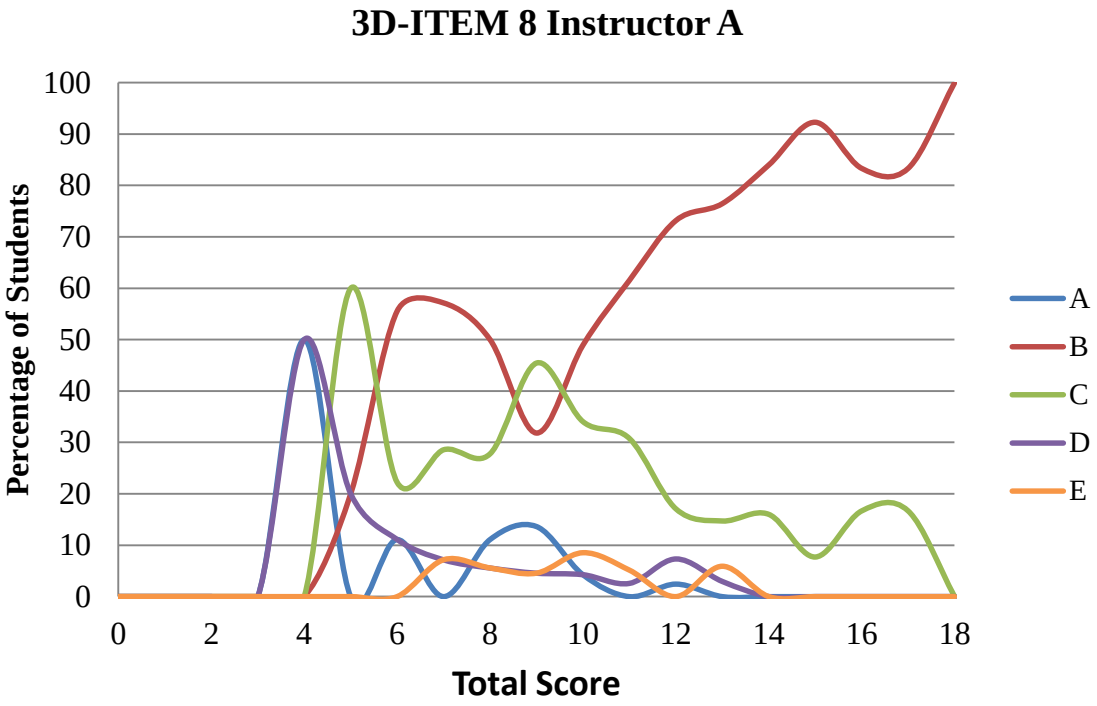


Figure 6. IRC for 3D-ITEM 8 given by Instructor A on Exam 4 (*N* = 283).

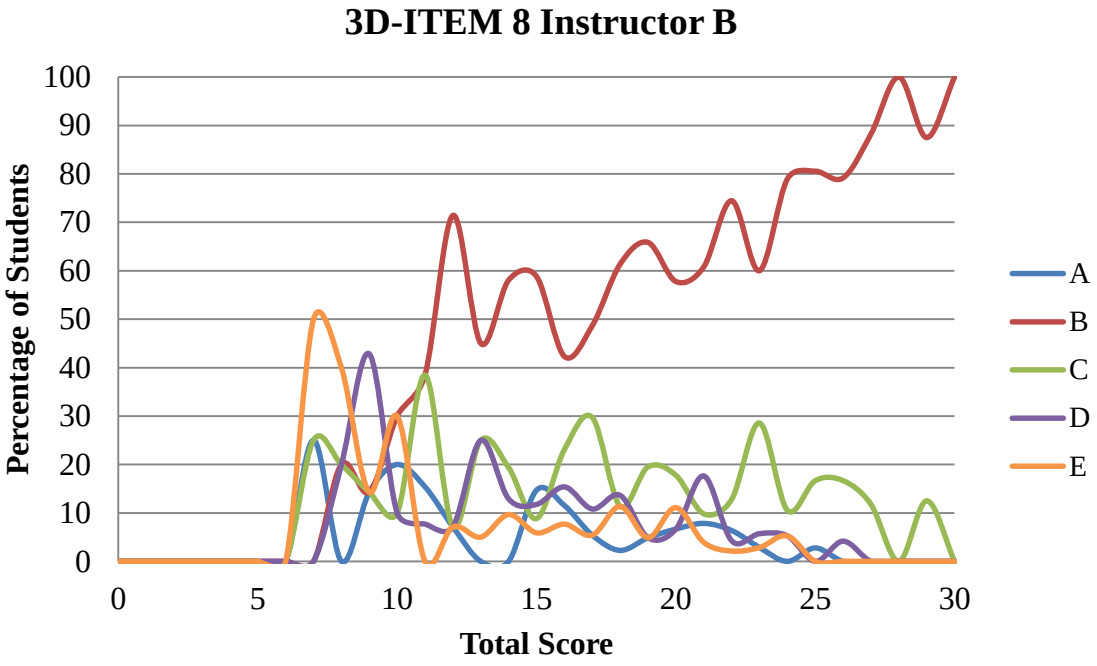


Figure 7. IRC for 3D-ITEM 8 given by Instructor B on Exam 4 (*N* = 607)

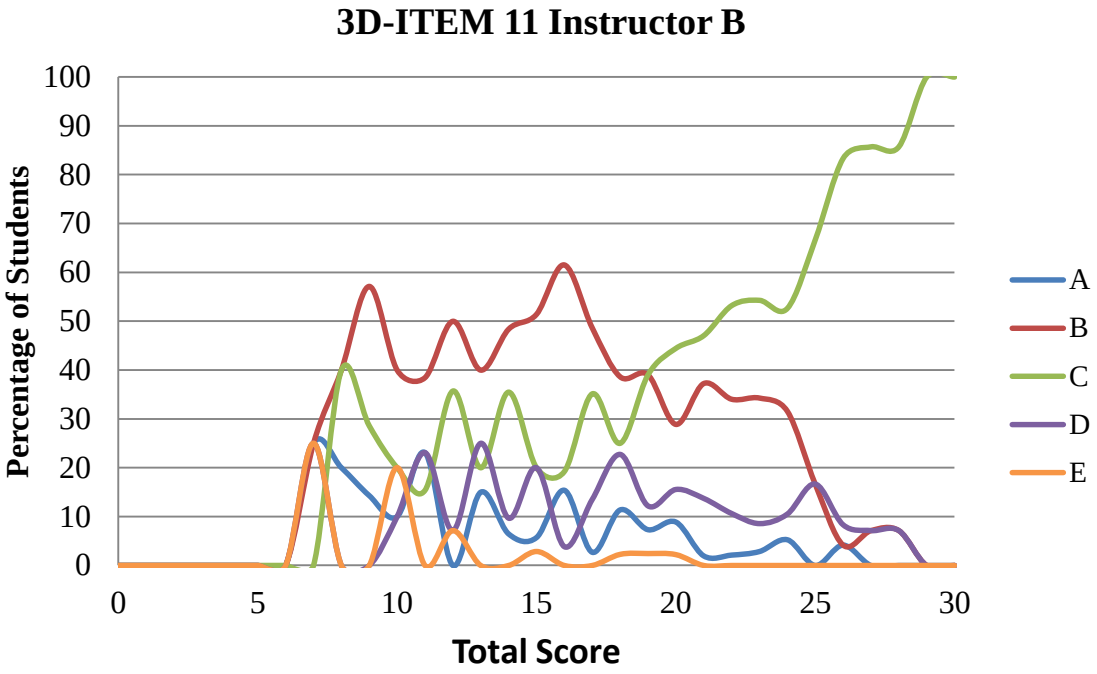


Figure 8. IRC for 3D-ITEM 11 given by Instructor B on Exam 4 (N = 607).

APPENDIX

The items created through the use of the 3D-LAP rubric as well as exam performance data and item-response curves for all 3D-LAP created items are included in this appendix.

3D-ITEM 1:

The density of Mg_2SiO_4(s) is 3.4 g/cm^3 and the density of $MgCO_3$ (s) is 3.1 g/cm^3. If you have a dump-truck sized sample of each form of rock, and the samples are essentially the same mass which one occupies a greater volume and why?

A) The volume of $MgCO_3$ is greater because volume is mass / density.
B) The volume of $MgCO_3$ is greater because it contains CO_2 gas.
C) The volume of Mg_2SiO_4 is greater because volume is mass x density.
D) The volume of Mg_2SiO_4 is greater because it has more atoms per formula unit.
E) It is impossible to determine which volume is greater because the difference in density is too small to measure with accuracy.

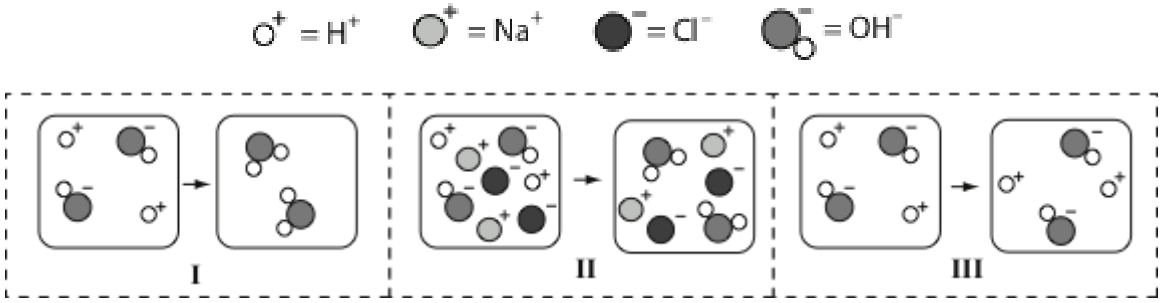
3D-ITEM 2:

A laboratory spill of hydrochloric acid, HCl (aq), can be neutralized by either baking soda, $NaHCO_3$, or washing soda, Na_2CO_3. Suppose a rather large spill of HCl occurs in the lab and you have a box of either baking soda or washing soda available. Which box would you choose to more efficiently neutralize the acid spill and why?

A) baking soda because it will dissolve easier in the acid
B) baking soda because it has a smaller molar mass
C) baking soda because hydrogen carbonate is a stronger base than carbonate
D) washing soda because each mole of carbonate neutralizes two moles of acid
E) washing soda because the increase in sodium present also help neutralize the acid.

3D-ITEM 3:

Sodium hydroxide also neutralizes acid. Which diagram best represents the ***net ionic equation*** for the neutralization of HCl by NaOH? Which of the statements below best describes the reason for your choice?



A) Diagram I because it produces the minimum number of products.
B) Diagram I because it shows reactive ions and products.
C) Diagram II because it shows the chemical species that are actually present.
D) Diagram III because it shows the minimum change between reactants and products.
E) Diagram III because it shows the participating ions are both reactants and products.

3D-ITEM 4:

Suppose there are 75 grams of warm water at 80 °C in a thermally insulated, 100 mL container. You have a 20 gram cube of aluminum at 5 °C and you have 20 grams of water also at 5 °C. If your goal is to get the water in the insulated container as cool as possible, which should you add and why?

A) The aluminum cube because it has a lower heat capacity and therefore absorbs more heat.
B) The aluminum cube because it has a higher heat capacity and therefore absorbs more heat.
C) The cold water because it has a lower heat capacity and therefore absorbs more heat.
D) The cold water because it has a higher heat capacity and therefore absorbs more heat.
E) The cold water because it will mix better with the initial warm water.

3D-ITEM 5:

The heat of formation of $N_2O(g)$ is +82.05 kJ/mol. Which statement is true about the decomposition of dinitrogen monoxide into nitrogen and oxygen and why?

A) It is endothermic because an O_2 molecule must be broken into O atoms.
B) It is endothermic because decomposition reactions are always endothermic.
C) It is exothermic because the heat capacity of N_2O is higher than either N_2 or O_2.
D) It is exothermic because the decomposition is the opposite of the formation reaction.
E) It is exothermic because O_2 would be in the products rather than its usual place in the reactants.

3D-ITEM 6:

In a discussion of bonding during recitation, a student claims that the octet rule states ***"atoms will do anything to get eight electrons."*** Select the response you would give to best critique this student's claim.

It would be better to say…

A) The octet rule means an atom *wants* 8 electrons.
B) The octet rule only applies when drawing Lewis structures.
C) The octet rule is only true for atoms with a formal charge of 0.
D) The octet rule describes a tendency for an atom to have 8 valence electrons.
E) The octet rule means that an atom will gain or lose electrons to have 8 electrons in *total*.

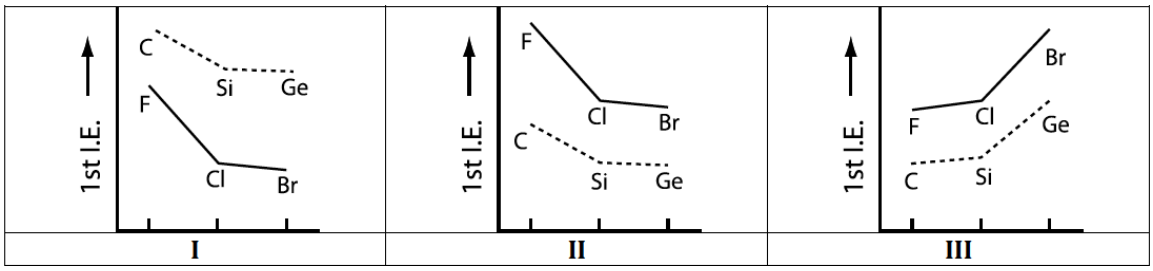
3D-ITEM 7:

One use for potassium nitrate, KNO_3, is as an active ingredient in toothpaste used to treat sensitive teeth. Which statement best describes the bonds within KNO_3, and why?

A) KNO_3 has only ionic bonds because it can be broken into K^+ and NO_3^- ions.
B) KNO_3 has only metallic bonds because potassium metal loses an electron to form K^+.
C) KNO_3 has both ionic and metallic bonds since it contains ions and also potassium metal.
D) KNO_3 has only covalent bonds because the potassium, nitrogen, and oxygen atoms are all sharing electrons equally.
E) KNO_3 has both covalent and ionic bonds since it contains ions and has covalent sharing of electrons within the NO_3^- ion.

3D-ITEM 8:

Which graph correctly depicts the first ionization energy of three elements in groups 14 (dashed line) and 17 (solid line), and what explains the trend that is graphed?



A) Graph I, because group 14 is further away from a noble gas configuration so it will have higher ionization energies.
B) Graph II, because effective nuclear charge is higher in group 17 than it is in group 14 within any given period on the periodic table.
C) Graph II, because the Coulomb's Law forces in the group 17 must be larger than in group 14 because there are more protons and electrons.
D) Graph III, because larger atoms have higher ionization energy both within a period and between groups 14 and 17.
E) Graph III, because in the graph ionization energy becomes more exothermic from left to right as is observed as a periodic trend.

3D-ITEM 9:

Chlorofluorocarbons (CFCs) are molecules that contain only chlorine, fluorine, and carbon. While CFCs are inert in the lower atmosphere, they dissociate in the stratosphere when exposed to high energy UV radiation (UV-C) to form chlorine radicals. These chlorine radicals actively destroy some of the ozone in the stratosphere.

Which question, if answered, would provide the most direct scientific evidence as to why chlorine radicals are generated?

A) What is the Lewis structure of a CFC?
B) What are the relative strengths of bonds in CFCs?
C) What are the bond dipoles in the C-Cl and C-F bonds?
D) What is the VanDerWaal radius of fluorine versus chlorine?
E) What differences are there in the reactions of fluorine and chlorine radicals once they are formed?

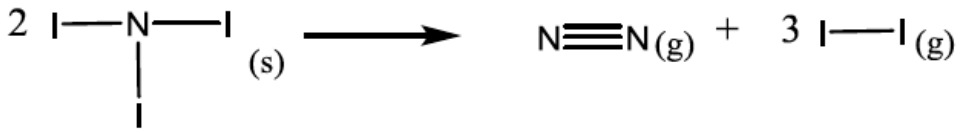
3D-ITEM 10:

20th Century scientists argued that line spectra implied that electrons have quantized energy levels. What about line spectra supports this argument?

A) Electrons move so fast that their positions can't be measured, only their frequencies are shown in the line spectra.
B) Electrons all have the same mass, so they must have the same energy in the line spectra as represented by the equation E = mc2.
C) Since electrons can only transition between certain energies only specific frequencies are observed in the line spectra.
D) If electrons orbited the nucleus, core electrons would have to exceed the speed of light or collide with the nucleus because of Coulomb's law of attractions. Therefore, the line spectra show quantized energy levels.
E) The electron attractions to the positively charged protons in the nucleus must be offset by repulsion from other electrons. This means the speed of orbiting electrons can only be specific values, as shown in the line spectra.

3D-ITEM 11:

Nitrogen triiodide, NI3, is unstable and will spontaneously detonate to form a bright purple cloud of nitrogen and iodine gases accompanied with a loud "bang" and a release of energy. Which is the best explanation for why the decomposition of nitrogen triiodide is exothermic?



A) The formation of gaseous products from solid reactants releases energy.
B) The breaking of the nitrogen-iodine bonds in the reactants releases a lot of energy.
C) The formation of the nitrogen triple bond in the products releases a lot of energy.
D) The decomposition of a compound into stable elements always releases energy.
E) There are more molecules on the products side than on the reactants side, so energy is released.

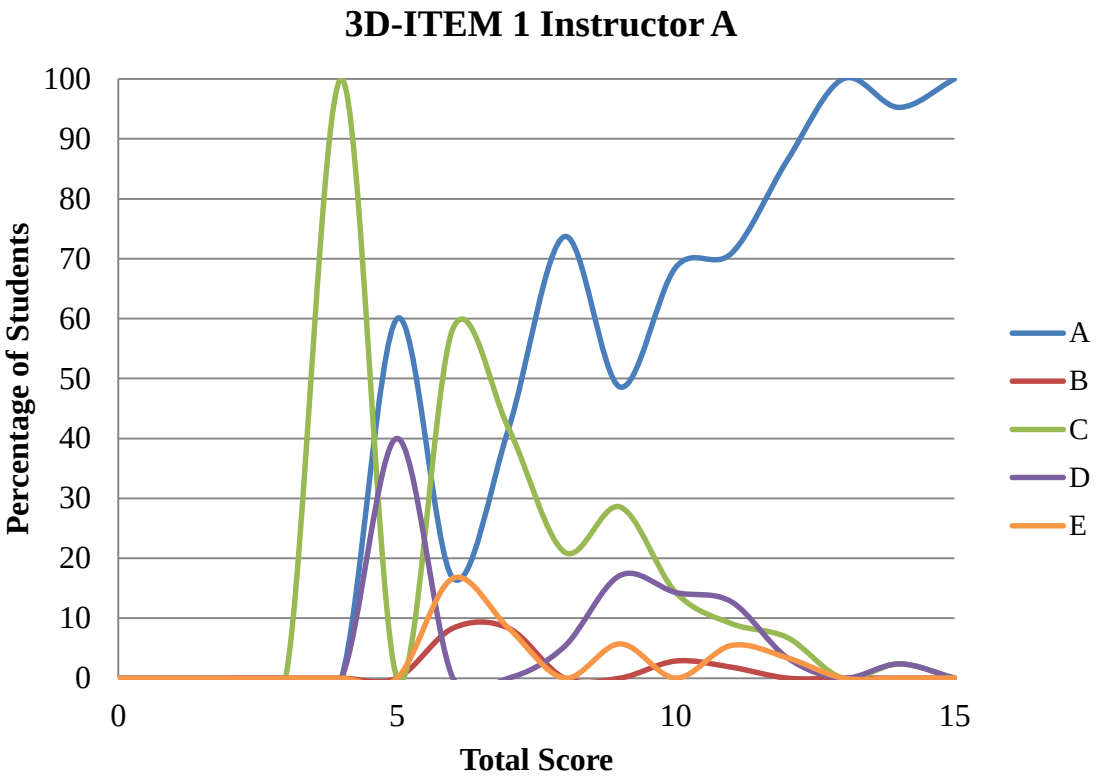


Figure 1. IRC for 3D-ITEM 1 on Exam 1 from Instructor A (*N* =342).

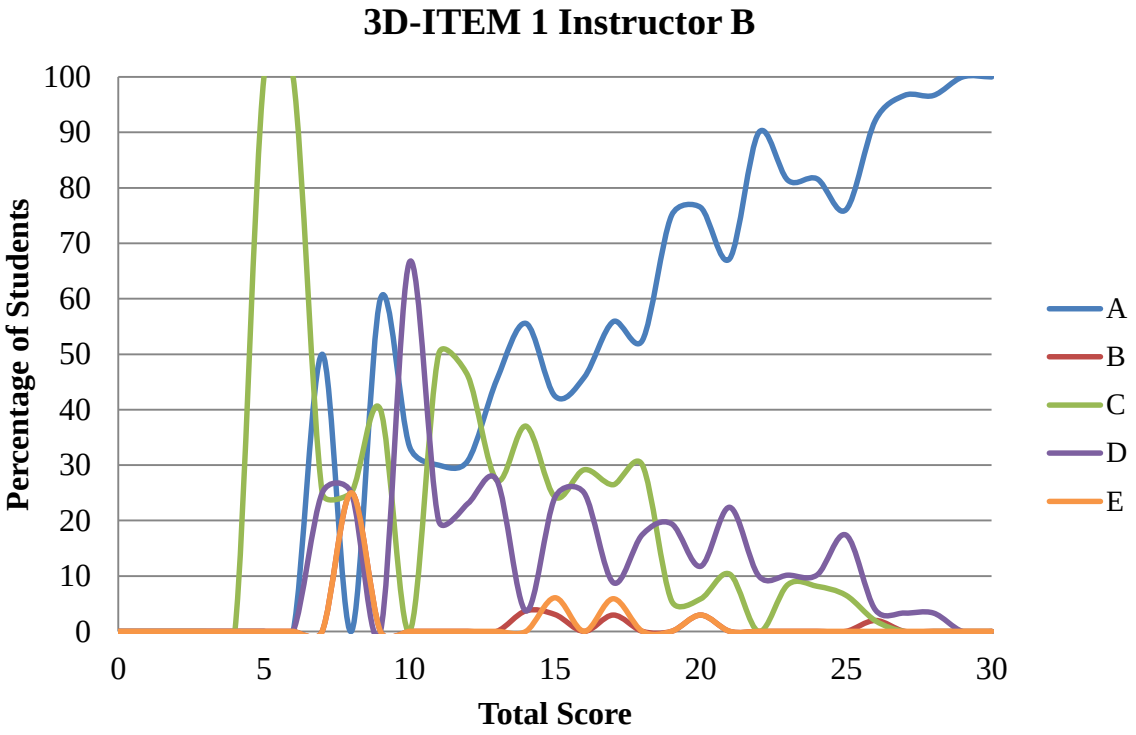


Figure 2. IRC for 3D-ITEM 1 on Exam 1 from Instructor B (*N* = 675).

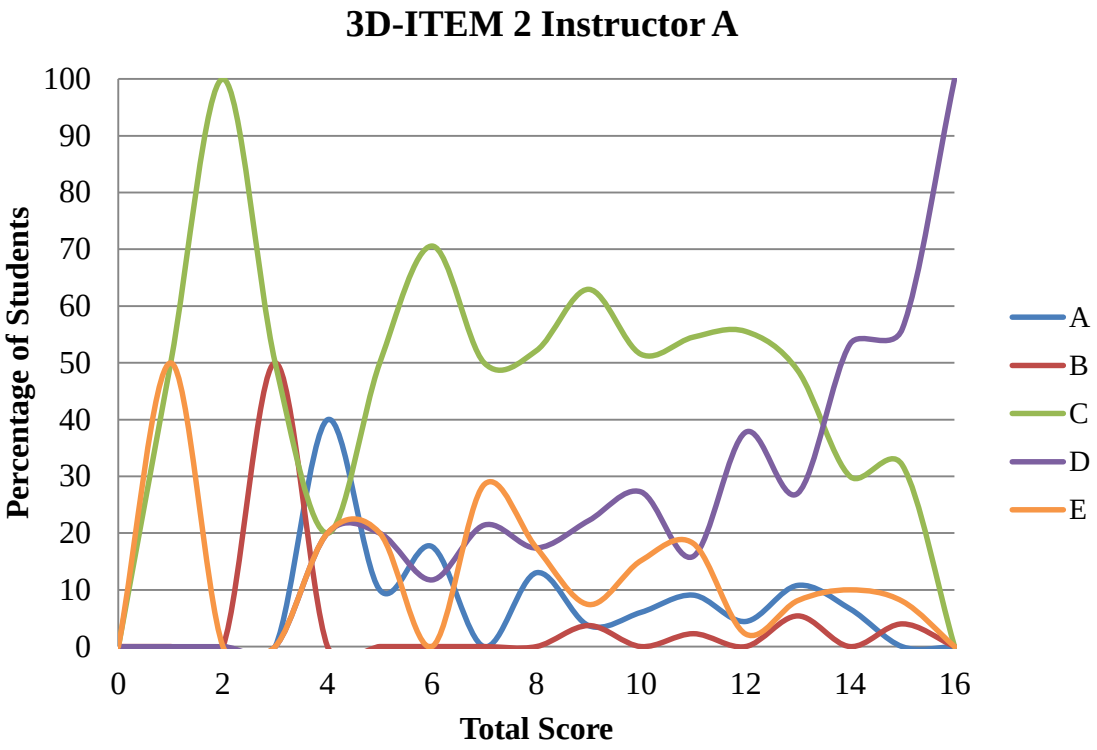


Figure 3. IRC for 3D-ITEM 2 on Exam 2 from Instructor A (*N* = 331).

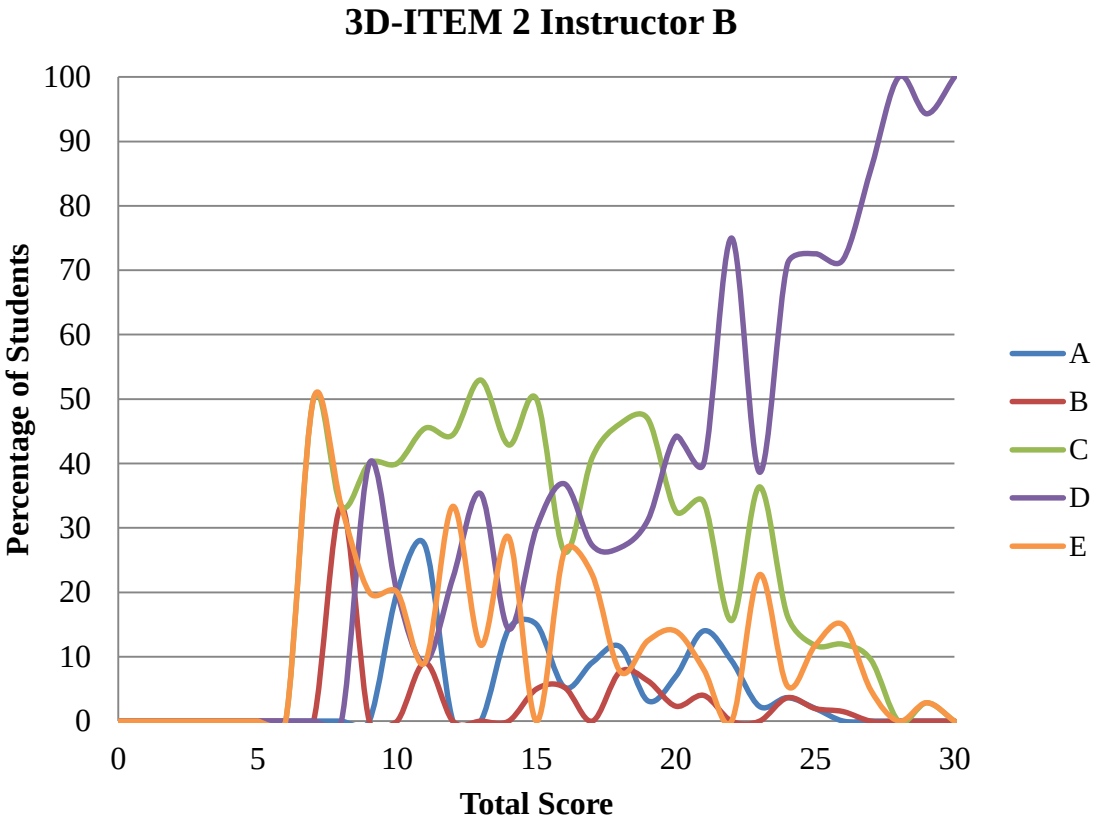


Figure 4. IRC for 3D-ITEM 2 on Exam 2 from Instructor B (*N* = 661).

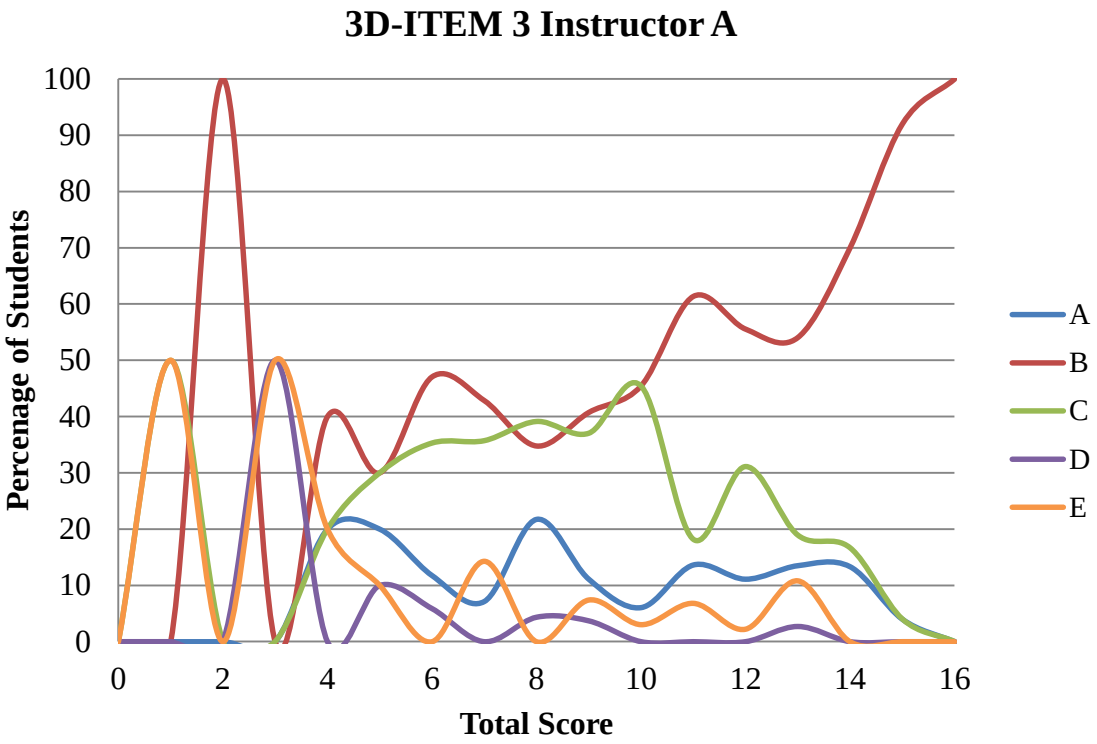


Figure 5. IRC for 3D-ITEM 3 on Exam 2 from Instructor A (*N* = 331).

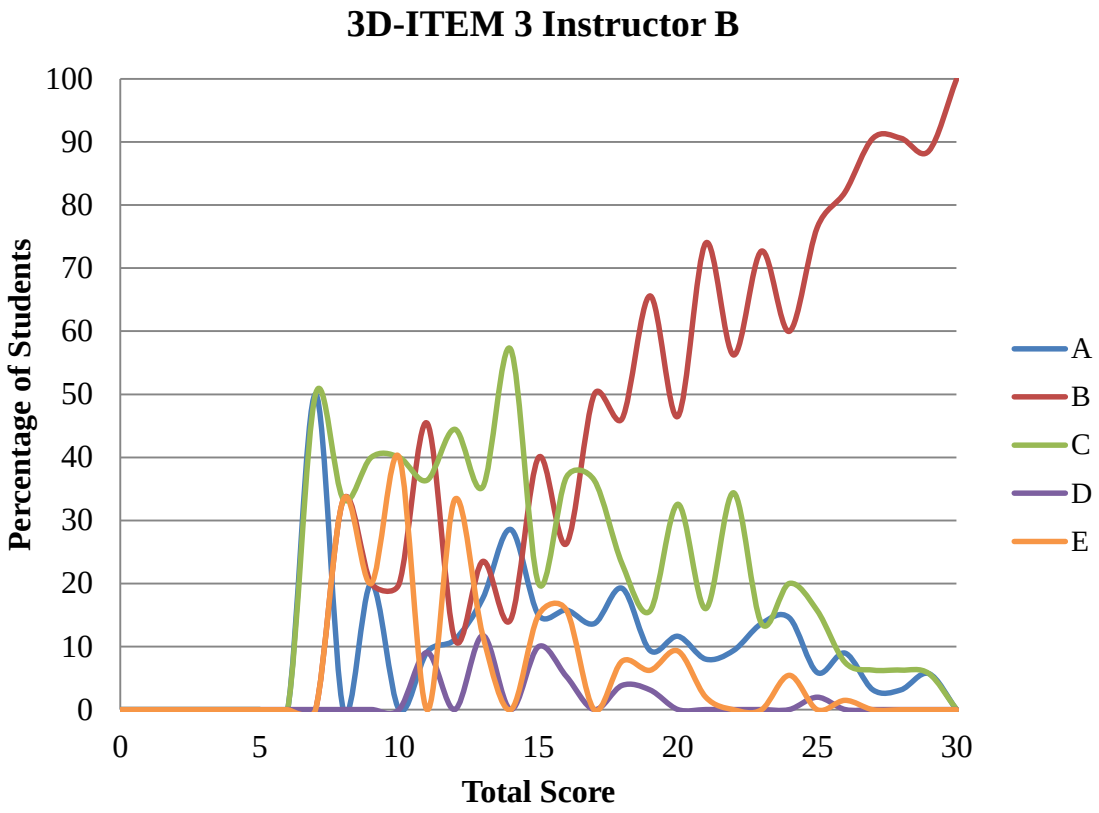


Figure 6. IRC for 3D-ITEM 3 on Exam 2 from Instructor B (*N* = 651).

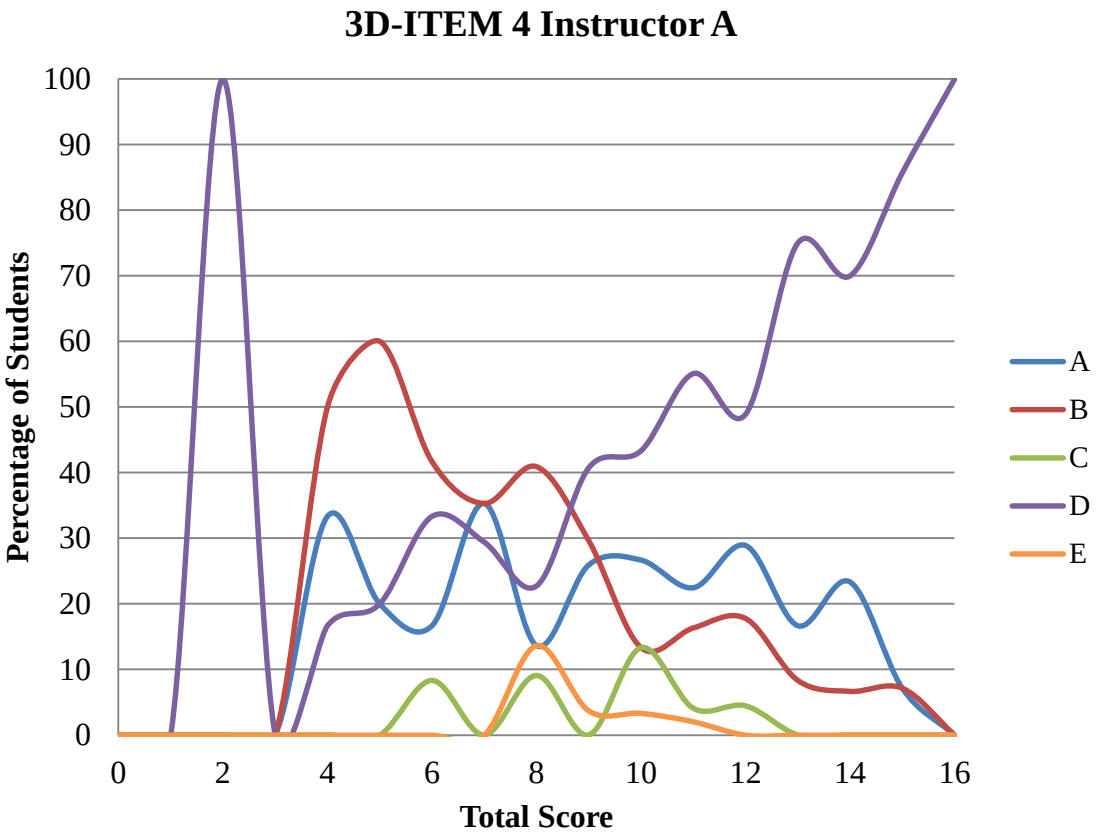


Figure 7. IRC for 3D-ITEM 4 on Exam 3 from Instructor A (*N* = 301).

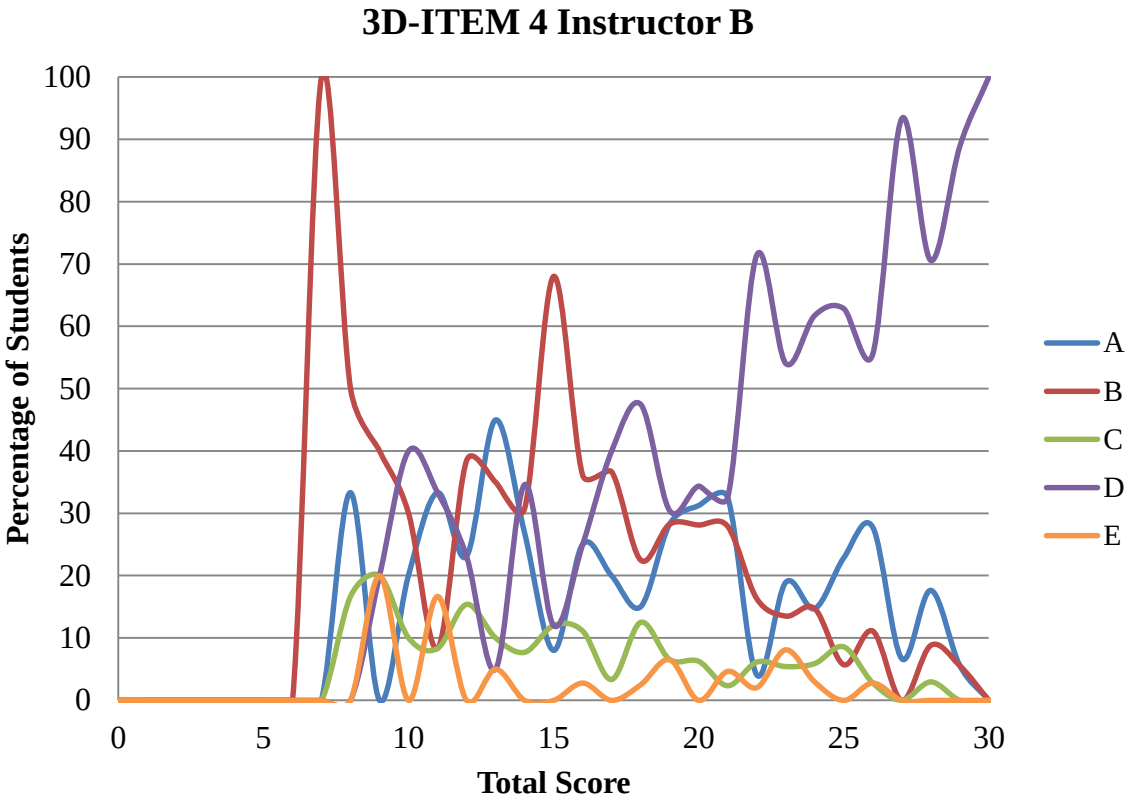


Figure 8. IRC for 3D-ITEM 4 on Exam 3 from Instructor B (*N* = 617).

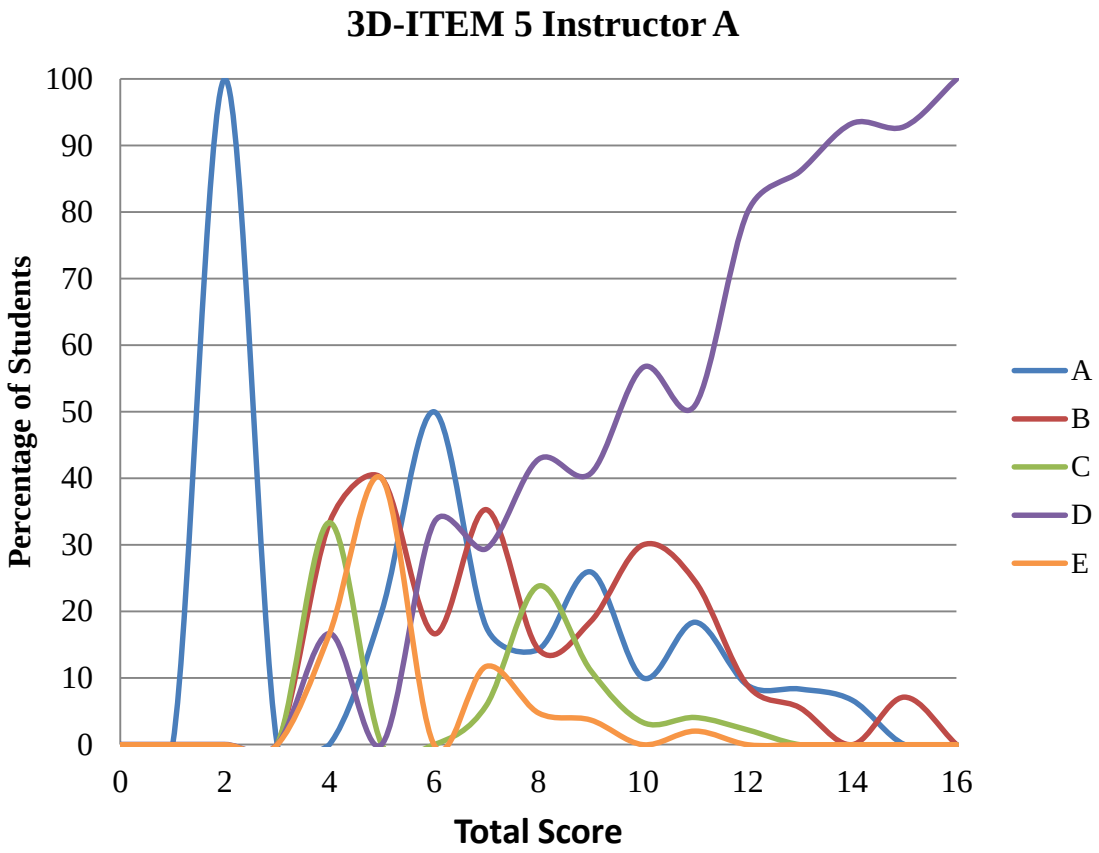


Figure 9. IRC for 3D-ITEM 5 on Exam 3 from Instructor A (*N* = 300).

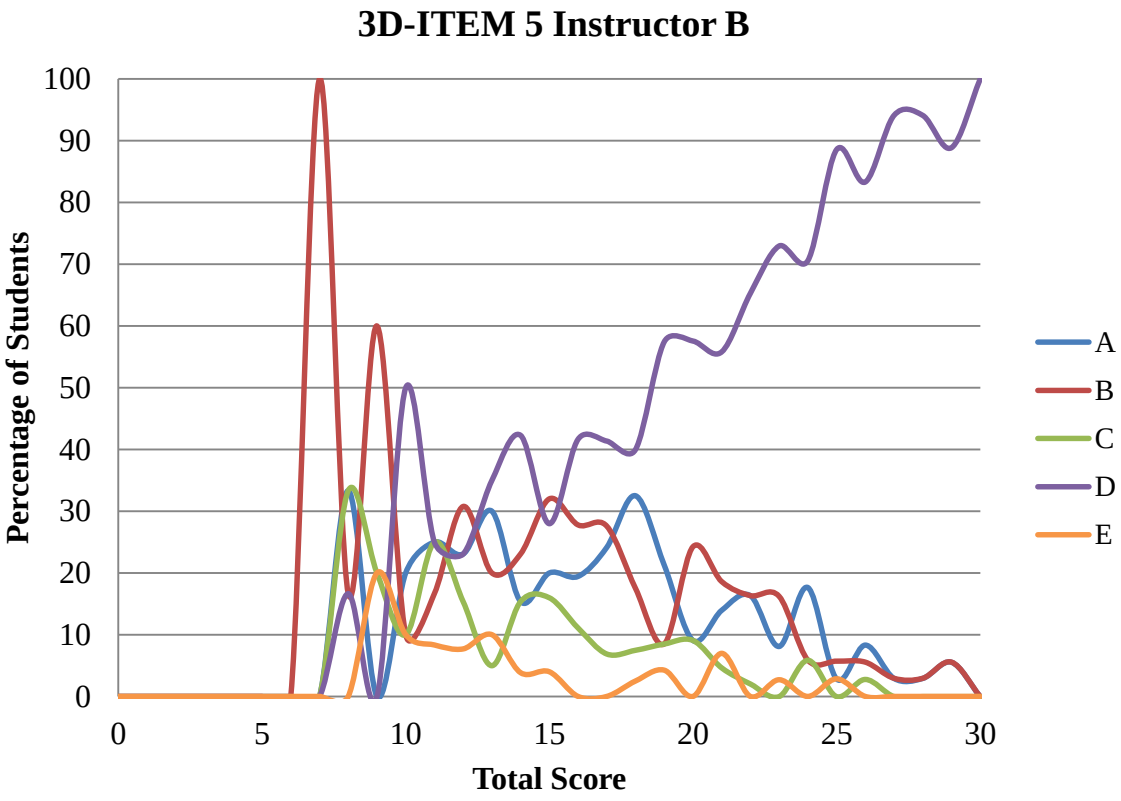


Figure 10. IRC for 3D-ITEM 5 on Exam 3 from Instructor B (*N* = 637).

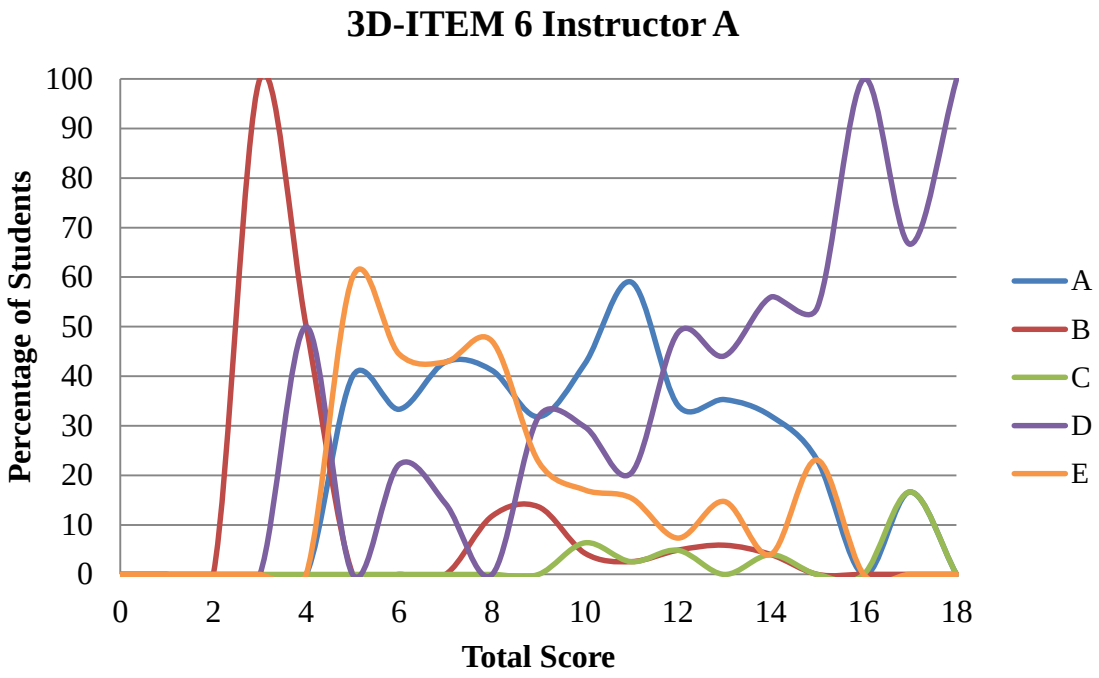


Figure 11. IRC for 3D-ITEM 6 on Exam 4 from Instructor A (*N* = 282).

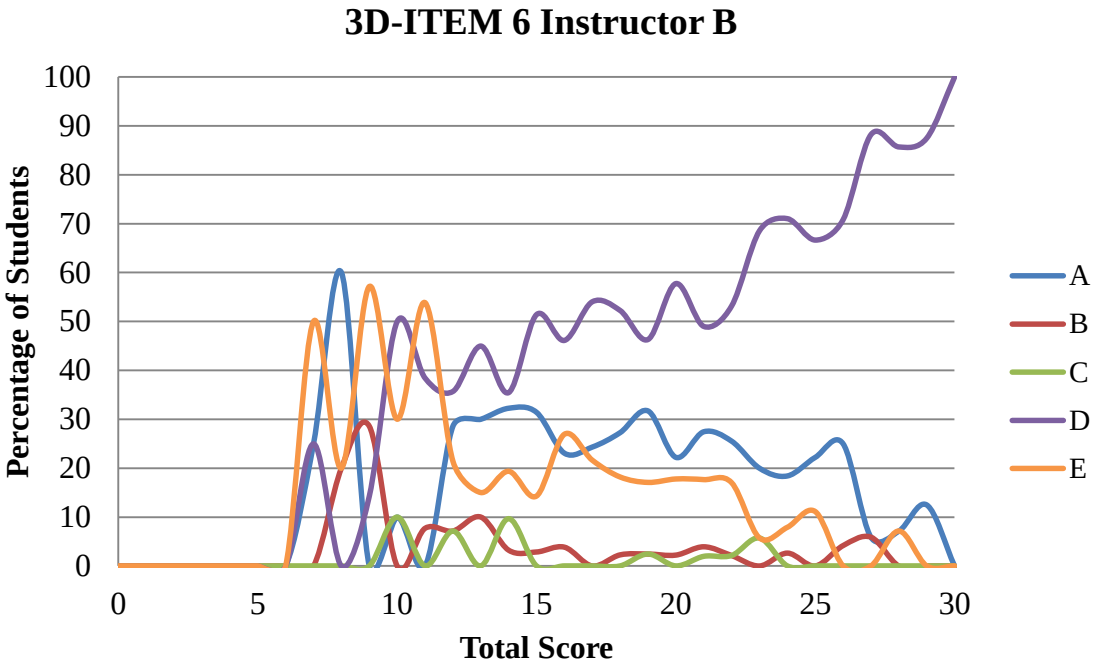


Figure 12. IRC for 3D-ITEM 6 on Exam 4 from Instructor B (*N* = 610).

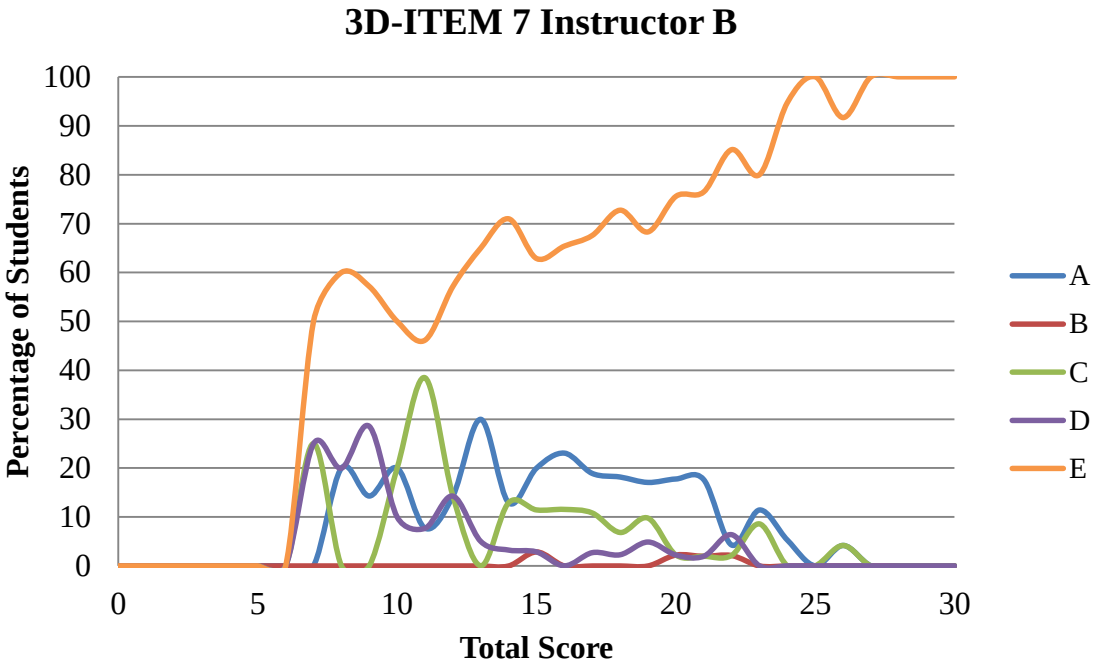


Figure 13. IRC for 3D-ITEM 7 on Exam 4 from Instructor B (*N* = 607).

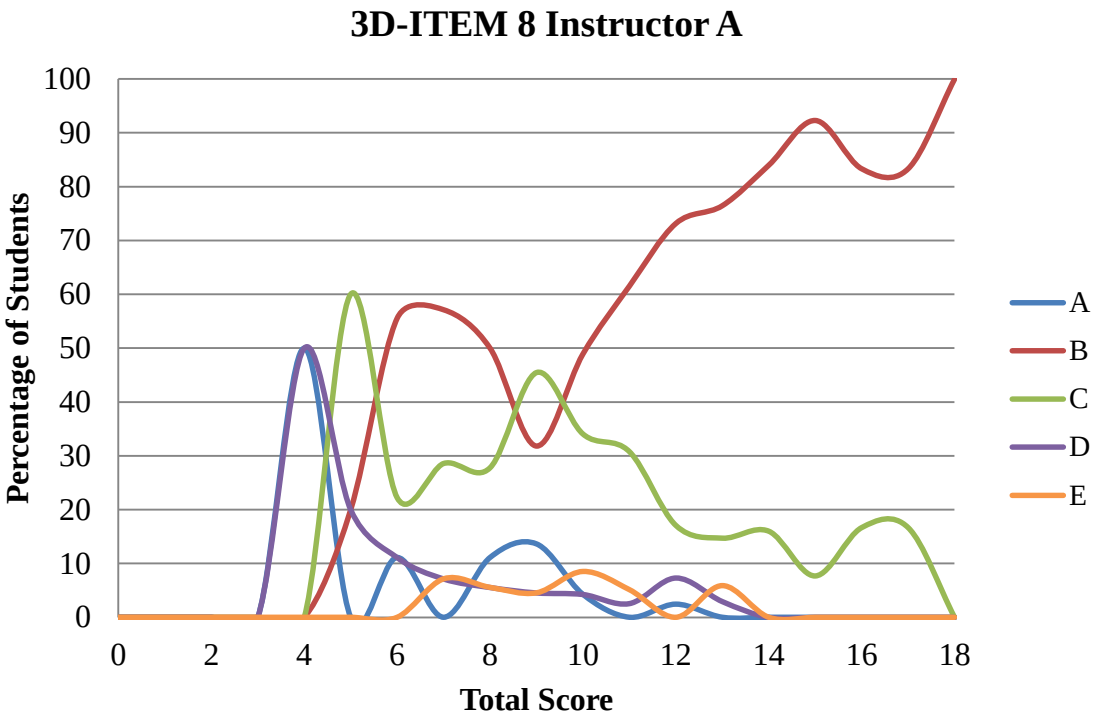


Figure 14. IRC for 3D-ITEM 8 on Exam 4 from Instructor A (*N* = 283).

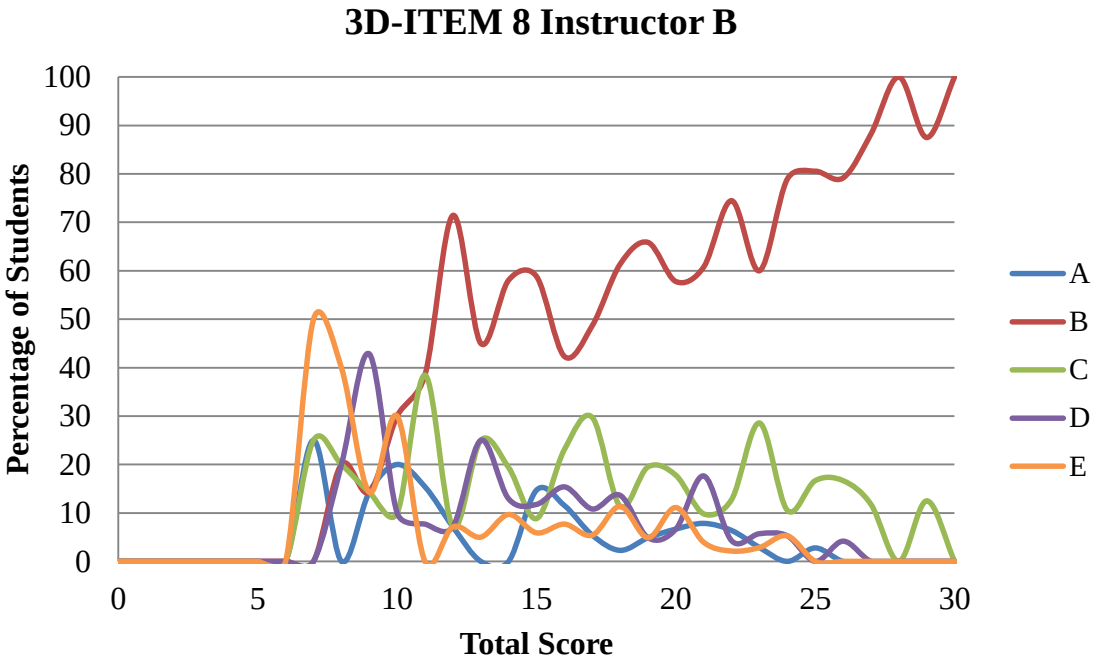


Figure 15. IRC for 3D-ITEM 8 on Exam 4 from Instructor B (*N* = 609).

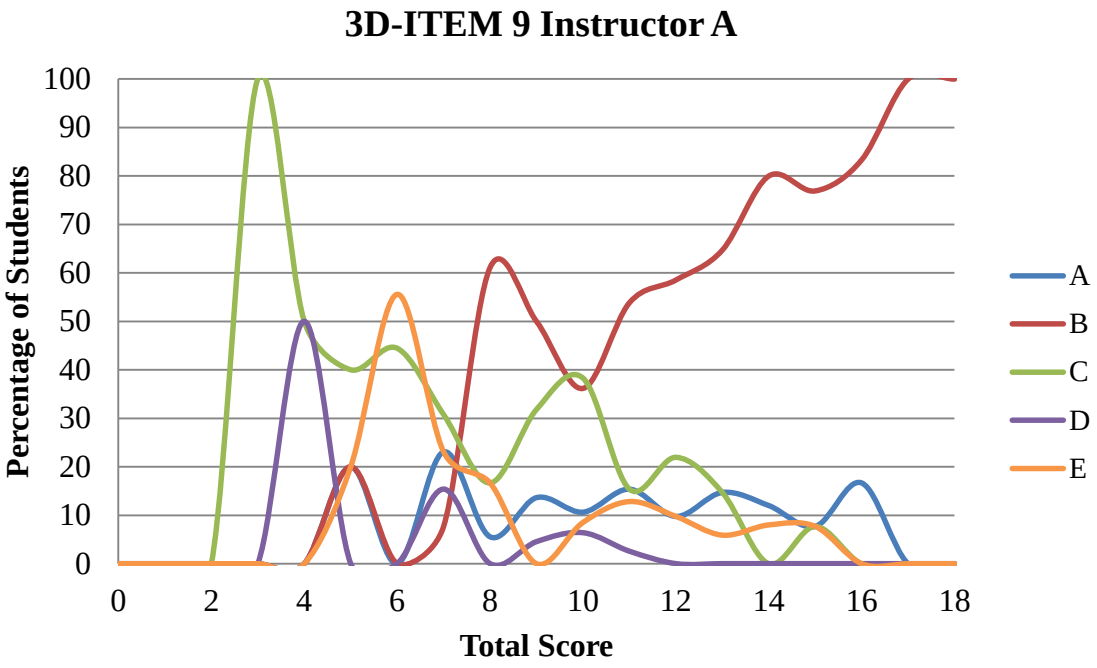


Figure 16. IRC for 3D-ITEM 9 on Exam 4 from Instructor A (N = 282).

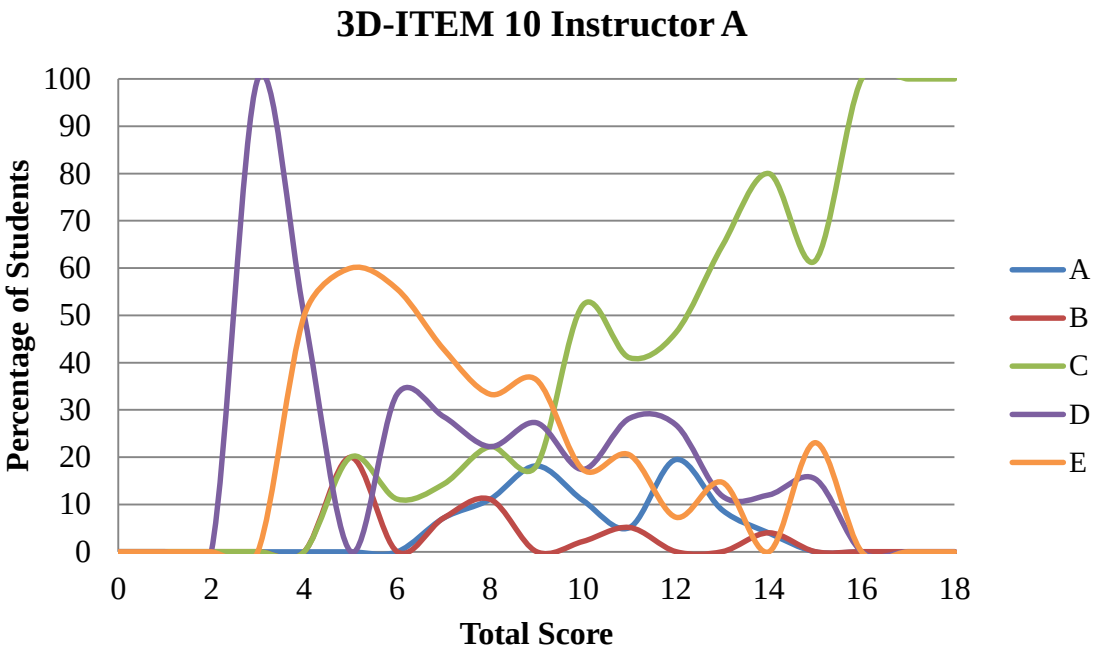


Figure 17. IRC for 3D-ITEM 10 on Exam 4 from Instructor A (*N* = 282).

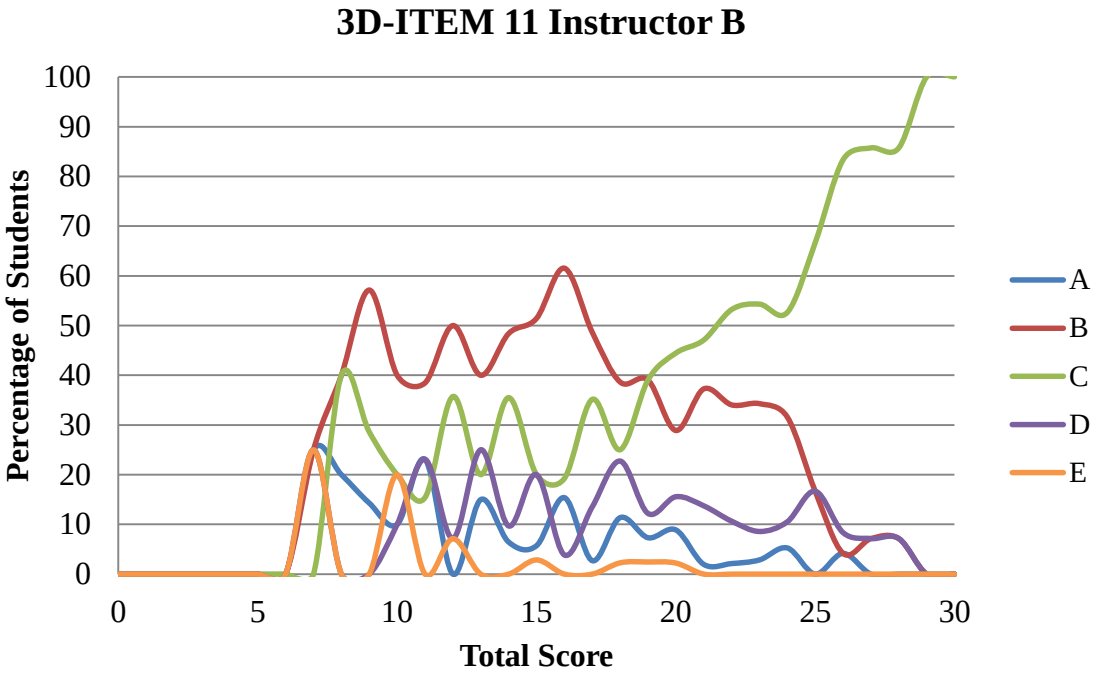


Figure 18. IRC for 3D-ITEM 11 on Exam 4 from Instructor B (*N* = 607).

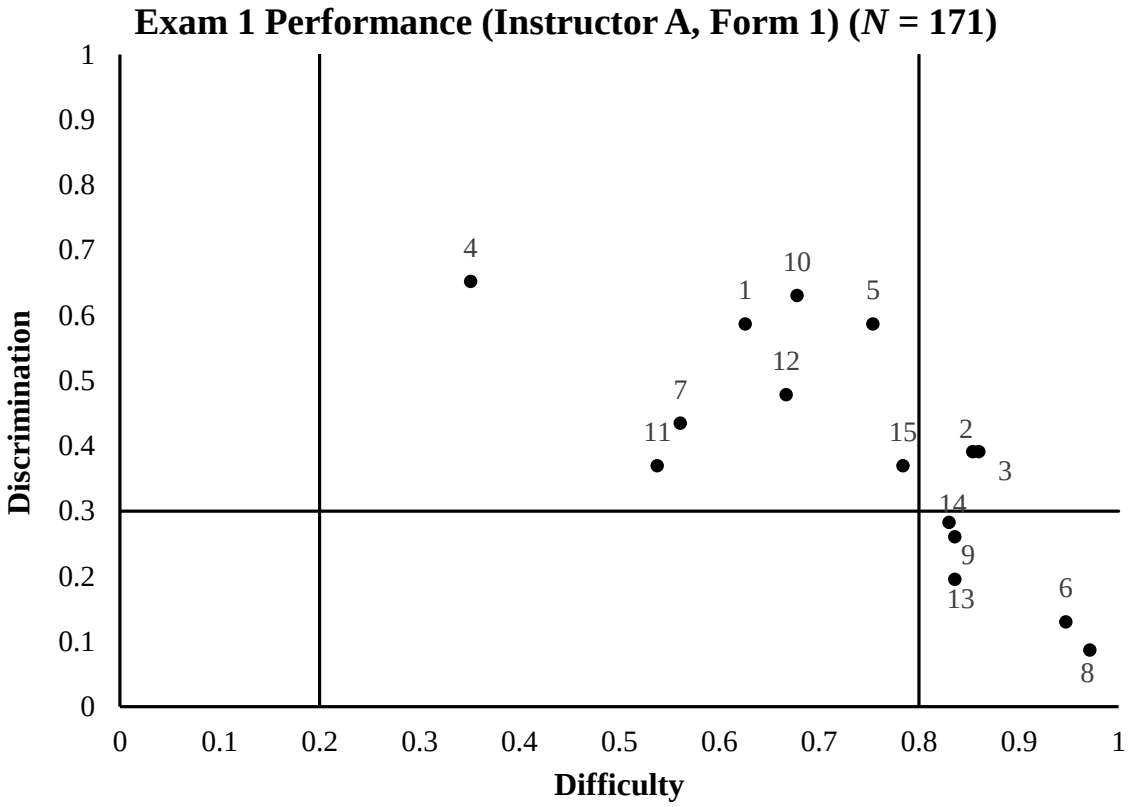


Figure 19. Exam 1 performance on Form 1 given by Instructor A. Item 15 was created for research purposes to incorporate a science practice. Item 11 was created by the instructor and also includes a science practice.

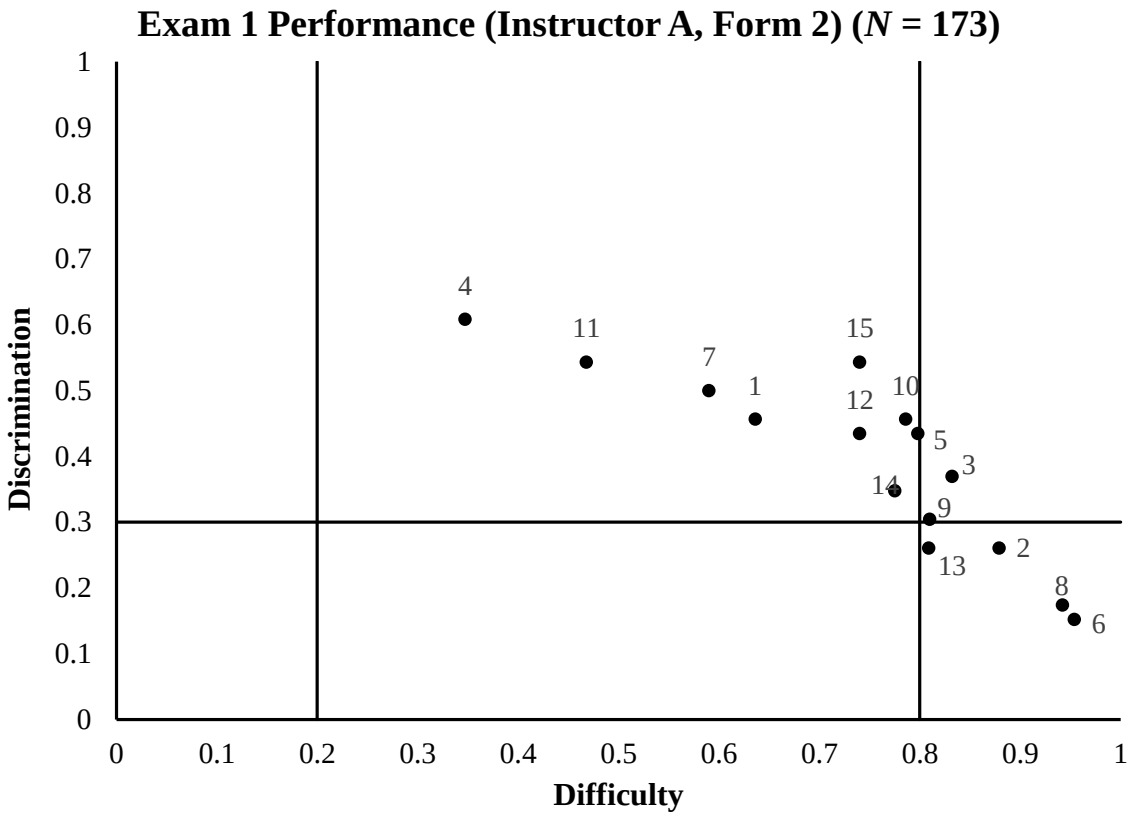


Figure 20. Exam 1 performance on Form 2 given by Instructor A. Item numbering has been aligned to match that of Form 1 for ease of comparison. Item 15 was created for research purposes to incorporate a science practice. Item 11 was created by the instructor and also includes a science practice.

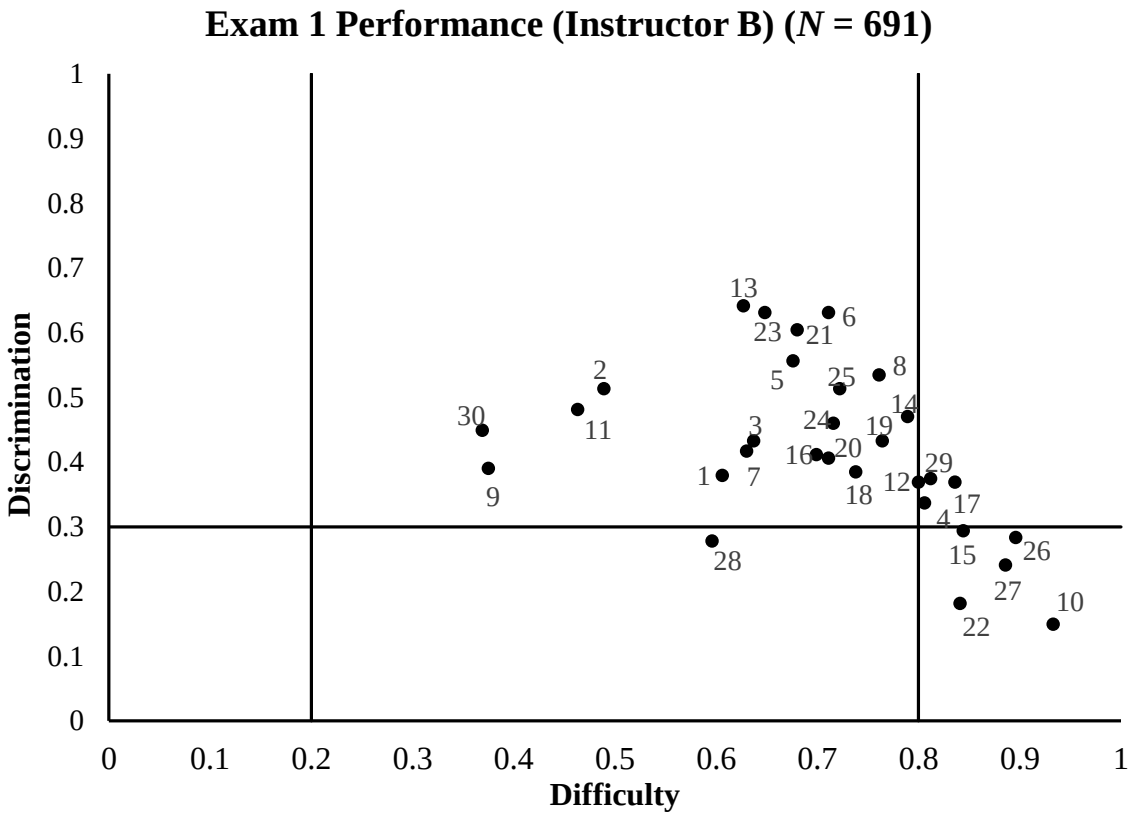


Figure 21. Exam 1 performance for Instructor B. Only one form was given. Item 24 was created for research purposes to incorporate a science practice. Items 9 and 25 were created by the instructor and also include a science practice.

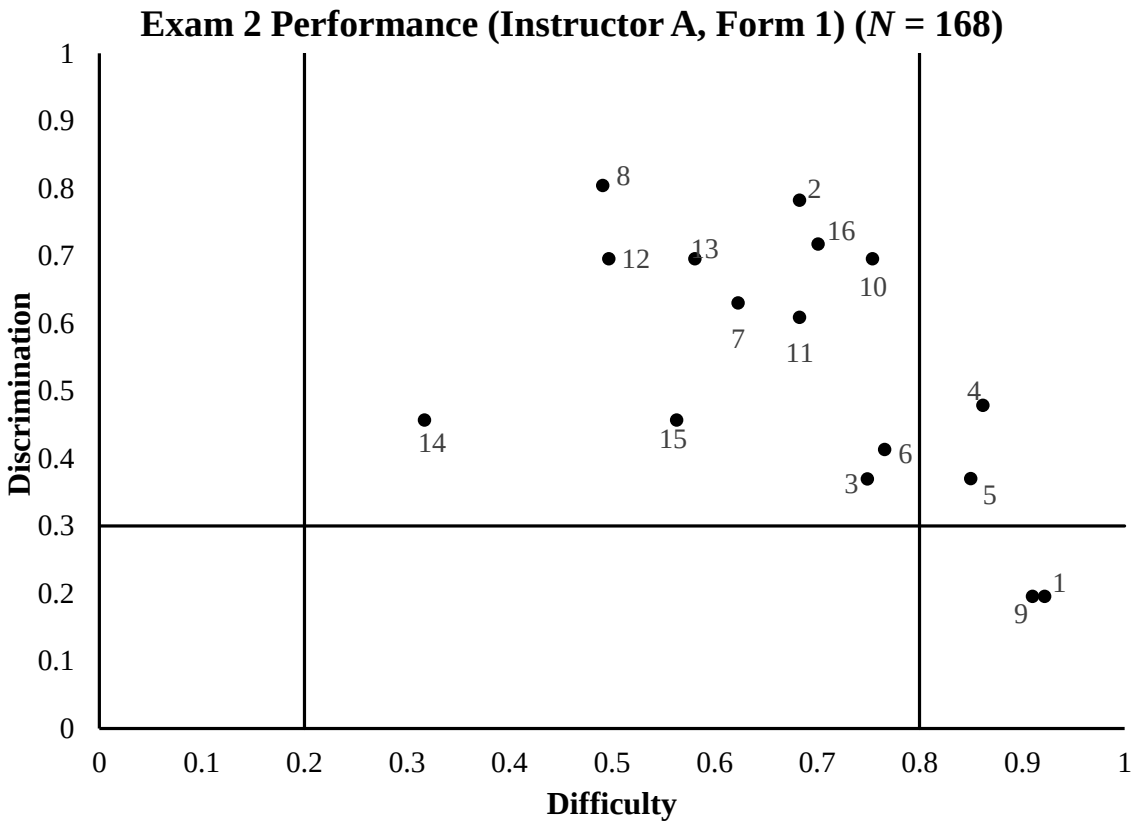


Figure 22. Exam 2 performance on Form 1 given by Instructor A. Items 14 and 15 were created for research purposes to incorporate a science practice. Items 11 and 13 were created by the instructor and also include a science practice.

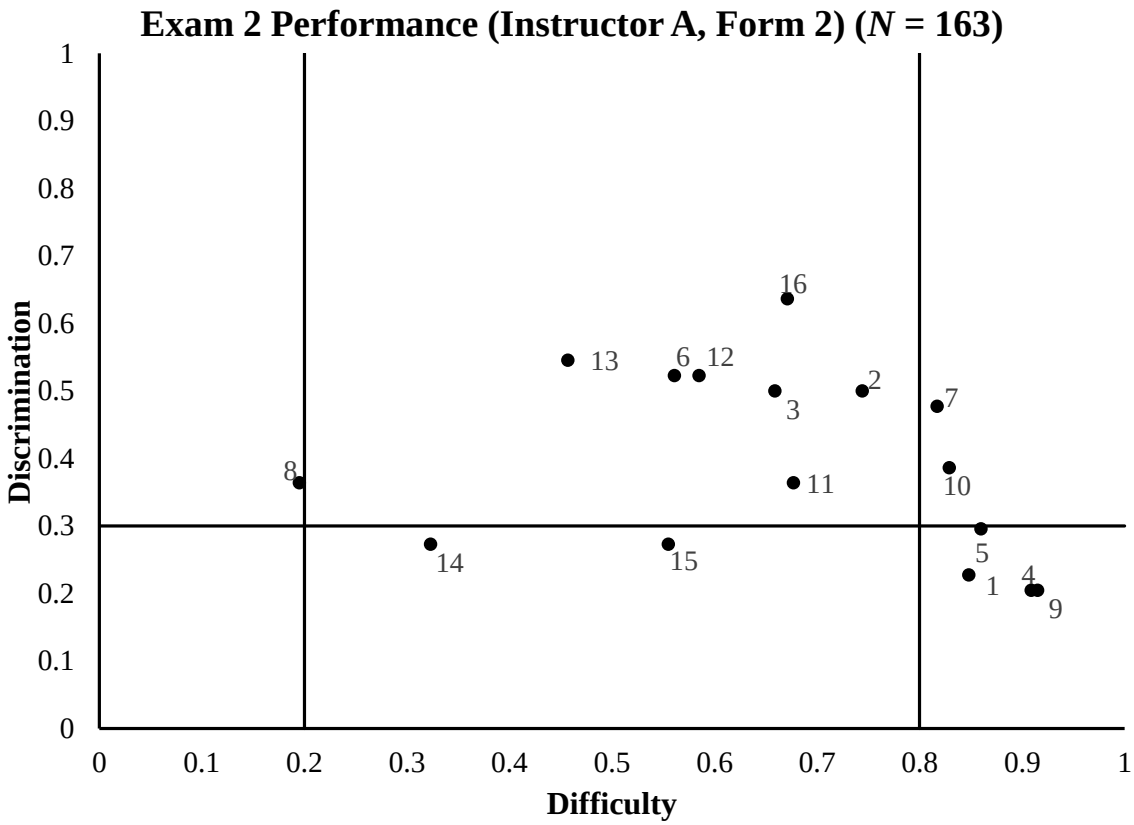


Figure 23. Exam 2 performance on Form 2 given by Instructor A. Item numbering has been aligned to match Form 1 for ease of comparison. Items 14 and 15 were created for research purposes to incorporate a science practice. Items 11 and 13 were created by the instructor and also include a science practice.

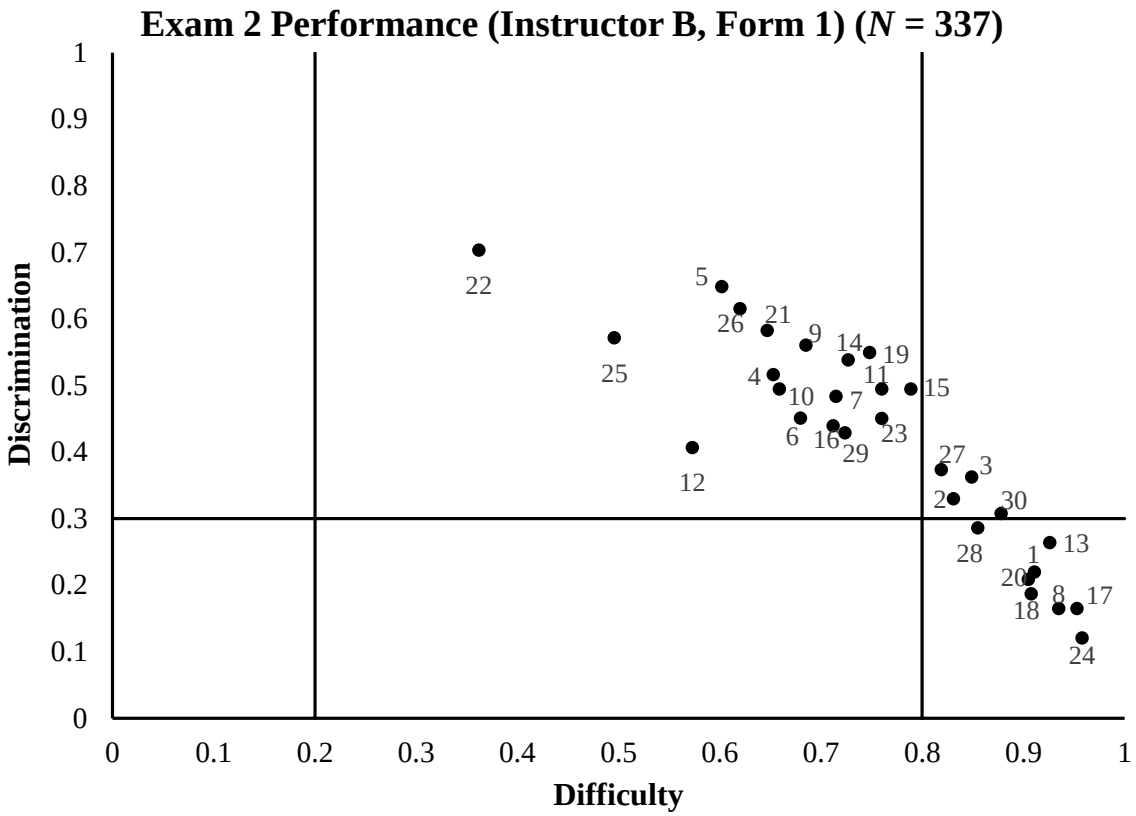


Figure 24. Exam 2 performance on Form 1 given by Instructor B. Items 5 and 6 were created for research purposes to incorporate a science practice. Items 18, 25, 28 and 29 were created by the instructor and also include a science practice.

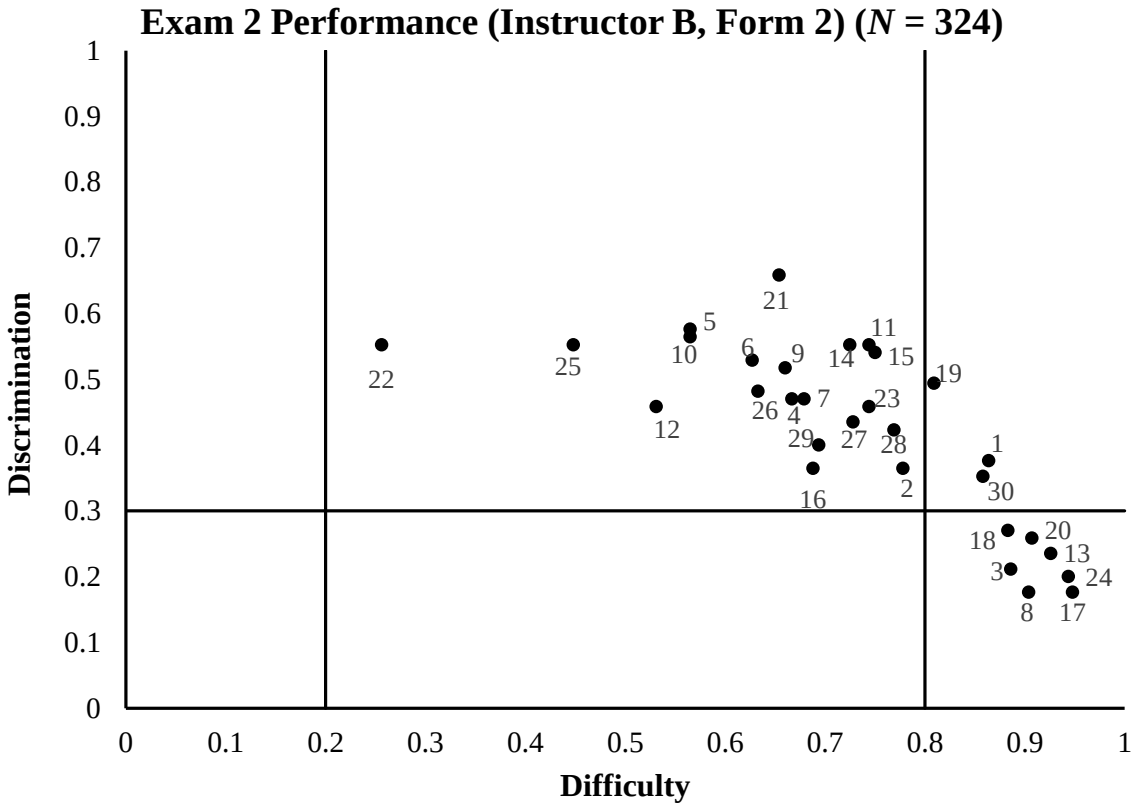


Figure 25. Exam 2 performance on Form 2 given by Instructor B. Item numbering has been aligned to match Form 1 for ease of comparison. Items 5 and 6 were created for research purposes to incorporate a science practice. Items 18, 25, 28 and 29 were created by the instructor and also include a science practice.

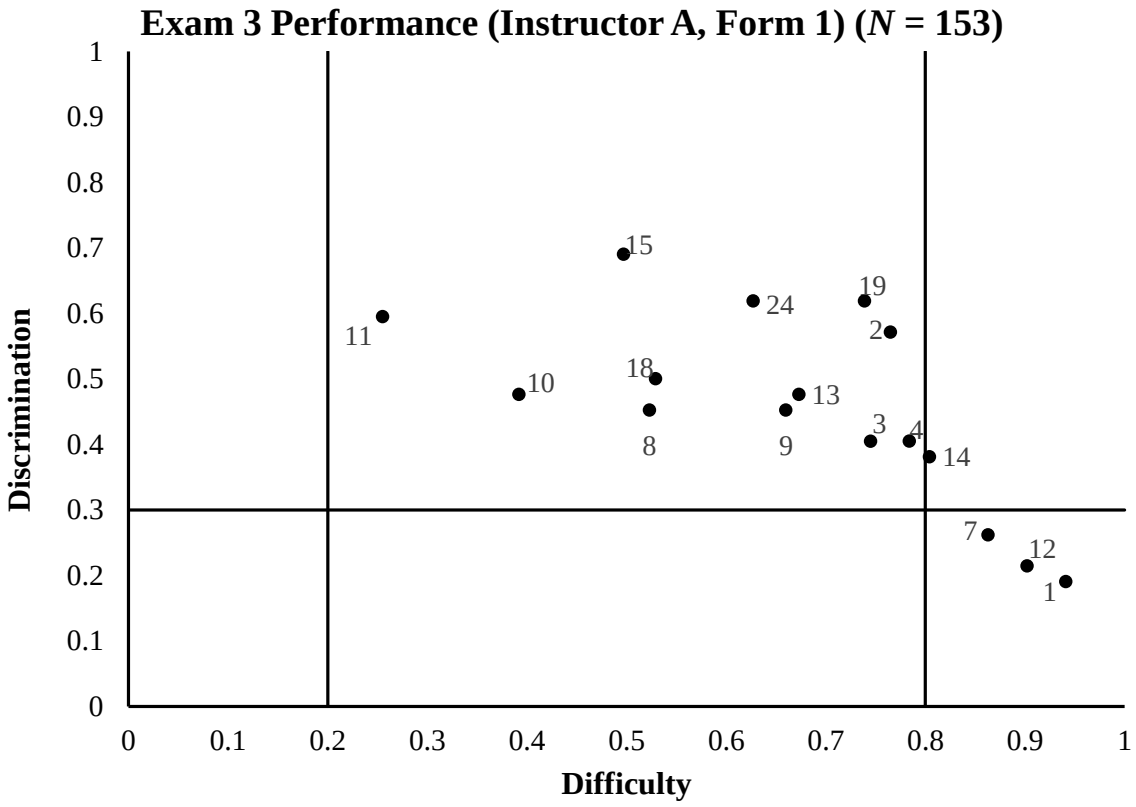


Figure 26. Exam 3 performance on Form 1 given by Instructor A. Items 15 and 16 were created for research purposes to incorporate a science practice. Item 9 was created by the instructor and also includes a science practice. Forms 1 and 2 of Exam 3 written by Instructor A vary only in response choice order.

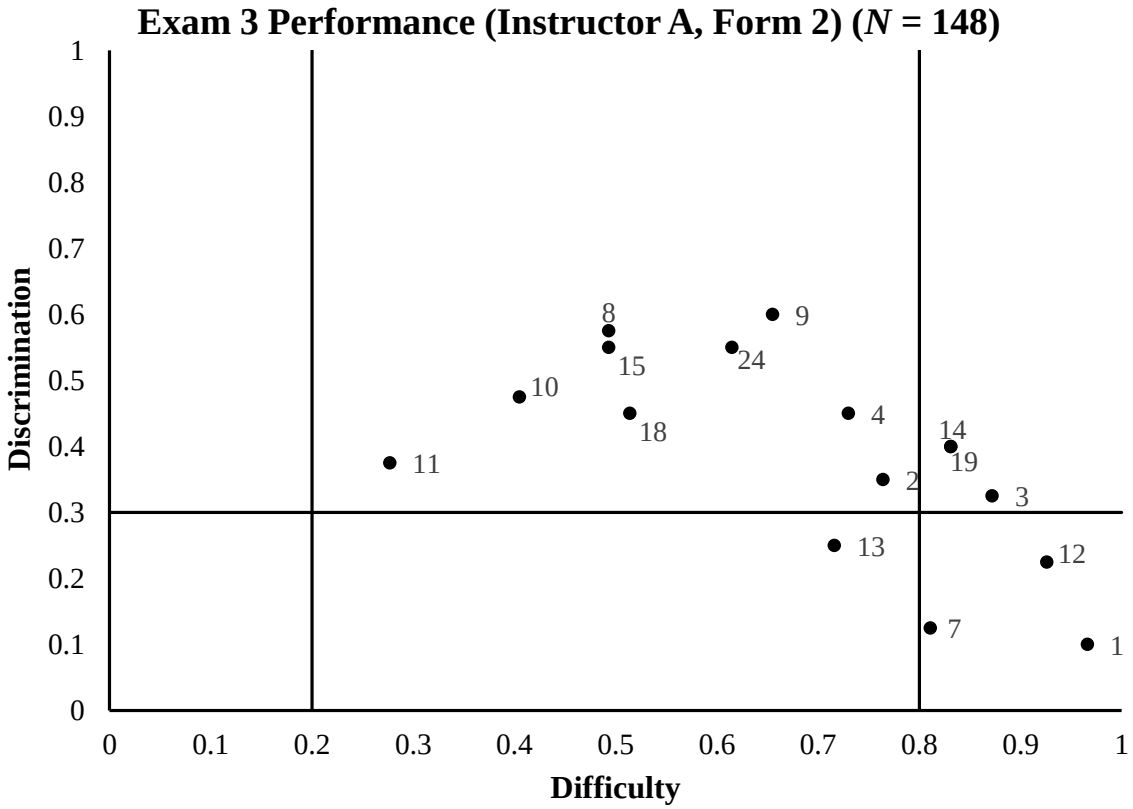


Figure 27. Exam 3 performance on Form 2 given by Instructor A. Items 15 and 16 were created for research purposes to incorporate a science practice. Item 9 was created by the instructor and also includes a science practice. Forms 1 and 2 of Exam 3 written by Instructor A vary only in response choice order.

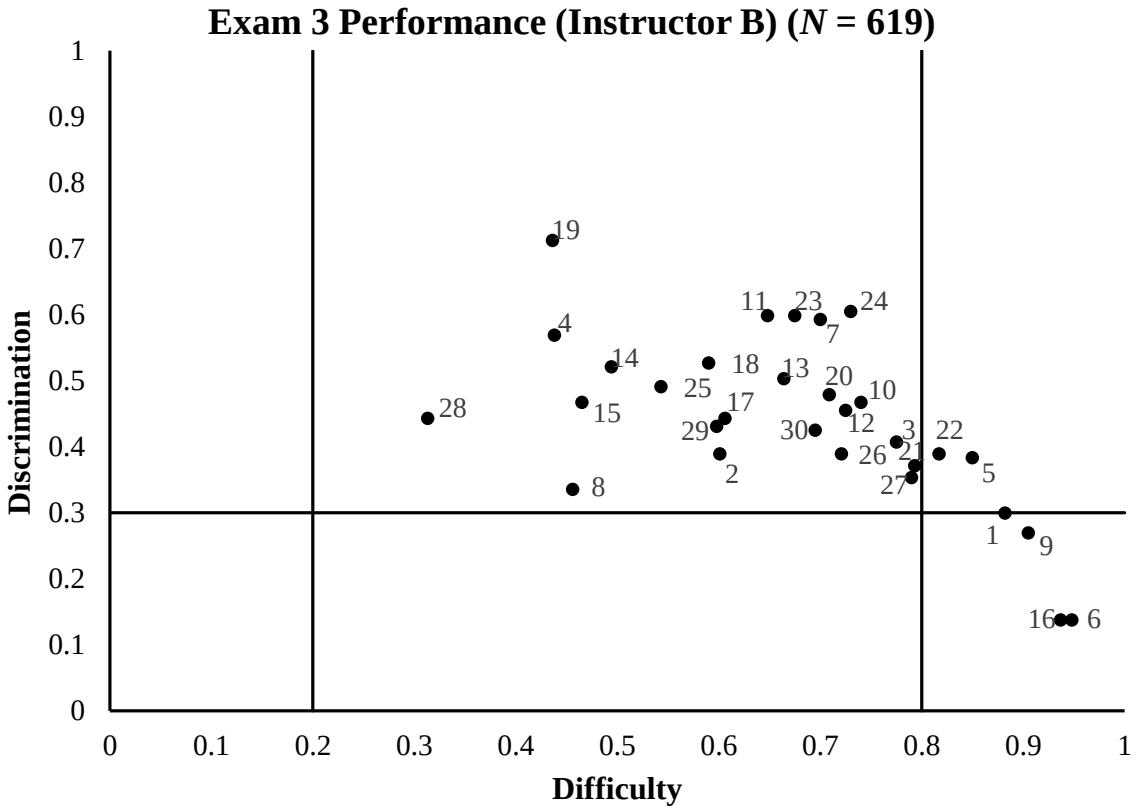


Figure 28. Exam 3 performance for Instructor B. Only one form was used. Items 14 and 15 were created for research purposes to incorporate a science practice. Items 10, 22, 23, and 27 were created by the instructor and also include a science practice.

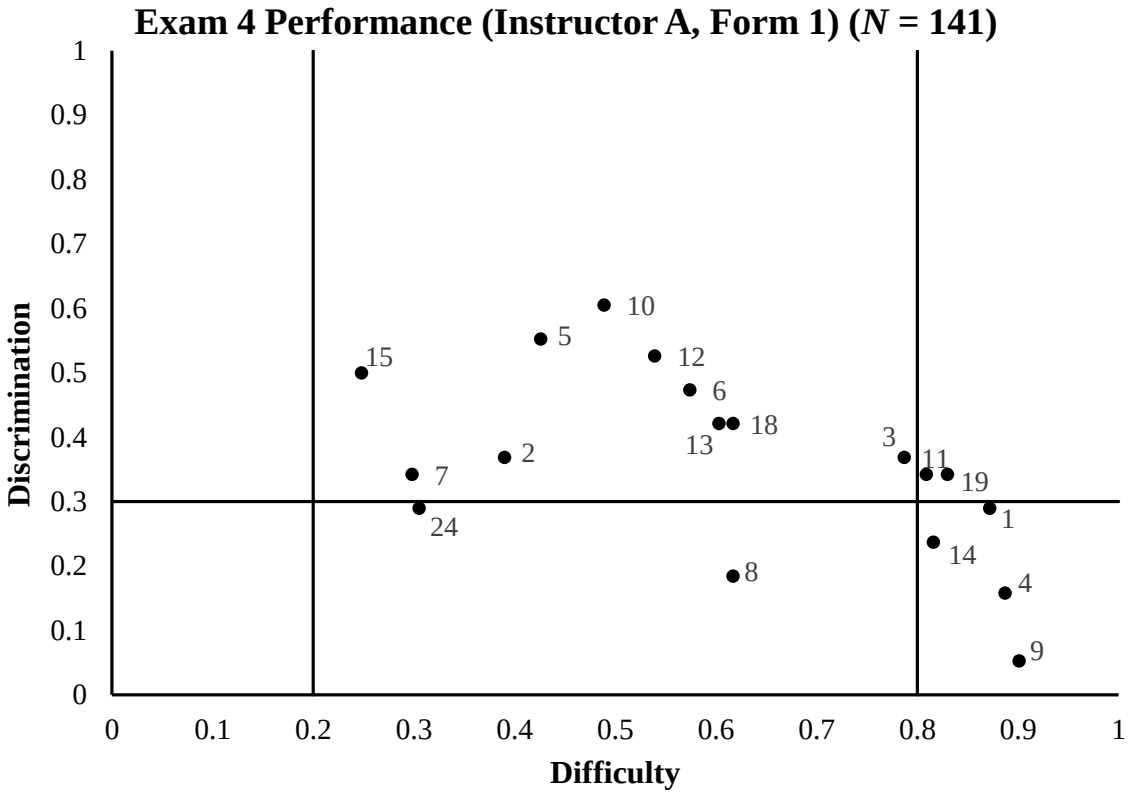


Figure 29. Exam 4 performance on Form 1 given by Instructor A. Items 9, 10, 11, and 17 were created for research purposes to incorporate a science practice. Item 16 was created by the instructor and also includes a science practice.

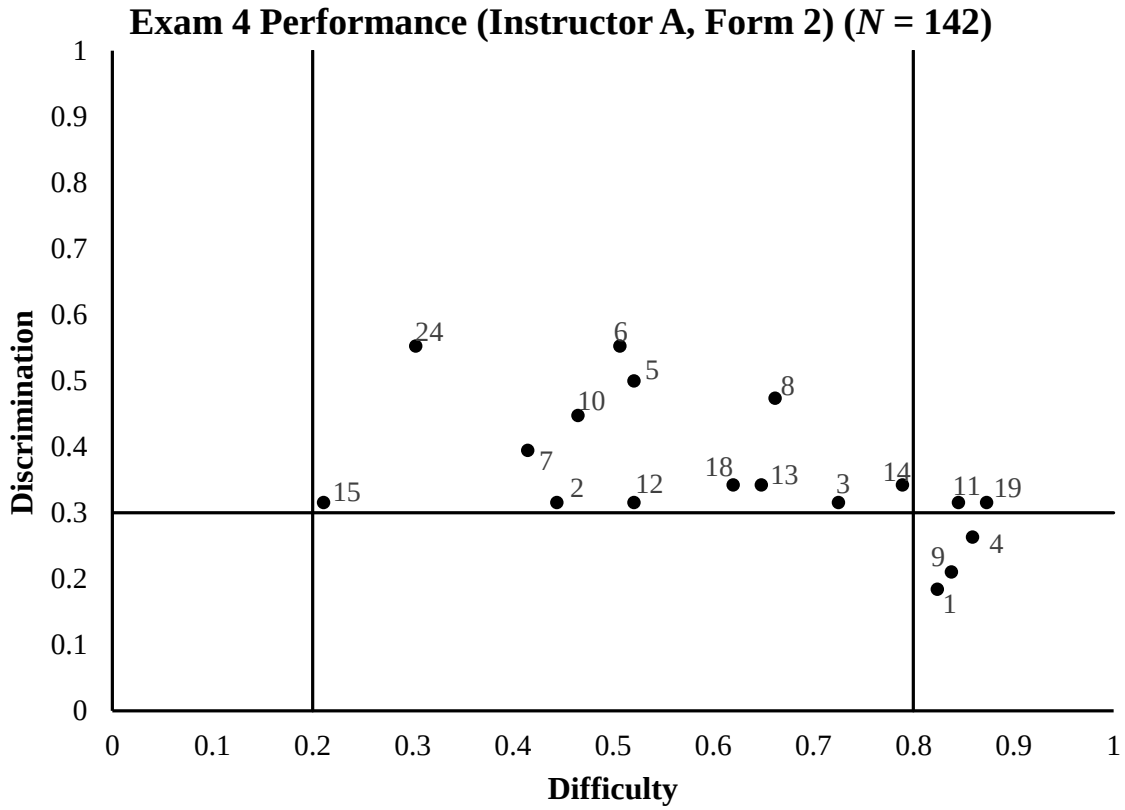


Figure 30. Exam 4 performance on Form 2 given by Instructor A. Item numbering has been aligned to Form 1 for ease of comparison. Items 9, 10, 11, and 17 were created for research purposes to incorporate a science practice. Item 16 was created by the instructor and also includes a science practice.

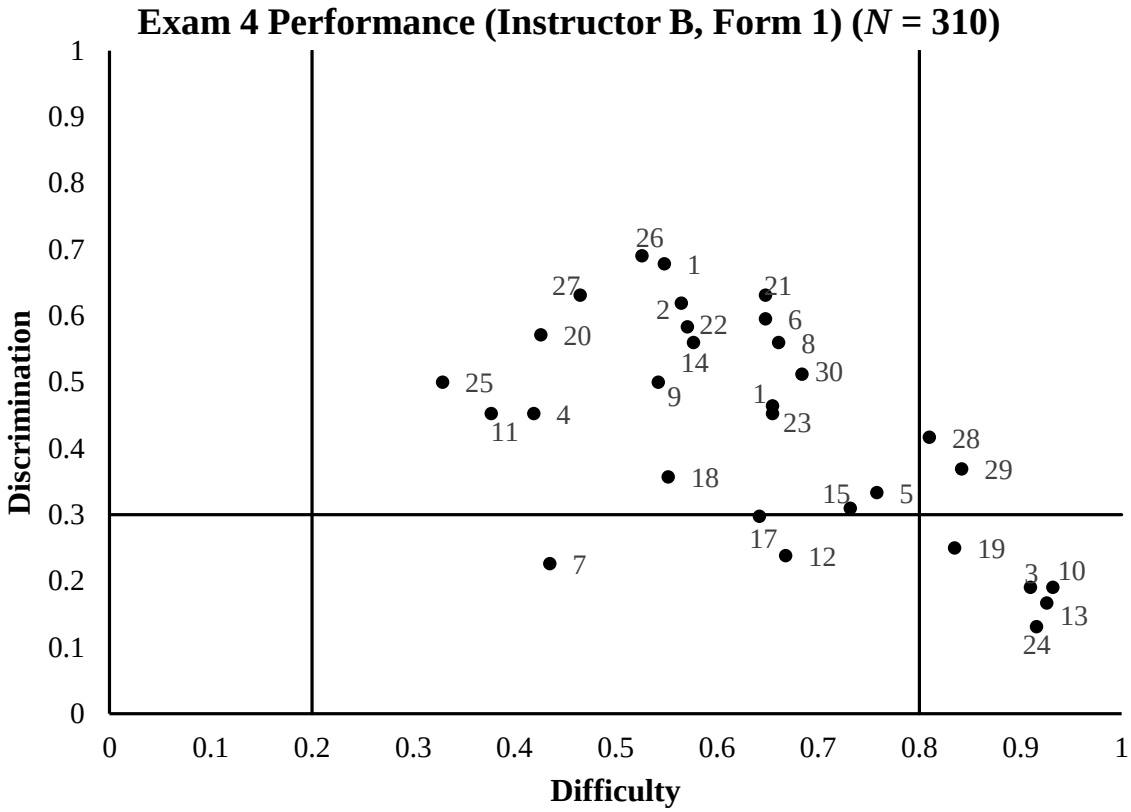


Figure 31. Exam 4 performance on Form 1 given by Instructor B. Items 7, 15, 17, and 27 were created for research purposes to incorporate a science practice. Items 10, 11, 13, 14, and 19 were created by the instructor and also include a science practice.

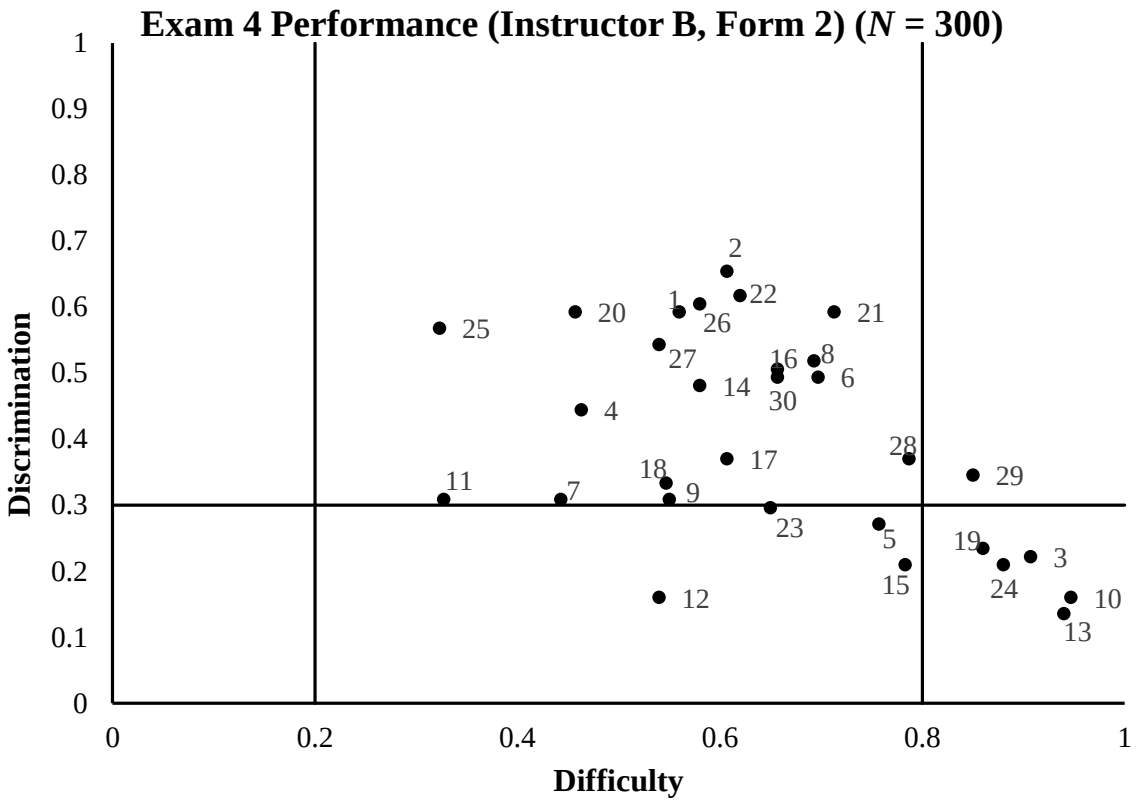


Figure 32. Exam 4 performance on Form 2 given by Instructor B. Item numbering has been aligned to Form 1 for ease of comparison. Items 7, 15, 17, and 27 were created for research purposes to incorporate a science practice. Items 10, 11, 13, 14, and 19 were created by the instructor and also include a science practice.

CHAPTER 6: CONCLUDING REMARKS

Summary of Research Findings

As demonstrated in the previous chapters, the development of science practices within chemistry courses is not only valued by instructors, but inherently embedded within the content assessed in chemistry examinations. Yet, little has been done to explicitly express the value of such practices to students or make them explicit components of course instruction. The work herein aims to provide a foundation for future studies of science practices by investigating the current status of incorporation of these practices in chemistry assessments at the college level.

Qualitative interviews with general chemistry instructors revealed that they value a variety of goals and skills in their courses, particularly goals related to life skills, laboratory skills, and appreciation of chemistry. These goals were found to be in accordance with other published goals of science education (Duschl, 2008; Hodson, 2003; Longbottom & Butler, 1999; Norris, 1997). Additional evidence of the value placed upon content independent skills was supported through the results of a national survey which contained questions about how often instructors incorporate various goals and skills into classroom instruction, and how they assess those goals. From these results, it was apparent that the development of goals and skills beyond content proficiency is of value to the chemistry community, yet instructors did little to assess these skills beyond traditional forms of assessment such as exams, quizzes, and laboratory reports. This was consistent with previous studies of faculty awareness of assessment (Emenike, Raker, & Holme, 2013; Raker, Emenike, & Holme, 2013; Towns, 2009), and provided a glimpse of the tension that often arises between the desire to assess non-content goals and the

necessity to assess content knowledge. These results suggested that assessments of the future will need to consider combining measures of content and skill simultaneously. In order to understand how to move forward with design of new assessment materials, a framework to support the design of assessment materials that incorporate measures beyond content was necessary. The report entitled *A Framework for K-12 Science Education* (National Research Council, 2012), and the Next Generation Science Standards (NGSS) (Achieve, 2013)derived from it, served as a guiding framework due to the nature in which content, practices, and crosscutting concepts were intertwined. This framework provided a lens for analysis called three-dimensional learning. In order to analyze assessment materials for the presence of science practices, a rubric based upon the NGSS was refined and used. The Three-dimensional Learning Assessment Protocol (3D-LAP) (Cooper, 2014; Underwood, et al., 2014) was used as a starting point to analyze a variety of chemistry assessments from the American Chemical Society Examinations Institute for incorporation of science practices. Fewer than half of the items analyzed contained a science practice, and while this may seem modest, since the items were not intended to explicitly incorporate practices these data suggest that science practices are inherently embedded within chemistry content and are of value to be assessed. Additionally, the 3D-LAP rubric was used to create items that incorporated science practices to be included in general chemistry course exams. Items created to incorporate science practices were significantly more difficult for students, implying that instruction focused on the development of science practices is likely necessary to close the gap. Even though student performance on items with science practices was lower than on items without, the items with science practices still had acceptable psychometric

properties, suggesting that items designed to intentionally incorporate science practices with content are viable options for future assessment developments.

In short, there is a desire to measure goals and skills beyond content in chemistry courses, yet the community tends to settle for measures strictly of content for a variety of reasons, primarily that measuring such skills is often not readily feasible in a time-constrained curriculum. The research herein suggests that measures of skills beyond content knowledge proficiency are possible with the use new and emerging tools. In this sense, practitioners and chemistry education researchers can get more out of the measurements they make if they are willing to invest the time and effort to engage with these new tools.

Implications for Assessment

The items analyzed which contained science practices had psychometric data to support their use as valid assessment items. Future assessment developers likely need not be concerned that incorporation of science practices into multiple-choice items will inherently have a negative effect on student performance or its measurement. This is also not to suggest that every item constructed needs to have an associated science practice. Some areas of chemistry content do not readily mesh with any of the eight science practices, and are more suited toward rote learning as it relates to declarative knowledge of facts. While there is a place for rote learning within chemistry, or any field, current calls for reform aim to shift the focus away from measures of discrete facts to measures that provide evidence of what students can do with the knowledge. In practice, tests that intentionally incorporate science practices will likely be a combination of items with and without science practices.

Future Work

While there is evidence to suggest that test items can be constructed to measure content and science practices together, there is also an apparent gap between student performance on items with and without science practices. To understand this gap, additional research on the effects of instruction is necessary. Assessments that incorporate science practices ought to be aligned with instruction that values and emphasizes such practices in order for student learning to improve (Pellegrino, et al., 2014; Stiggins, 1994). Additional evidence about the link between explicit instruction and performance on assessment items incorporating science practices is necessary to encourage practitioners to support such endeavors. Currently, there are few redesigned curricula to support three-dimensional learning in the post-secondary chemistry classroom (Cooper & Klymkowsky, 2013; Talanquer & Pollard, 2010). As more curricular revisions and transformations that include a more holistic approach to chemistry learning occur, assessments to measure skills beyond content knowledge will become more mainstream. In this sense, it is up to the chemistry education research community to model how such assessments should be constructed and provide evidence of their efficacy.

As the community progresses in these endeavors, this project lays the foundation for future studies, such as the following examples:

1) Comparison of Multiple-Choice and Free-Response Assessment Items

Additional future work should examine how student performance on multiple-choice items that incorporate science practices correlates with performance on free-

response items which incorporate similar content and practices. Ideally, multiple-choice and free-response items to measure the same science practice and similar content would be incorporated onto the same exam, and results cross-validated. If this level of homogeneity is not possible, at the very least the stakes for student performance must be carefully controlled in studies about the outcomes of different item formats for science practices. Students' responses to the free-response item would need to be qualitatively coded and compared to patterns related to overall exam performance and performance on multiple-choice items measuring the same science practice. Student responses to free-response questions could be used to identify misconceptions for use in the subsequent creation of distractors for multiple-choice items, as part of a larger program of test development. By identifying whether students are able to perform similarly on items that only vary in their format, the evidence that adjudicates the ability to measure science practices with multiple-choice items is obtained.

2) Qualitative Evaluation of Students' Use of Science Practices

Qualitative interviews with students about how they are thinking and interacting with item content while answering items with and without science practices would provide additional evidence to validate both the 3D-LAP rubric, and the ability to assess the development of science practices within traditional formats of assessments such as pencil and paper exams. Qualitative interviews would confirm whether or not the assessment is functioning as the researchers intend it to, and would likely provide evidence to suggest revisions of the assessment. Ideally, the interviews would help to identify how students engage in each of the eight science practices in high stakes assessment situations. The sample could potentially consist of students from only general

chemistry courses, or students from each level of the undergraduate chemistry curriculum. The information gained from a study of this nature would aid the design of future assessments by substantiating how students actually engage in science practices compared to how practitioners and researchers expect them to engage.

3) Longitudinal Studies of Students' Development of Science Practices

In order to better understand how students develop science practices, research on the longitudinal development of these practices is necessary. Time-scales to be considered for these studies are those related to development of practices within a single chemistry course versus development of practices across the undergraduate chemistry curriculum. Longitudinal studies of this nature would be no small undertaking, and thus, would likely require the work of a team of researchers across several years. The effects of the curriculum and instructor would need to be considered and accounted for in addition to changes affecting the incoming college student population as K-12 curricula that emphasize science practices become more common. Nonetheless, these types of studies should be considered in order to measure the efficacy of instruction and assessment of science practices, and three-dimensional learning in total. In this sense, longitudinal studies could be used to establish the measurement of science practices and crosscutting concepts independent of particular content. Longitudinal evaluations of students' development of science practices could serve a variety of purposes such as evaluating the efficacy of new curricula that support three-dimensional learning in chemistry, determining how the undergraduate chemistry curriculum fosters the growth of science practices, and identifying whether the incorporation of specific science practices marginalizes any sub-population of the course.

Assessment across the undergraduate chemistry curriculum is available with ACS examinations. The use of the 3D-LAP rubric to evaluate additional ACS exams across the undergraduate curriculum would be a beneficial step in this research in order to understand the current status of incorporation of science practices across chemistry disciplines. Additionally, the conclusions drawn from the simultaneous investigation of ACS exam items and items constructed to intentionally incorporate science practices can be used to help ACS exam development committees consider the adoption of strategies to enhance the role of science practices in multiple-choice ACS exams. The rubric could be used to aid in construction of items to incorporate science practices particularly in upper-level chemistry courses.

Final Remarks

As reforms of science education continue to challenge traditional modes of assessment and instruction to provide additional evidence of student learning beyond content proficiency, chemistry education researchers and practitioners alike need to consider what this evidence will look like and how it will be measured. Additionally, researchers will need to aide practitioners to understand how to develop items to elicit such evidence. Publications and professional development opportunities that promote the changes and improvements of assessments to measure beyond content will likely help to engage chemistry practitioners in evidence-based pedagogy and assessment practices.

References

Achieve. (2013). Next generation science standards. Washington, DC: National Academies Press.

Cooper, M.M. (2014) Personal Communication.

Cooper, M. M., & Klymkowsky, M. (2013). Chemistry, life, the universe, and everything: a new approach to general chemistry, and a model for curriculum reform. *Journal of Chemical Education, 90*(9), 1116-1122.

Emenike, M., Raker, J. R., & Holme, T. (2013). Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology. *Journal of Chemical Education, 90*(9), 1130-1136.

Hodson, D. (2003). Time for action: Science education for an alternative future. *International Journal of Science Education, 25*(6), 645-670.

Longbottom, J. E., & Butler, P. H. (1999). Why teach science? Setting rational goals for science education. *Science Education, 83*(4), 473-492.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: National Academies Press.

Norris, S. P. (1997). Intellectual independence for nonscientists and other content-transcendent goals of science education. *Science Education, 81*(2), 239-258.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*: National Academies Press.

Raker, J. R., Emenike, M. E., & Holme, T. A. (2013). Using structural equation modeling to understand chemistry faculty familiarity of assessment terminology: Results from a national survey. *Journal of Chemical Education, 90*(8), 981-987.

Stiggins, R. J. (1991). Assessment Literacy. *Phi Delta Kappan, 72*(7), 534-539.

Talanquer, V., & Pollard, J. (2010). Let's teach how we think instead of what we know. *Chemistry Education Research and Practice, 11*(2), 74-83.

Towns, M. H. (2009). Developing learning objectives and assessment plans at a variety of institutions: Examples and case studies. *Journal of Chemical Education, 87*(1), 91-96.

Underwood, S. M., Cooper, M. M., Krajcik, J., Cabellero, D., & Ebert-May, D. (2014). *Designing a rubric to characterize assessments*. Paper presented at the 248th National Meeting of the American Chemical Society.