Investigating the Impact of College-level General Chemistry Curricula on General Chemistry Students’ Conceptions of Organic Acidity and Oxidation-Reduction

A Thesis Presented

by

Christian Rodriguez

to

The Graduate School

in Partial Fulfillment of the

Requirements

for the Degree of

Master of Science

in

Chemistry

Stony Brook University

August 2018

The Graduate School

Christian Rodriguez

We, the thesis committee for the above candidate for the Master of Science degree, hereby recommend acceptance of this thesis.

Dr. Gregory Rushton – Thesis Advisor

Associate Professor, Department of Chemistry & Institute for STEM Education, Stony Brook University

Dr. Katherine Aubrecht – Second Reader

Assistant Professor, Department of Chemistry, Stony Brook University

Dr. Ross Nehm – Outside Member

Associate Professor, Department of Ecology & Evolution, Stony Brook University

This thesis is accepted by the Graduate School

Charles Taber
Dean of the Graduate School

Abstract of the Thesis

Investigating the Impact of College-level General Chemistry Curricula on General Chemistry Students' Conceptions of Acidity and Oxidation-Reduction

by

Christian Rodriguez

Master of Science

in

Chemistry

Stony Brook University

August 2018

Chemistry students have historically struggled with conceptually understanding organic acidity and oxidation-reduction. Previously dominant approaches towards remediating students' misconceptions has been challenged by *Explanatory Coexistence*, which eludes to a competition between conceptions held within individuals. Conceptual reprioritization may be associated with the restructuring of conceptual dominance hierarchies, which may occur once a conceptual competition concludes. Investigation of conceptual reprioritizations of general chemistry students' conceptions of organic acidity and oxidation-reduction performed across multiple demographics using Rasch analysis, student interviews and argumentation quality assessment. Student samples belonged to two different general chemistry courses that used different curricula. One used a reform-based curriculum, that compared to the traditional curriculum, focused on discussion and argumentation. Student conceptions were captured, and tracked via repeated measures, using the *ACIDI* and *ROXCI* concept inventories. Results indicated both inventories were capable of detecting conceptual reprioritizations after instruction from both curricula. Student achievement was consistent across multiple demographic characteristics. Evidence of argumentation quality and its association with conceptual reprioritizations of organic acidity and dominant, scientifically accepted redox conceptions was collected. Individual interviews suggested conceptual reprioritizations may be attributed to their respective curricula, while also adding insight into thought processes that arose while taking both inventories. Suggestions for future work is also discussed, highlighting the development of community standards, *ACIDI* and *ROXCI* responses databases to assess general student representation, and modification of both inventories.

Table of Contents

List of Figures

Chapter 2

Chapter 3

Chapter 4

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

List of Abbreviation

ACC: Anterior cingulate cortex

AD: Alzheimer’s disease

BA: Basic argument

fMRI: Functional magnetic resonance imaging

Logit: Log-odds unit

RBC: Reform-based general chemistry curriculum

ROXCI: Oxidation-reduction concept inventory

STEM: Science, technology, engineering and math

TAP: Toulmin’s argumentation pattern

TC: Traditional general chemistry curriculum

VLPC: Ventrolateral prefrontal cortex

Acknowledgments

My deepest gratitude go to all friends, family and co-workers who aided me during the process of this work:

Rushton research group members for listening to my work for the past two years;

Chemistry main office staff for helping answer all of the questions I asked and their administrative guidance;

Dr. John Michael Linacre for helping me understand Winsteps;

Dr. Ross Nehm for being flexible and guiding my understanding of statistics;

Dr. Gena Sbeglia for helping teach me how to code and use R;

Dr. Gregory Rushton and Dr. Lisa Shah for the countless hours of professional development and helping me become an independent thinker through their support and constant challenges that helped me grow as a person.

Chapter 1: Introduction

Student Conceptions of Organic Acidity and Oxidation-Reduction

Topics within chemistry, such as acid-base chemistry and oxidation-reduction, are challenging for students of varying education levels (Bretz & McClary, 2015; Brandriet & Bretz, 2014). The evidence of the difficulty of these topics can be found through the presence of acid-base and oxidation-reduction misconceptions, or those that differ from scientifically-accepted ones (Garnett & Treagust, 1992; Rosenthal & Sanger, 2012; Brandreit & Bretz, 2014; Bretz & McClary, 2014; Bhattacharyya, 2006), inappropriate application of theories related to acid-base and oxidation-reduction chemistry (Cartrette & Mayo, 2011; Treagust et al. 2009; Garnett & Treagust, 1992; Bradley & Mosimege, 1998) and difficulty applying knowledge to problems that are presented to students in unfamiliar contexts (Cartrette & Mayo, 2011).

Previous studies investigating conceptions and mental models of acid-base chemistry have uncovered various conceptions that chemistry students have struggled with. Graduate organic chemistry students have expressed mental models of organic acids, which had little predictive power in the absence of provided data, that did not evolve past their undergraduate organic chemistry courses. This may interfere with the graduate students' ability to combine organic chemistry theory and practice (Bhattacharyya, 2006). Undergraduate organic chemistry students have expressed inappropriate use of pi bonds or lone pairs of electrons as support for Bronsted-Lowry's theory of acids and bases, lack a firm conceptual understanding of nucleophiles and electrophiles due to inexpression of the physical reasoning for why a nucleophile attacks an electrophile and had difficulty using declarative knowledge to solve problems in unfamiliar contexts (Cartrette & Mayo, 2011; Kousathana, Demerouti & Tsaparlis, 2005). College-level students that were planning to become teachers have shown the presence of acid-base misconceptions, as Bradley and Mosimege (1998) discovered that a portion of these students thought Cl^- was a strong base. Nakhleh (1994) reviewed misconceptions held by K-12 and undergraduate students. Students at the high school level held many misconceptions of the particulate nature of matter. Reported misconceptions include the following: molecules expand when heated, inappropriate representations of air molecules and acids or bases, and attribution of molecular properties to singular atoms. Similarly, college-level students were unable to properly articulate particle interaction when commonly using the Bohr model (Nakhleh, 1994). Misconceptions of particle kinetics were also reported, such as inappropriate molecular diagrams of chemical reactions and static perception of equilibrium (Nakhleh, 1994). Nakhleh cautioned the drawbacks to the persistence of such misconceptions, as they may serve as a shaky foundation that students build future conceptions upon (Nakhleh, 1994).

Previous literature regarding student difficulty with oxidation-reduction conceptions have uncovered similar results to acid-base misconceptions, as there have been reports of misconceptions regarding the particulate nature of matter for oxidation-reduction (Rosenthal & Sanger, 2012). While some valuable insight is present regarding possible obstacles to the propagation of scientifically accepted conceptions to students, such as teaching activities that may not convey the necessity of new conceptions, or their intelligibility, plausibility or fruitfulness (De Jong, Acampo & Verdonk, 1995). Stains and Talanquer (2008) reported that novice, and some intermediate, college-level chemistry students had displayed reliance on explicit surface features of chemical equations to classify chemical reactions, while implicit features were utilized for classification more as level of chemistry expertise increased. Garnett and Treagust (1992) reported high school students (i) inappropriately applying oxidation numbers to monoatomic ions and (ii) tracking charge of polyatomic species rather than oxidation numbers of atoms within each species to identify oxidation-reduction reactions. Rosenthal and Sanger (2012) identified more misconceptions held by high school students regarding oxidation-reduction using computer animations and semi-structured interviews. The most striking misconceptions they had identified include: ions in solution forming neutral ion-pairs as opposed to dissociating, adding or losing valence electrons does not affect size or charge, and water drives oxidation-reduction reactions. Further discussing a major factor that may have led to such misconceptions: misinterpretation of the computer animations. Some students had revealed, in interviews, that they had incorrectly assigned molecules to different colored shapes in the computer animation. This warrants valid instruments to be used to uncover conceptions held by students, as a potential cause of misconceptions may not be instruction or topics that tend to be difficult to conceptualize. Instead, a cause of observed misconceptions may be the instrument used to uncover conceptions.

Conceptual Reprioritization

Classical approaches to misconceptions, or scientifically inaccurate conceptions, held by students has been to eliminate or “reconstruct” misconceptions, or change students’ relationships with contexts and the misconceptions related to those contexts (Linder, 1993). These approaches may begin with the presentation of a cognitive conflict that exposes weak links in the armor of misconceptions by providing examples or contexts in which those misconceptions do not hold true or cannot be successfully applied to solve a problem (Potvin, 2017; Posner et al. 1982). This approach exemplifies an attempt to de-value misconceptions to make the adoption of scientifically accurate conceptions more easily achieved. The exposure of weak links in misconceptions is part of a process that often begins with the instructor collecting the conceptions of the students and identifying if they are or are not scientifically accepted (Potvin, 2017; Posner et al. 1982). Then, instructional approaches can be implemented to expose weak links, introduce students to more useful, scientifically accepted conceptions, and then re-assess student conceptions after instruction (Potvin, 2017; Posner et al. 1982). The tracking of conceptions before and after instruction may provide evidence for the occurrence of "conceptual change." This would be indicated by differing conceptions held before and after instruction (Libarkin, 2008). The term "conceptual change" implies that previous conceptions held by students regarding particular phenomena are no longer the same. Rather, they have undergone an evolution, while also implying a lack of multiple conceptions being present for the same phenomena (Potvin, 2017; Shtulman & Lombrozo, 2016).

However, recent studies refute a "change" of conceptions within individuals. Rather, the results of these studies support a "coexistence" of conceptions, meaning multiple conceptions may be present within an individual for the same phenomena (Shtulman & Lombrozo, 2016). Masson et al. (2014) had explored certainty and uncertainty in novice undergraduate students' (N = 22) conceptions of electrical circuits using functional magnetic resonance imaging (fMRI) and observed that, during times of uncertainty, the anterior cingulate cortex (ACC), which has been linked to conflict detection, was activated. Potvin and associates had then posited that activation of the ACC, in absence of external negative feedback to participants, may be due to conflict of knowledge gained during school and misconceptions held by participants.

Foisy et al. (2015) had explored association between inhibition, indicated by regional brain activation obtained using fMRI, and conceptions of mechanics (i.e. heavier ball falling faster than a lighter ball) held by novices and experts in science. Novices were defined as participants that held misconceptions of mechanics and experts were defined as participants who had theoretically undergone the process of conceptual change due the absence of misconceptions related to mechanics. fMRI results had indicated that experts activated the ventrolateral prefrontal cortex (VLPC), an area of the brain associated with inhibition, significantly more than novices when observing a video of a heavier ball falling and hitting the ground before a lighter ball (a common mechanical misconception). This suggests that experts in mechanics may rely more on a brain region linked to inhibition, when explicitly evaluating stimuli that is not scientifically accepted (i.e. heavier ball falling faster and hitting the ground before a lighter ball), than novices. Furthermore, this explicit evaluation of a scientifically inaccurate phenomenon, that has shown increased VLPC activation in experts compared to novices, may serve the purpose of inhibiting scientifically inaccurate conceptions of mechanics. This would then suggest the presence of multiple conceptions of mechanics, as the experts, by definition, also hold scientifically accepted conceptions of mechanics.

Lombrozo, Kelemen and Zaitchik (2007) had investigated the presence of teleological explanations, commonly held by children, in patients with Alzheimer’s disease (AD) to determine if such explanations are outgrown as one gets older. Teleological explanations are defined as functional modes of explanation (i.e. a heart exists to pump blood). However, not all teleological explanations are accepted by adults, as they may even be rejected when adults are evaluating causal explanations (Lombrozo & Carey, 2006). Therefore, teleological explanations may be outgrown as one gets older to more appropriately determine causal explanation for observed phenomena, and a re-emergence of teleological explanations during adulthood may suggest that such explanations may be outcompeted by causal explanations rather than outgrown by them (Lombrozo & Carey, 2006). Lombrozo, Kelemen and Zaitchik (2007) had discovered that AD patients “broadly accept and prefer teleological explanations,” while healthy adults did not broadly accept and prefer unwarranted teleological explanations, as AD patients had expressed that the purpose of rain is to provide water for plants and animals, and the existence of trees is to provide shade. This suggests a preference for teleological explanations is not truly outgrown and may re-emerge once the ability to apply causal explanations for the same phenomena is compromised. These results also suggest the existence of multiple conceptions for the same phenomena and the ability of either conception to compete may impact which one wins the competition and becomes the dominant conception.

Recently, *explanatory coexistence*, has emerged, after many years of the coexistence claim being present (Linder, 1993), as a theory that challenges old theories of “conceptual change” (Shtulman & Lombrozo, 2016). Explanatory coexistence contrasts more traditional ideas of conceptual change by positing an existence of multiple conceptions held within an individual due to their usefulness (Shtulman & Lombrozo, 2016). A competition between multiple useful conceptions that are used to explain the same phenomena then takes place, and the result is the emergence of a dominant conception (Shtulman & Lombrozo, 2016). The emergence of dominant conceptions after competition implies a reprioritization process that takes place within an individual and a creation of a conceptual dominance hierarchy, with respect to the conceptions that participated in the competition.

Explanatory coexistence lays the groundwork for how this study defines conceptual reprioritization and the proposed mechanism that may promote it . Conceptual reprioritization is defined in this study as the restructuring of conceptual dominance hierarchies within an individual. Conceptual reprioritization is proposed to occur after a competition of useful conceptions occurs and a “winning” conception emerges (Shtulman & Lombrozo, 2016). However, the margin of victory of a conception above others may relate to the persistence of “losing” conceptions, as it may be possible that a small margin of victory represents a lack of separation of usefulness between competing conceptions (Shtulman & Lombrozo, 2016).

Social Constructivism

Vygotsky’s theory of social constructivism posits that learning is socially achieved (Vygotsky, 1962). Social constructivism lays the theoretical structure for methodological aspects of this study, as certain learning environments that are explicitly structured in the involved courses place students into small cooperative learning groups, or a large discussion-oriented lecture. These environments theoretically promote the co-construction of chemistry knowledge through diverse conceptual exposure regarding oxidation-reduction and organic acidity conceptions, as they promote discussion and argumentation (Nussbaum, Sinatra & Poliquin, 2008; Shah et al. 2018; Shtulman & Lombrozo, 2016; Erduran, Simon & Osborne, 2004). Discussion and argumentation allow students to be exposed to various conceptions and perspectives, and may provide them a strong opportunity to undergo conceptual reprioritization, or reconstruct their conceptual dominance hierarchies regarding the topic of discussion (Shtulman & Lombrozo, 2016). Opportunities for conceptual reprioritization may increasingly occur as competitions between useful conceptions increase in frequency, which may be due to increased conceptual exposure in discussion-oriented lecture or cooperative learning environments (Shtulman & Lombrozo, 2016). Furthermore, diverse learning groups have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001), which may increase student exposure to useful conceptions in such learning groups, therefore increasing the students' opportunities to undergo conceptual reprioritization.

Cooperative Learning

Cooperative learning is a form of active learning that is centered around student collaboration and its pedagogical goals are consistent with social constructivist theory through the promotion of socially constructed knowledge (Bowen, 2000; Vygotsky, 1962; Johnson, Johnson & Smith, 1991). Cooperative learning environments place students into small groups and encourage student collaboration, accountability, and group processing (Bowen, 2000; Johnson, Johnson & Smith, 1991). Cooperative learning environments have been linked to mixed learning outcomes in chemistry at the secondary and tertiary level, as some course and test improvement and reduction have been reported (Bowen, 2000). These mixed results suggest that more research must be done to better understand, and potentially predict, learning outcomes of students in cooperative learning settings. Previously, researchers were concerned with unequal benefits extracted from cooperative learning environments depending on differing student ability (Slavin, 1996). However, Slavin (1996) reviewed the cooperative learning literature and discovered that cooperative learning has the same positive impact across high, middle, and low performing students. Thus reducing the fear of unequal student benefits due to variation in student ability, and serving as evidence for the creation of diverse cooperative learning groups that contain students of unequal student ability. The cooperative learning environments involved in this study were designed to increase group diversity (different genders, races, ethnicities, etc), rather than student ability due to lack of student ability indicators for first-year students (i.e. undergraduate GPA). Diverse learning groups have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001), which may increase student exposure to useful conceptions. Furthermore, diversity of thought may play a role in conceptual exposure, as it may enhance student discussions through exposure to different student perspectives (Powell & Kalina, 2009). Diversity of thought may be achieved through diversifying cooperative learning groups and focusing the topics of student-student discussions on a mix of group assignments and less formal discussions that spread cultural understanding (Powell & Kalina, 2009).

Student Argumentation and Discussion

Student discussion and argumentation are linguistic mechanisms that offer students the opportunity to voice their conceptions, be exposed to conceptions held by their peers, and co-construct conceptions in group settings (i.e. small cooperative learning groups, large discussion-oriented lectures). Argumentation provides students a methodological base to explicitly voice their conceptions, as students may connect argumentation components such as a claim, data, and warrant together to argue on their conception's behalf (Toulmin, 1958; Kulatunga et al. 2014; Shtulman & Lombrozo, 2016; Shah et al. 2018). Student argumentation has been linked to increased learning outcomes in college-level physics courses, and positive outcomes in college-level general chemistry (Nussbaum, Sinatra & Poliquin, 2008; Shah et al. 2018). Furthermore, student argumentation may be enhanced through diverse learning groups, which have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001), and diversity of thought, which may increase exposure to diverse conceptual perspectives (Powell & Kalina, 2009). Enhancement of student argumentation in the aforementioned manners may increase the odds of students being exposed to useful conceptions, and perhaps increase the odds of conceptual reprioritization occuring (Shtulman & Lombrozo, 2016).

Arguments can be extracted from student discussions using audio and video equipment, as they provide a record what was discussed. These arguments can then be transcribed and coded to determine the presence of argumentation components. One of the most commonly used argumentation coding scheme is Toulmin’s argumentation pattern, or TAP (Heng et al. 2014; Erduran, Simon & Osborne, 2004). TAP breaks down arguments into the following components: claim, data, warrant, rebuttal, qualifier, and backing. The combination of a claim, data and warrant has been previously defined as a basic argument by Kulatunga et al. (2014), and their creation may serve as a simple way to practice argumentation. A more sophisticated argument is a basic argument with the addition of qualifiers or rebuttals (Heng et al. 2014; Eduran, Simon & Osborne, 2004). Definitions of these components and a basic argument can be found below in table 1.

Table 1. Argumentation components adapted from Toulmin (1958) & Kulatunga et al. (2014).

Component	Definition
Claim	An assertion put forth to the public regarding the topic/question of interest.
Data	Facts or information used to support a claim.
Warrant	A justified connection between data and a claim.
Backing	Assumptions under which the warrant holds power.
Qualifier	Conditions under which a claim is true.
Rebuttal	Refutations that may undermine a previous claim.
Basic Argument (BA)	A verbal utterance that contains a claim, data and warrant connecting the data to the claim.

TAP can be used to detect basic arguments and components of argumentation when paired with activities that elicit student-student discussions, such as cooperative learning activities or discussion-oriented lectures (Nussbaum, Sinatra & Poliquin, 2008; Shah et al. 2018; Kulatunga et al. 2014; Heng et al. 2014; Talanquer & Pollard, 2010). Heng and associates (2014) claimed “the construction of scientific arguments requires cognitive involvements, such as analyzing and making sense of the data, generating explanations, supporting the idea, and challenging the validity of an idea” as a possible path towards previous notions of conceptual change. This may also serve as a path towards conceptual reprioritization through the creation of competitions between useful conceptions, as the conceptual product of a scientific argument may be seen as useful (Shtulman & Lombrozo, 2016). However, the participation of students in these discussions may affect the benefits students extract from collaboration and it is tough to increase student participation in scientific argumentation (Cohen, 1994; Sampson & Clark, 2009). Active participation in collaboration and discussion may not be necessary to for a student to be exposed to useful conceptions, as one may be able to collect useful conceptions by strictly listening (passive participation) to conceptions that are put forth in discussion by their group mates or classmates (Cohen, 1994; O'Connor et al. 2017). Therefore, active participation in group discussion and argumentation may not be necessary to undergo conceptual reprioritization. Although it may be possible that active participation increases the likelihood of one undergoing conceptual reprioritization, as active participation in collaborative activities has been linked to increased student performance when applying ideas that arise from collaboration to problem solving tasks (Cohen, 1994).

Although TAP is commonly used, there are some noteworthy limitations to the coding scheme (Erduran, Simon & Osborne, 2004; Heng et al. 2014). First, TAP does not assess the correctness of an argument, it simply is in place to detect the presence of arguments and their components (Heng et al. 2014). Second, it does not assess the quality of an argument (Erduran, Simon & Osborne, 2004). However, Erduran and co-workers have utilized TAP as a foundation to create an analytical framework that assess the quality of arguments (Erduran, Simon & Osborne, 2004). Erduran's analytical framework uses TAP to detect argumentation components, and then contextualizes the use of those components, or combination of components, into 5 levels of argumentation quality (i.e. claim v claim, claim v counter-claim). Erduran's analytical framework for argumentation quality can be found below in table 2.

Table 2. Argumentation quality framework adapted from Erduran, Simon & Osborne (2004).

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

Argumentation quality and its relation to student ideas, at various education levels, has been alluded to in multiple previous studies (Bell & Linn ,2000; McNeil & Pimentel, 2010; Dawson & Venville, 2009). However, it has only been linked to chemistry students' conceptions of organic acidity in a recently published paper by Shah et al. (2018), as increased argumentation quality of reform-based general chemistry curriculum students seemed to be associated with conceptual understanding of organic acidity.. Similar to Shah et al. (2018), argumentation quality, assessed using Erduran's analytical framework, is used in this study as a framework that may be able to associate argumentation quality and conceptual reprioritization, as more sophisticated arguments (i.e. higher quality arguments) may lead to the exposure of more useful conceptions and an eventual reprioritization of students' redox or organic acidity conceptions.

College-level Chemistry Curricula

Van Berkel et al. (2000) had assessed previous chemistry curricula and pedagogy, building upon Kuhn’s assessment of chemistry textbooks (Kuhn, 1963). Kuhn had discovered that chemistry textbooks had not varied much, aside from education level and pedagogy (Kuhn, 1963). Van berkel and associates had discovered similar results, while focusing on the propositions and reliance on algorithms of chemistry curricula at the secondary and tertiary level, claiming the following (Van Berkel et al. 2000, p 152):

> “The structure of the currently dominant school chemistry curriculum is accurately described as a rigid combination of specific substantive structure… the structure of dominant school chemistry as a whole suffers from a sevenfold isolation: from common sense, everyday life and society, history and philosophy of science, technology, school physics, and from chemical research.”

Van berkel and associates caution that deviating from such a structure may only be achieved via replacement of the underlying structures of school chemistry. This emphasis on students’ algorithmic ability does not promote the exposure of student conceptions, therefore reducing the likelihood of a student undergoing conceptual reprioritization (Talanquer & Pollard, 2010; Shtulman & Lombrozo, 2016). Similar thoughts are discussed by Bulte et al. (2006), discussing the abstract, and potentially meaningless, concepts presented in contexts other than their origin that are taught in modern chemistry curricula; and how modern chemical knowledge has evolved from the late nineteenth and early twentieth centuries along with chemical practices, therefore making those teachings out-of-date. Talanquer (2013) discusses constraints that previous ideas for chemistry curricula place upon improvement of learning core chemistry ideas and practices. Talanquer had noted that explanatory constructs of fundamental chemistry concepts (i.e. chemical reactions and elements) being separated from practical uses in chemistry as an obstacle to chemistry teaching, while suggesting the need to eliminate this separation and focus on the identification of fundamental questions of chemistry that may be answered through intellectual and practical tools within the discipline (Talanquer, 2013).

Bulte et al. (2006) investigated how high school chemistry students may legitimize the learning of a water quality unit in their high school course, or the “need-to-know” principle. The “need-to-know” principle has been defined in the following manner: “the context must legitimize the learning of chemical theory from the perspective of the students and thus make their learning intrinsically meaningful” (Bulte et al., 2006). A water quality unit was chosen due to its easily applicable content to societal functions, which may appeal to students. The study designed the unit to begin with a leading context-question (i.e. is the water clean enough in our neighborhood?”), followed by sub-questions that were expected to align with the “need-to-know” principle, and ultimately end in a collective reflection on what was discussed. These questions underwent several revisions, which uncovered two noteworthy issues that may occur when trying to modernize chemistry curricula to promote conceptual understanding of traditional chemistry concepts. First, the attempt to align traditional concepts with modern chemistry contexts or practices may not solve the problem of conceptual misalignment, as chemistry concepts may still not align with modern chemistry contexts or practices. Second, the students may not be able to identify the connection between activities and the concepts that are associated with those activities, as students may not have enough chemistry expertise to be able to identify associations. However, alignment of theory, practice, and concepts may reduce the abstract and unconnected notion of chemistry course material. Bulte and associates (2006) had also reported an association between a guided problem-posing approach with student identification of rationale to elicit adoption of chemistry knowledge.

Recently, Talanquer and Pollard (2010) created a reform-based general chemistry curriculum that is designed to promote conceptual understanding of general chemistry. This new curriculum, *Chemical Thinking*, does so by centering the students' chemistry learning experience around fundamental questions of modern chemistry that link course units together. The goals of this curriculum are for students to create meaningful ways of thinking like a chemist through ideas that guide chemical thoughts of analysis, synthesis, transformation and modeling (Talanquer & Pollard, 2010). Talanquer and Pollard then propose levels of progression for understanding core concepts in *Chemical Thinking's* curriculum. Table 3, below, is a sample of learning progression levels from the *Chemical Thinking* curriculum for molecular interactions and physical properties of molecular compounds.

Table 3. Chemical Thinking learning progression levels and referred units.

Level 1: Recognizes that differences in physical properties can be explained based on differences in the strength of attractive forces between submicroscopic particles (Unit 1).
Level 2: Relates the differences in the strength of intermolecular forces to differences in molecular structure and composition (Units 1 and 3).
Level 3: Explains differences in the strength of intermolecular forces based on differences in charge distribution in a molecule (Unit 3).
Level 4: Predicts differences in physical properties based on analysis of molecular structure and charge distribution (Unit 3).

It is important to recognize that these progressions are based on a students' level of thinking, rather than algorithmic proficiency, and offer a new framework that can be referenced when creating assessment tools (i.e. clicker questions, quizzes) to measure students' levels of chemical thinking (Talanquer & Pollard, 2010). The promotion and alignment of *Chemical Thinking* with conceptual understanding of chemistry incentivizes measurement of conceptual understanding of students who take a general chemistry course that uses *Chemical Thinking*. Concept Inventories, cognitive instruments (discussed in more detail below), may be capable of capturing the conceptual understanding of students in courses that use *Chemical Thinking* or other general chemistry curricula. Appropriate comparison of conceptions held by students educated by *Chemical Thinking* and traditional chemistry curricula may provide insight into the effectiveness of *Chemical Thinking* and conceptual understanding and reprioritization.

Concept Inventories

Concept inventories are multiple-choice cognitive instruments that are designed to measure conceptions (Libarkin, 2008). They can be used for diagnostic purposes to gain insight into how students, primarily at the college-level, are thinking about domains within various disciplines (Libarkin, 2008). Concept inventories are designed to represent misconceptions and scientifically accepted conceptions in the instrument's multiple-choice answers (Bretz & McClary, 2014; Libarkin, 2008). These represented conceptions allow for diagnostic assessment of student conceptions, which may be used for instructional purposes, as it allows instructors to be aware of prominent misconceptions held by their students (Libarkin 2008; Bretz, & McClary, 2014; Shah et al. 2018). Once aware of prominent misconceptions, instructors may be able to adjust what domains may require more time to address misconceptions and reduce their conceptual dominance (Libarkin, 2008; Steif & Hansen, 2007; Shah et al. 2018; Bretz & McClary, 2014).

Concept inventories can also be used for assessment purposes in repeated-measures studies to capture potential impact of instructional intervention or curricula on student conceptions (Libarkin, 2008). Changes in responses from pre to post-test on a concept inventory may be evidence of conceptual reprioritization, as different responses may represent dominant conceptions held by a respondent at different points in time. Conceptual reprioritization may be attributed to the instructional intervention or curriculum if respondents indicate instruction or curricula as primary factors influencing their conceptions.

Many STEM disciplines (i.e. physics, biology, astronomy, etc) have used concept inventories for instructional purposes (Libarkin, 2008). However, until recently, the chemistry education community has lagged behind other STEM disciplines for creating and utilizing concept inventories, as concept inventory construction involves many rounds of validity and reliability (Libarkin, 2008). The purpose of constructing concept inventories is to evaluate conceptual understanding of novice students (Libarkin, 2008). The chemistry education community has recently began developing concept inventories that cover various domains of chemistry (Libarkin, 2008). A challenge to the use of concept inventories has been establishing reliability, according to traditional indicators of reliability (i.e. cronbach's alpha), of the developed instruments. The problem that arises with reliability of concept inventories has been argued to be due to the nature of the questions (Bretz & McClary, 2014), as they are not intended to be related to each other as there may be many different conceptions for the same questions on a concept inventory (Bretz & McClary, 2014). This variance in conceptions, manifested by answer choices, may contribute to low covariance among item responses, therefore reducing the value of reliability estimates (i.e. Cronbach's alpha). This lack of alignment between theory and measurement for concept inventories in chemistry may indicate a need to develop a more appropriate form of reliability assessment for these instruments (Shah et al. 2018). As of the writing of this paper, there are concept inventories present within the discipline of chemistry that address the following domains of chemistry: particulate nature of matter (Nyachwaya et al. 2011; Stains et al. 2011), covalent and ionic bonding representations (Luxford & Bretz, 2014), kinetic particle theory (Treagust et al. 2010), solution chemistry (Adadan & Savasci, 2012), acid-base chemistry (Bretz & McClary, 2014; Rahayu et al. 2011), oxidation-reduction (Brandreit & Bretz, 2014), chemical equilibrium (Ozmen, 2008; Voska & Heikkinen, 2000), electrolysis (Sia et al. 2012), and enzyme-substrate interactions (Bretz & Linenberger, 2012). This study will address, and utilize, the *ACIDI* (Bretz & McClary, 2014) and *ROXCI* (Brandreit & Bretz, 2014) concept inventories in subsequent chapters.

Previous study

This study stems from the previous study by Shah et al. (2018), which had investigated the conceptions held by first-year, reform-based general chemistry curriculum students at a large public research university in the northeastern region of the United States. *Chemical Thinking*'s impact on first-year general chemistry students' conceptions of organic acidity was investigated, while also compiling evidence for possible association of argumentation quality and scientifically accepted conceptions of organic acidity. Shah and associates utilized the *ACIDI* concept inventory to capture conceptions of organic acidity before, after, and a prolonged time after instruction by participating students (i.e. pre, post and delayed post-test). Item-level analysis of inventory responses (i.e. pre-post comparison) indicated a on students' scientifically accepted conceptions, and potential conceptual reprioritizations, of induction and resonance (i.e. quantity of resonance structures) by the *Chemical Thinking* curriculum. Furthermore, a prolonged retention of conceptual gains from pre to delayed post-test was reported. Evidence of a positive association (R = 0.72) between high quality group argumentation and correct *ACIDI* post-test responses was also collected. Student interviews provided insight into how first-year students were thinking about induction and resonance, as a mixture of scientifically accepted and alternative conceptions were present. First-year students seemed to better understand the inductive effect, compared to resonance, reporting difficulty comparing the quantity of resonance structures of phenol to the quality of resonance structures of acetylacetone. First-year students had indicated the quantity of resonance structures, of phenol's conjugate base, makes it more acidic than acetylacetone's, even though the resonance structures of acetylacetone's conjugate base are more stable than the resonance structures of phenol's conjugate base. This resulted in responses that indicate the presence of a misconception.

This study builds upon Shah et al. (2018) by first investigating the association between *Chemical Thinking* and conceptual reprioritization using item response theory (i.e. Rasch analysis) and its indicators of conceptual reprioritization (i.e. person ability logit scores) while accounting for demographics such as gender (male v female) and student generation (first-gen vs non-first-gen). Second, association between argumentation quality and person ability at the group level, also indicated by person ability logit scores, was investigated. Third, insight into potential obstacles that may influence reform-based general chemistry curriculum students' conceptions of resonance was revealed. Finally, this study translates some methodology and revised analytical frameworks from Shah et al. (2018) to *ROXCI*, a concept inventory that targets student conceptions of oxidation-reduction (Brandreit & Bretz, 2014), for both reform-based and traditional general chemistry curriculum students.

Gaps in Literature

This study addresses some gaps in the chemistry education literature in four ways. First, potential association of group argumentation quality with undergraduate students' conceptions of oxidation-reduction in addition to organic acidity was investigated, as it has not been previously assessed. Second, collection of first-generation, compared to non-first-generation, undergraduate students' conceptions of organic acidity and oxidation-reduction was performed and may act as a base for comparison of other first-generation undergraduate students. Third, assessment of student performance on the *ACIDI* and *ROXCI* instruments using item-response theory (i.e. Rasch analysis) has not been performed before this study, while also adding to previous arguments for the preferred use of item-response theory over classical test theory for such instruments. Finally, this study addresses future investigation to aid in closing gaps that still remain in the chemistry education literature, such as the creation of community standards for undergraduate students' conceptions of organic acidity and oxidation-reduction, potential refinement to the *ACIDI* and *ROXCI* instruments, and suggestions for comparison of different student populations.

Chapter 2: Conceptual Reprioritization of Organic Acidity using *Chemical Thinking*

Introduction

Background

Acid-base chemistry has traditionally been one of the most difficult topics for students to retain knowledge, perform under unfamiliar question contexts, and conceptually understand at both the graduate and undergraduate level (Bretz & McClary, 2015; Bradley & Mosimege, 1998; Cartrette & Mayo, 2011; Bhattacharyya, 2006). Previous, studies have reported (under)graduate students having difficulty qualitatively applying mental models of organic acids, inaccurately using Bronsted-Lowry theory, and conceptually understanding nucleophiles and electrophiles (Bhattacharyya, 2006; Cartrette & Mayo, 2011, Bradley & Mosimege, 1998). These difficulties, paired with time pressure on college students' need to learn course material in a short amount of time, may lead to memorization and heuristics as major routes for test preparation, rather than strong conceptualization of chemistry material (Talanquer, 2018; Bhattacharyya, 2006). These perceptually quicker routes have reportedly lead to instances of dependence on heuristics (e.g., OH functional group being indicative of an acid), vulnerability to test-taking mistakes, incomplete mental models, scientifically unaccepted conceptions, and reduced sensitivity to long-term retention of chemistry concepts (Bhattacharyya, 2006; Cartrette & Mayo, 2011; Bradley & Mosimege, 1998). Strong conceptual understanding of acid-base chemistry may be indicated by students demonstrating scientifically accepted conceptions (Bretz & McClary, 2015).

Cooperative learning, a form of active learning, has been commonly used to increase student performance, conceptual understanding, and been linked to content retention in college level chemistry courses (Bowen, 2000; Freeman et al. 2014; Shah et al. 2018). Cooperative learning environments encourage student collaboration and offer students opportunities to argue, which may promote conceptual understanding of learnt material (Nussbaum, Sinatra & Poliquin, 2008; Shah et al. 2018; Bowen, 2000). Argumentation, a form of discussion, may promote conceptual understanding and reprioritization of conceptions if students are exposed to useful conceptions (Shtulman & Lombrozo, 2016; Lombrozo & Carey, 2006; Lombrozo, Kelemen and Zaitchik, 2007). Conceptual reprioritization is defined in this study as the restructuring of conceptual dominance hierarchies within an individual. Useful conceptions of chemistry topics, when present within the mind of an individual, may compete with each other for dominance and the conclusion of such a competition may lead the individual restructuring their conceptual dominance hierarchies of chemistry. Diverse learning groups (i.e. groups with members of different races or ethnicities), which may be implemented in cooperative learning environments, have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001). Increased critical thinking and problem-solving skills may lead to higher quality arguments that may contain useful conceptions proposed by members in diverse learning groups. A potential way to detect quality arguments that may contain useful conceptions is through Erdruan's analytical framework for argumentation quality (Erduran, Simon and Osborne, 2004). This framework detects different levels (1-5) of vocalized arguments and may be applied across various STEM disciplines. Ascending levels of argument account for longer arguments, increasing presence of basic argumentation components (i.e. claim, data, warrant) and increasing argumentation components that are thought to be part of higher quality arguments, such as a rebuttal (Erduran, Simon and Osborne, 2004). Exposure to useful conceptions may be more likely as argumentation quality increases, and potentially lead to reprioritization of conceptions (Shah et al. 2018). However, variations in learning gains have been observed across instructional methods and contexts that utilize cooperative learning environments, urging researchers to investigate the reasons behind such variations (Bowen, 2000).

Traditional general chemistry curricula may not align well with learning environments (i.e. cooperative learning) that may increase a student's likelihood to undergo conceptual reprioritization, as traditional curricula place less emphasis on argumentation and more emphasis on algebraic proficiency to perform well (Talanquer, 2013; Van Berkel et al. 2000; Kuhn, 1963; Bulte et al. 2006; Shah et al. 2018). A reform-based general chemistry curriculum, *Chemical Thinking*, developed by Talanquer and associates, is organized into units based on fundamental questions of chemistry that guide chemical practices (i.e. How do we distinguish substances?) that encourages students to go beyond algebraic methods of learning and into conceptual understanding of chemistry material through instructor-student and student-student discussion (Talanquer, 2013; Sevian & Talquer, 2014; Talanquer, 2018). Participation, passive or active, in these discussions offer students the opportunity to reveal their conceptions, listen to conceptions of others, and the potential to appropriately reprioritize their conceptions (Cohen, 1994; Shah et al. 2018; Shtulman & Lombrozo, 2016). Previous reports of peer discussion encouraging argumentation, and potential reprioritization of conceptions, have shown to be at least as valuable as direct instruction when answering undergraduate physics questions (Nussbaum, Sinatra, Poliquin, 2008).

Recently, chemistry education researchers have developed concept inventories, instruments designed to capture student conceptions via multiple-choice questions (MCQs), that may be used to gather evidence of conceptual reprioritization of acid-base chemistry concepts (Libarkin, 2008; Steif & Hansen, 2007; Bretz & McClary, 2015; Rahayu et al. 2011). Captured conceptions, indicated by answers to MCQs, include both scientifically accepted conceptions, appropriate modes/models of thinking determined by the scientific community, and alternative conceptions, modes/models of thinking that aren't deemed appropriate by the scientific community (Bretz & McClary, 2015). Concept inventories, paired with curricula or instructional approaches that offer students increasing opportunities to reprioritize their conceptions through discussion and argumentation, may provide an effective and useful opportunity to detect conceptual reprioritization. Evidence of conceptual reprioritization may be collected through repeated measures methods by tracking changes in student responses to concept inventory items. To date, only two studies have investigated undergraduate students' conceptions of organic acidity using *ACIDI* (Bretz & McClary, 2015; Shah et al. 2018). This study aims to go beyond concept collection at the class level and into demographic sub-sections (i.e. gender, student-generation) to initially assess the presence, or lack thereof, of an achievement gap. Furthermore, this study adds uncommon IRT insight into the reliability and validity of the *ACIDI* concept inventory.

Theoretical Framework

Vygotsky's social constructivism theory argues that learning is socially constructed (Vygotsky, 1962). The social construction of knowledge provides the theoretical basis of cooperative learning and discussion-oriented instruction through support of achievement and learning being attained in a collaborative manner. Traditional theories regarding newly introduced concepts and ideas, which may be introduced in social settings such as cooperative learning environments, argue for a process of conceptual change (Linder, 1993; Posner et al. 1982). Implicit to the term conceptual change is the assumption that conceptions evolve into something that they previously were not. Most recently, a theory of conceptual coexistence has arisen, suggesting that concepts are never truly eliminated, or "change." Instead, coexistence claims conceptions accumulate within the mind of the individual (Shtulman & Lombrozo, 2016; Potvin, 2017). Conceptions, both new and old, may then compete for conceptual dominance within the individual (Shtulman & Lombrozo, 2016). After the conclusion of such a competition, conceptual reprioritization, or restructuring of conceptual dominance hierarchies, may take place.

Chemical Thinking, as an instructional approach in this study, primarily functions as a platform for students to become aware of chemistry conceptions through a discussion-oriented lecture and cooperative learning workshop (Talanquer, 2013; Talanquer & Pollard, 2010; Sevian & Talanquer, 2014). The goal of this approach is for students to become aware of many conceptions and different ways of thinking about chemistry, both normative, or scientifically accepted, and alternative, not scientifically accepted (Talanquer & Pollard, 2010). Ideally, the student is then setup for a competition to take place between acknowledged, and useful, conceptions. Opposing views may produce meaningful discourse through discussion, or argumentation, as argumentation may lead to increased exposure to useful conceptions

(Shtulman & Lombrozo, 2016). Although there is no agreed upon definition of argumentation, it has previously been defined as "a verbal activity oriented towards the realization of a goal" (Micheli, 2011). Previous research has shown positive learning gains and potential association of argumentation and student conceptions when arguments take place in small groups (Bell & Linn, 2000; McNeil & Pimentel, 2010; Shah et al. 2018; Nussbaum, Sinatra and Poliquin, 2008). This alignment of learning gains, potential increased exposure to useful conceptions through argumentation, and a reform-based teaching approach that promotes argumentation and the social construction of knowledge allows the ability to measure any associations between student performance and conceptual reprioritization with *Chemical* Thinking.

ACIDI Concept Inventory

ACIDI, developed by the Bretz group, is a cognitive instrument designed to detect undergraduate students' conceptions of organic acidity. *ACIDI* aims to identify conceptions held by students of trends in organic acid strength via nine multiple-choice questions (MCQs), six of which are two-tiered, and follow a three-question sequence (Bretz & McClary, 2015). Two-tiered questions aim to address reasoning question responses by extending one question into two separate questions. The first question, tier-one, is similar to a common multiple choice question as it simply asks for an answer to a problem. The second question, tier-two, is an extension of the first as it asks for a reason for why your first answer best solves the problem. Organic structures used in the inventory sequences are show below in figure 1.

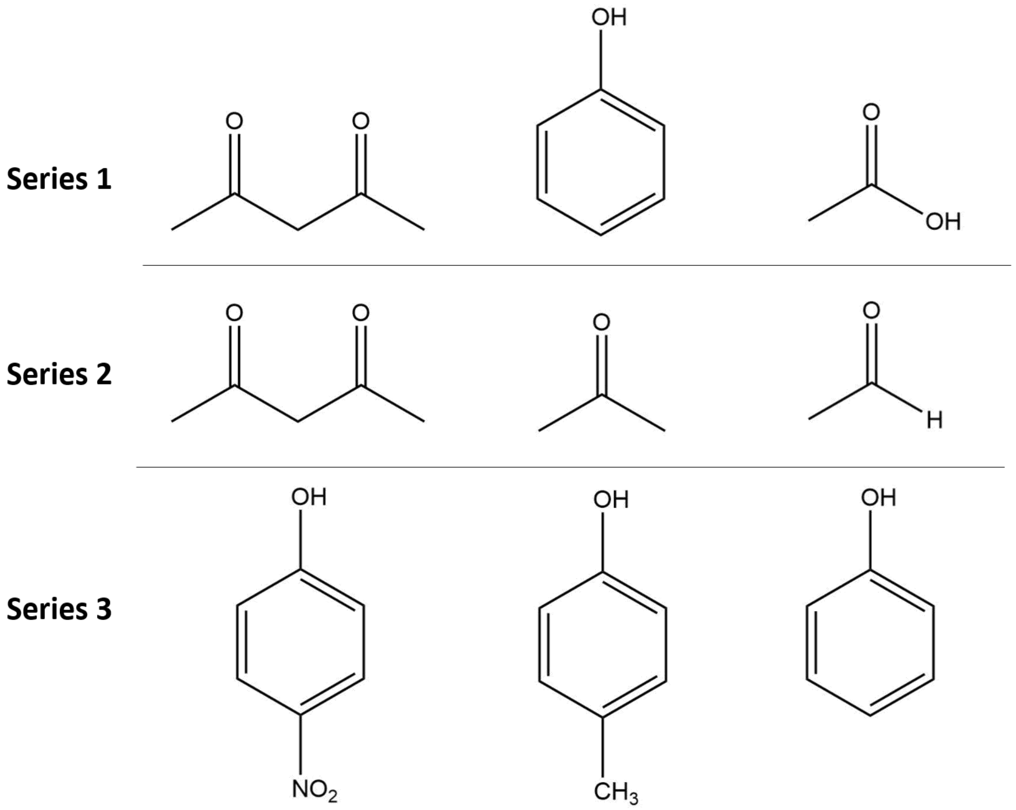


Figure 1. Organic species used in ACIDI.

The first question gives the respondent three organic structures, reveals the most acidic species, and then asks the respondent to choose an answer that they believe best represents the reason why that species is the most acidic. The second question, tier-one, asks the respondent to rank the acidity of the remaining compounds, and the third, tier-two, asks the respondent to choose a reason that best represents why they chose their previous ranking. Students are expected to use their knowledge and conceptual understanding of induction and resonance to answer all *ACIDI* items.

Rationale and Research Questions

This study aims to: First, increase the record of undergraduate students' conceptions of organic acidity, as there are very few. Second, this study aims to evaluate *Chemical Thinking* with regard to undergraduate students' conceptions of organic acidity, as alignment of the reform-based general chemistry curriculum with conceptual reprioritization may provide insight to what chemistry curricula may be linked to undergraduate students' scientifically accepted conceptions of organic acidity. With these goals in mind, the following research questions were formulated.

1. What conceptions of organic acidity are held by undergraduate general chemistry students after *Chemical Thinking* instruction?
2. Can *ACIDI* detect reprioritizations of undergraduate general chemistry students' conceptions of organic acidity? If so, to what degree?

Methods

Research Design

This study took place at a large pubic university in the northeastern region of the United States. The following protocol was executed after receiving IRB approval, package 917004-8, at the university where this study took place. A first-year, reform-based general chemistry class was given the *ACIDI* concept inventory three times, 15 minutes each, during the workshop section of their course. First, a pre-test was administered to measure organic acidity conceptions held by students before relevant instruction. Second, a post-test was given to measure reprioritizations of organic acidity conceptions held by students after three weeks of relevant instruction. Finally, *ACIDI* was given third time as a delayed post-test to measure retention of scientifically accepted organic acidity conceptions, approximately 10 weeks after relevant instruction. Each *ACIDI* iteration used the same item order, as six of the nine items were two-tiered questions and needed to be presented in the same sequential order. In addition to the concept inventory administrations, 13 students were individually interviewed for about 30 minutes to gain insight into how the student population may be interpreting *ACIDI* items, item choices and how students were thinking about *ACIDI* material.

The reform-based general chemistry course offered students three discussion-oriented lecture periods per week, for a total of 160 minutes/week, and one cooperative learning workshop (two separate sections) class per week, for a total of 80 minutes/week, for 14 weeks. Lecture and workshop activities followed the *Chemical Thinking* curriculum. Lecture and workshop material (i.e. lecture slides, workshop activities) were matched with *ACIDI* material to determine timing of *ACIDI* administrations. Relevant topics of instruction included:

identification of the most acidic proton, resonance (de)stabilization, and induction. In lecture, students were offered opportunities to participate in discussions, led by the instructor, that allowed students to voice their conceptions, listen to conceptions of their peers, and evaluate the usefulness of all exposed conceptions.

In the cooperative learning workshop, students were placed into groups of 3-4 students. Three workshop groups per section, totaling six groups, were audio and video recorded for the whole semester (14 weeks). All students were given a demographic questionnaire to collect information regarding race, ethnicity, gender, first-generation student status, and first language. Demographic questionnaire responses were used to create diverse workshop groups based on race, ethnicity and gender (Hurtado, 2001). However, most demographic variables were not subject to analysis due to inadequate sample size. Workshop activities were designed to promote student discussion, and offered opportunities for students to argue their answers with group members and the class, through explicit workshop prompting (i.e. identify the most acidic proton in each of these molecules. Justify your choice.). Workshop classes were facilitated by undergraduate and graduate teaching assistants (UGTAs & GTAs), and allowed students to work in their groups for approximately 30 minutes before going over the first half of workshop activity questions for 10 minutes, followed by a second cycle for the second half of workshop activity questions. Students were expected to use information, gained primarily in lecture, in the workshop class and collaborate with group members to complete all workshop questions each week.

After collecting the conceptions held by students before instruction, students were given the opportunity to become aware of useful organic acidity conceptions during relevant lecture and workshop classes of their general chemistry course. If useful conceptions were collected, competed with each other and conceptual hierarchies were restructured, then conceptual reprioritization theoretically occured. Evidence of conceptual reprioritization would then be evident in the change of *ACIDI* responses from pre to post-test. Conceptual retention, or longevity of the restructured conceptual hierarchies, would also be evident in the lack of change in student responses from post to delayed post-test.

Consenting students who completed both the pre and post-test, and the demographic questionnaire were given extra credit equivalent to one full workshop quiz grade due to time equivalency of full participation and one workshop quiz. The workshop quizzes were weekly homework assignments that students were expected to complete before coming to that week's workshop class. Extra credit awarded to fully participating students equated to 0.55% of their overall course grade. A research design summary can be found below in table 1.

Table 1. ACIDI research design summary.

Reform-Based General Chemistry (Week of Fall Semester)	**Action**
6	*ACIDI* Pre-Test
7 & 9	Audio and Video Recording relevant workshop discussions (80 mins/week)
10	*ACIDI* Post-Test
15 & 16	Individual Interviews (30 mins)
2 (Spring)	*ACIDI* Delayed Post-Test

Validity and Reliability

Assessment of content, construct and substantive validity is important, and relevant, to understand if *ACIDI* is truly testing organic acidity concepts (construct), properly testing organic acidity concepts (content), and if students are appropriately interpreting (substantive) *ACIDI* items (Libarkin, 2008, AERA, 2014). Bretz and McClary (2015) previously assessed content and construct validity of *ACIDI* by consulting graduate and undergraduate organic chemistry instructors at a research university and liberal arts college, both in the United States. Construct validity stems from construct validity theory and has been defined as "evaluation of the extent to which a measure assesses the construct it is deemed to measure" (Strauss & Smith, 2009). To assess content and construct validity of *ACIDI* for this study, undergraduate organic chemistry instructors at the university where this study took place reviewed the instrument's questions and structural representations to determine if they are consistent with previous organic chemistry students' course representations and appropriate for their level of knowledge after instruction. In addition, 13 individual interviews of the reform-based students were audio and video recorded, then transcribed and probed to assess students' interpretations of the *ACIDI* questions to determine substantive validity.

Previously, Bretz and McClary (2015) had assessed reliability of *ACIDI* using Cronbach's alpha, while cautioning the use of traditional forms of reliability assessment due to the nature of concept inventories. Bretz and McClary (2015) received Cronbach's alpha values ranging from 0.39 - 0.54. Although these values are low by conventional Cronbach alpha standards, it is important to keep in mind that Cronbach's alpha is more of a confirmatory measure of dimensionality (Tavakol & Dennick, 2011) and is heavily influenced by the number of questions an instrument contains. Therefore, it is less likely to achieve high Cronbach alpha values from *ACIDI*, as the instrument contains a mere nine questions. *ACIDI* was subject to Rasch analysis reliability and validity measurements to determine reliability and validity of *ACIDI*. Reliability was assessed using the Rasch dichotomous model, and validity was assessed using dimensionality assessment. Dimensionality assessment of an instrument suggests latent constructs tested by the instrument, providing evidence that may or may not support the validity of the instrument (Linacre, 2018).

Reliability and validity was assessed using the Rasch dichotomous model rather than using Cronbach's alpha to assess reliability due to its primary use a confirmatory indicator of reliability, as opposed to a determinant indicator of reliability, and the concept inventory nature of *ACIDI*. A suggestive indicator of dimensionality, such as principal component analysis, is capable of measuring latent constructs that the instrument tests to assess its validity. The Rasch dichotomous model was used because all student responses were converted to a dichotomous (correct/incorrect) data set, in Winsteps version 3.68.2. The ITEM: dimensionality function was used to determine dimensionality of the instrument via principal components analysis (PCA). Dimensionality of *ACIDI* was assessed via PCA of residuals in the first contrast to determine if student responses to *ACIDI* assessed more than one latent construct, which would suggest that *ACIDI* tests more than one facet of chemistry, even though it was created to only test the single faacet of organic acidity (Bretz & McClary, 2015; Wold, Esbensen & Geladi, 1987) Eigenvalues, or latent roots (Marcus and Minc, 1965), are capable of suggesting the number of dimensions, or latent variables, tested by instruments. Eigenvalues greater than 2 indicate the presence of multiple dimensions, or multiple latent variables tested by the instrument. Presence of eigenvalues less than 2 suggest unidimensionality, or one latent variable tested by the instrument (Boone, Staver & Yale, 2013).

Person and item reliability scores were also investigated using the Rasch Dichotomous model in Winsteps version 3.68.2. Person reliability is an indicator of an instrument's ability to separate the test population into levels, and depends on sample variance, number of test items, categories per item, and sample-item targeting (Linacre, 2018). Item reliability is an indicator of the test population's capability of addressing items on the instrument's latent variable, and depends on sample size and item difficulty variance (Linacre, 2018). Person reliability scores may be affected in a similar manner as Cronbach's alpha, as *ACIDI* has only 9 items and different categories (conceptions) tested per item, which may lead to a lower person reliability score. Students who scored 100% or 0% were excluded from analysis due to inability of assessing person measurement error, as both of their errors are considered to be infinite. If a student correctly answers all *ACIDI* items, knowledge beyond what *ACIDI* tests regarding organic acidity cannot be determined. Similarly, students who incorrectly answer all *ACIDI* items, untested knowledge by *ACIDI* items regarding organic acidity cannot be assessed. Person and Item reliability scores range from 0, least reliable, to 1, most reliable (Boone, Staver & Yale, 2013).

Data Sample

Consenting students belonged to an advanced sequence, first-year, reform-based general chemistry curriculum course. The reform-based general chemistry course consisted of 179 students who were given the *ACIDI* concept inventory three times over the course of 12 weeks, across two semesters. The advanced sequence is expected to be composed of students who have taken at least two years high school chemistry (e.g., introductory high school chemistry with additional Advanced Placement or International Baccalaureate chemistry). Of the 179 total students enrolled in the course, 107 (59.78%) students consented and took the pre and post-tests to completion, and 86 (48%) students consented and took the pre, post and delayed post-tests to completion. Consenting students, who completed the pre and post-test to completion, were disaggregated into gender and generation demographics. Of the 107 pre-post participants, 44 (41%) reported as male, 37 (35%) reported as female, and 26 (24%) did not report a gender. With regard to student generation, 19 (18%) reported as a first-generation student, 58 (54%) reported as a non-first-generation student, and 30 (28%) did not report their student generation status.

Sample Representation Assessment

Consenting students' representation of their general chemistry class was assessed based on consenting students' pre and post-test scores and pre and post-test scores of the whole class. The post-test was used as an assessment of representation of student's knowledge regarding *ACIDI* after instruction. The pre-test was used as an assessment of representation of students' knowledge regarding *ACIDI* before instruction. An independent samples t-test, assuming equal variances, was performed on pre and post-test mean scores.

Analytical Framework

Classical test theory has been commonly used to analyze chemistry concept inventory scores due to its robustness in interpreting results from its analyses, and its relatively simplistic assumptions (DeVellis, 2006; Bretz & McClary, 2015; Brandreit & Bretz, 2014; Voska & Heikkinen, 2000; Prince, Vigeant & Nottis, 2012; Bradley & Mosimege, 1998; Kousathana, Demetouti & Tsaparlis, 2005; Rahayu et al. 2011). However, a major implicit assumption within classical test theory, relative to instruments such as concept inventories, is that all items on the instrument contribute to the final score equally, also known as the assumption of parallel items (DeVellis, 2006). For example, on an exam, if a student is to arbitrarily get 2 questions (#'s 1 and 3) correct out of 5 total questions, and another student is to arbitrarily get a separate 2 questions correct (#'s 2 and 4), both scores (2/5) are equivalent. This assumption also determines that the difference in student ability between a 1/5 and 2/5 is equivalent to the difference between a 4/5 and a 5/5, due to the equivalent differences among both pairs of reported scores being 1 question. However, this study does not assume this to be true, as suggested by previous users of item-response theory rather than classical test theory (Wei et al. 2012; Pentecost & Barbera, 2013; Barbera, 2013), as there are questions within the *ACIDI* concept inventory that are more difficult than others due to the sophistication and variation of problem solving skills needed to answer different questions. For example, the difficulty in answering a ranking question correctly on the inventory is easier than answering a reasoning question. This is also held true by the probability of guessing an item correct. A student has a 50-50 chance for correctly guessing a ranking question, due to two possible answer choices, and a 25% chance of correctly guessing a reasoning question, due to four possible answer choices. Furthermore, if an instrument is intended to return a stratification of student scores to distinguish high from low performing students, then the questions within the test ought not be of the same difficulty. These questions should aim to be of equivalent difference in difficulty to create a hierarchy of questions that appropriately separates high performing from low performing students. If multiple items are of the same difficulty, they may be redundant. However, the determination of redundancy on a concept inventory should be contextualized for each particular question (Boone, 2016). It is suggested to remove a redundant question if (1) the question has an equivalent item difficulty as another and (2) targets the same concept/topic (Boone, 2016). Previous research regarding *ACIDI* has not established this hierarchy or presented reported varying item difficulties for *ACIDI* items (Bretz & McClary, 2015), therefore rendering the operation under the assumption of parallel items within classical test theory false. This creates a need to look elsewhere to create a more meaningful way of measuring and interpreting student scores on *ACIDI*. Similar thoughts have been expressed and investigated for the Force Concept Inventory, displaying the benefits of using Item-Response Theory rather than Classical Test Theory, as Item-Response Theory may remove the factor of guessed responses, account of unequal item difficulties, and provide more meaning to student scores (Bao, 2010).

To alleviate this issue, this study used Rasch analysis to assess the presence of an equivalent ruler, or scale, by which to measure differences among student ability on *ACIDI* (Boone, 2016). Rasch analysis, a form of item response theory (IRT), aims to create this ruler, or scale, based on the natural log probability of a student answering any given question correctly and item difficult. The 'tick marks' of this ruler are manifested as logit scores (item difficulties), and are used to measure person ability. A logit score is a log odds unit score (eq. 1 below). Person ability logit scores averages of individual logit scores that are based on the probability of a student correctly answering each question on the testing instrument. Item difficulty logit scores are individual logit scores based on item difficulty and person ability scores. The formula by which Rasch analysis is based on is shown below (Boone, 2016).

1. $B_i - D_i = \mathit{ln}\ (P / 1\text{-}P)$

Where B_i is the person ability score, D_i is item difficulty, and P is the probability of answering an item correctly. On instruments that have multiple items, the analysis will return average logit scores that represent students' probabilities of answering an average question correctly. The logit scale may average item difficulty or person ability to a logit score of 0, depending on if the researcher decides to constrain items or persons in their analysis. Constraining items creates an average item difficulty of 0, therefore allowing the average person ability to change in a repeated measures study, which in turn provides the researcher with evidence of person ability change. Constraining means to restrict values so they average to 0, allowing you to focus changes in another variable's values to interpret change more easily. Constraining persons sets the average person ability to 0, therefore allowing the average item difficulty to change in a repeated measures study, which in turn provides the researcher with evidence of item difficulty change (Boone, 2016). This creates an easy to interpret manner of students' probabilities of answering any question on an instrument or changes in item difficulties for tested students. For example, assuming constraint of items, if the average person ability score for an instrument is 0 (for simplicity), then that student has a 50% chance of correctly answering an average item correctly (average Item difficulty logit = 0). Comparisons can then be done to get a quick idea of the probability of that same student to correctly answer other questions along the item difficulty logit scale. For example, keeping the assumption of the students' average person ability score of 0, if an item has a difficulty logit score of -1, then the student has a greater than 50% chance of correctly answering that particular question because 0 is greater than -1. If an item has a logit score of 1, then that same student has a less than 50% chance of correctly answering that particular question, because 0 is less than 1. This trend of comparison is held throughout the logit scale because it is centered around the comparison of logit person ability and item difficulty logit scores. For example, if a respondent has a person ability score of logit = 1, then that student has a 50% chance of answering an item of logit = 1, greater than 50% chance of answering items of logit < 1, and less than 50% chance of answering items of logit > 1 (Boone, 2016).

Rasch analysis also allows instructors/administrators to appropriately determine if the test, or questions within the test, are appropriate for the students being tested (Boone, 2016). Determination of a test being (in)appropriate for a population of students can be done through comparing the distribution of item and student logit scores, which would also be evident in item outfit scores (Boone, Staver & Yale, 2013). Item outfit scores are indicators of an item's ability to fit a model, and they can overfit or underfit that model. Item outfit scores greater than 1 suggest underfit, and excess unexplained variance by your model, and outfit scores less than 1 suggest overfit, and risk overprediction by your model and possibly exaggerated reliability statistics (Boone, Staver & Yale, 2013). If the distribution of student ability covers the distribution of item difficulty, then the test is appropriate for the students being tested, and the item outfit scores fall within the acceptable range of 0.5-1.5 (Boone, Staver & Yale, 2013), although it is common practice to use an item outfit range of 0.7-1.3 (Boone, 2016). If there are a few questions that are above or below the distribution of student ability, they may be inappropriate for the students as they may be too easy (below the distribution), or too difficult (above the distribution), making the ability to separate high from low performers less efficient. Similar to outfit scores for items, outfit scores of persons can also be determined using Rasch analysis (Boone, Staver & Yale, 2013). Person outfit scores use the same range of 0.5-1.5 for their data to be considered "productive for measurement," and follow the same trends as item outfit scores (Boone, Staver & Yale, 2013). Outfit scores that lay outside the aforementioned ranges may be removed from analysis to increase quality of results, however the removal of persons may impact statistical power of results due to a reduction in sample size (Boone, Staver & Yale, 2013).

For this study, Rasch analysis was used to measure the change in students' probabilities of correctly answering questions on *ACIDI* before and after relevant instruction. This can be done through anchoring post-test item difficulty to the pre-test analysis. Anchoring post-test item difficulty to the pre-test means to keep the same post-test item difficulty scores for both the pre and post-test This creates a consistent scale for both pre and post-test item difficulty logit scores. Therefore, person ability changes along the scale (pre to post-test) are appropriately representing a change in person ability after instruction. The post-test item difficulties are used in this test-retest method to evaluate instruction because the post-test item difficulties are more indicative of a prepared student to answer instrument items, after being taught the appropriate material. The function of *ACIDI* is to be able to indicate a student's organic acidity conceptions, including scientifically accepted organic acidity conceptions based on correctly answering *ACIDI* items. Therefore, the person ability logit score can be interpreted as the student's probability of holding a scientifically accepted conception, dependent on the question at hand. These changes can best be exemplified using a wright map. A wright map displays person ability and item difficulty logit scores adjacent to each other, lending simplicity to interpretation of learning gains (shown below in figure 2).

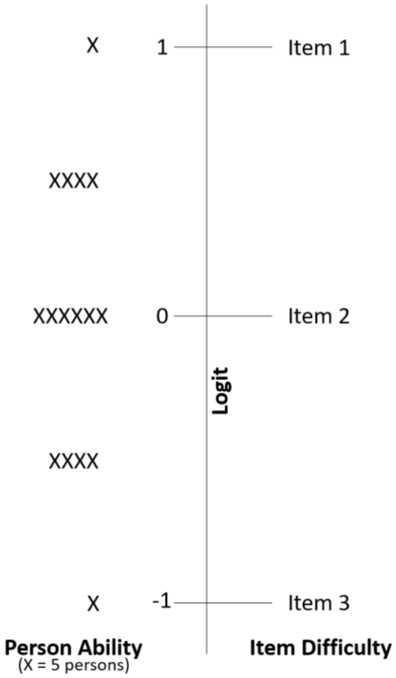


Figure 2. Sample wright map.

This study used Winsteps version 3.68.2 and Rstudio version 1.1.447 to perform Rasch analysis to analyze student responses & potential conceptions, determine person ability, and item difficulty of *ACIDI* items. IBM SPSS Statistics version 23 was used to perform inferential and descriptive statistical analyses on item responses and person ability logit scores. Microsoft excel version 16.0.9126.2152 was used to calculate Cohen's d for effect size to provide meaning to differences in person ability and item performance. Individual student interviews were performed for approximately 30 minutes each to assess substantive validity of *ACIDI* and get an understanding of how students think about questions, choices, and topics covered by the concept inventory.

Results

Validity and Reliability

To assess construct and content validity of *ACIDI*, organic chemistry instructors at the university where the study took place were consulted. They had asserted that the questions and structural representations of organic molecules within the instrument were scientifically correct and appropriate for a student population after relevant instruction and that each question was appropriately targeting its intended conception. Individual student interviews revealed students' interpretations of questions were consistent with the wording and intentions of each question, suggesting sufficient substantive validity of *ACIDI*. All interviewees had expressed they gained knowledge to answer the *ACIDI* items from the lecture or workshop section of the reform-based general chemistry course (See Supplemental).

To assess dimensionality of *ACIDI*, principal component analysis of residuals was performed in Winsteps version 3.68.2. The returned eigenvalue of the first contrast was less than 2, suggesting the instrument is unidimensional. Rasch dichotomous model analysis returned a person reliability value of 0.24 and item reliability value of 0.95, indicating low person reliability and high item reliability.

Sample Representation

An independent samples t-test, assuming equal variances, was performed on the consenting students' pre and post-test scores and the whole class's pre and post-test scores. The results of the independent samples t-tests indicated that the consenting students' scores were not significantly different than the class as a whole for the pre (df = 255, t-stat = 0.133, p = 0.894) and post-test (df = 255, t-stat = -0.042, p = 0.967). This serves as evidence that supports representativeness of consenting students for their whole class.

ACIDI Performance

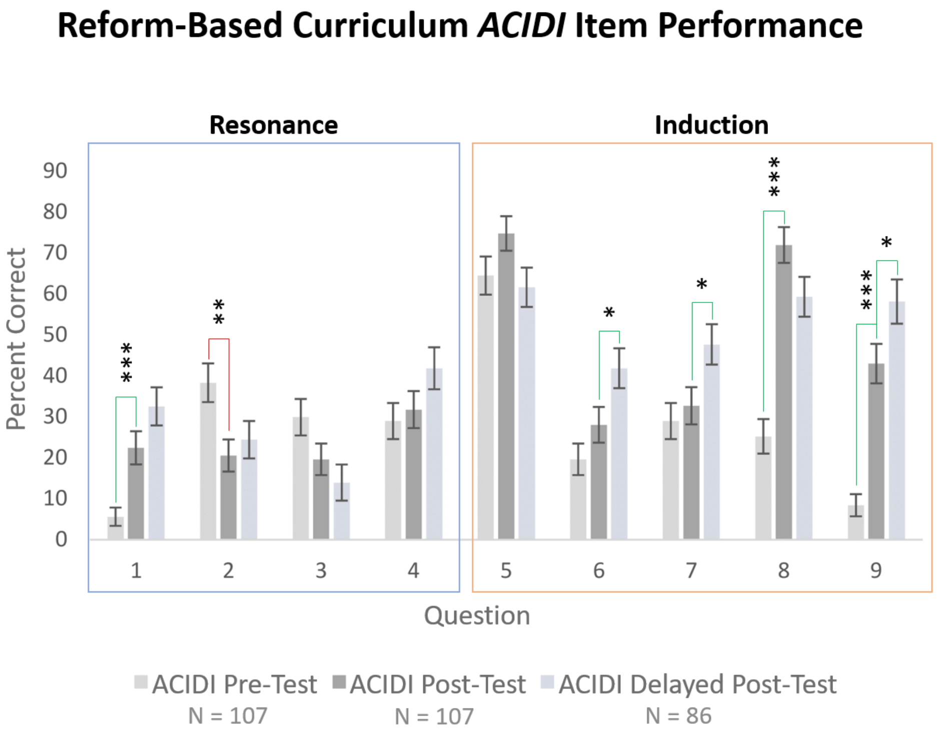


*Figure 3. ACIDI pre, post and delayed post-test item-level performance. *p<0.05, **p<0.01, ***p<0.001.*

Student performance at the item level for the pre, post and delayed post-test is summarized above in figure 3. From pre to post-test, students significantly increased their performance on 3 of the 9 *ACIDI* items (1, 8 and 9), while significantly decreased scores on one item (2). From post-test to delayed post-test, a period of no relevant course instruction, students significantly increased performance on 3 of the 9 *ACIDI* items (6, 7 and 9), while no significant decrease in item performance was observed.

*Table 1. ACIDI pre, post and delayed post-test item-level performance. Effect sizes calculated using Cohen's d. *p<0.05, **p<0.01, ***p<0.001. Pre/Post N = 107, Delayed-Post N = 86.*

Reform-Based Curriculum *ACIDI* Item Performance

Question	$\bar{X}_{Pre}$ (%)	$\bar{X}_{Post}$ (%)	$\bar{X}_{DPT}$ (%)	T-Stat$_{Pre\text{-}Post}$	P$_{Pre\text{-}Post}$	P$_{Post\text{-}DPT}$	$\delta_{Pre\text{-}Post}$	$\delta_{Pre\text{-}DPT}$
1	5.61	22.43	32.56	3.738	<0.001***	0.116	0.45 (S)	0.66 (M)
2	38.32	20.56	24.42	-2.931	0.004**	0.525	0.31 (S)	0.24 (S)
3	29.91	19.63	13.95	-1.828	0.070	0.301	0.19	0.31 (S)
4	28.97	31.78	41.86	0.446	0.657	0.149	0.05	0.22 (S)
5	64.69	74.77	61.63	1.939	0.055	0.050	0.18	0.05
6	19.62	28.04	41.86	1.531	0.129	0.044*	0.16	0.42 (S)
7	28.97	32.71	47.67	0.684	0.495	0.035*	0.07	0.31 (S)
8	25.23	71.96	59.30	7.662	<0.001***	0.065	0.86 (L)	0.61 (M)
9	8.41	42.99	58.14	6.489	<0.001***	0.037	0.77 (L)	1.11 (L)

A summary of performance differences between all three iterations of *ACIDI* is shown above in Table 1. Item performance changes from pre to post-test uncovered effect sizes between 0.05 and 0.86. Effect sizes for items in which students significantly improved from pre to post-test ranged from 0.45 (small) to 0.86 (large). Item 2, the only item that students significantly decreased performance from pre to post, had an effect size of 0.31 (small). Item performance changes from pre to delayed-post-test reveal effect sizes between 0.05 and 1.11. Effect sizes for items in which students significantly increased from pre to delayed post-test ranged from 0.22 (small) to 1.11 (large). The effect size for item 2 remained small, while decreasing from 0.31 (pre-post) to 0.24 (pre-delayed post).

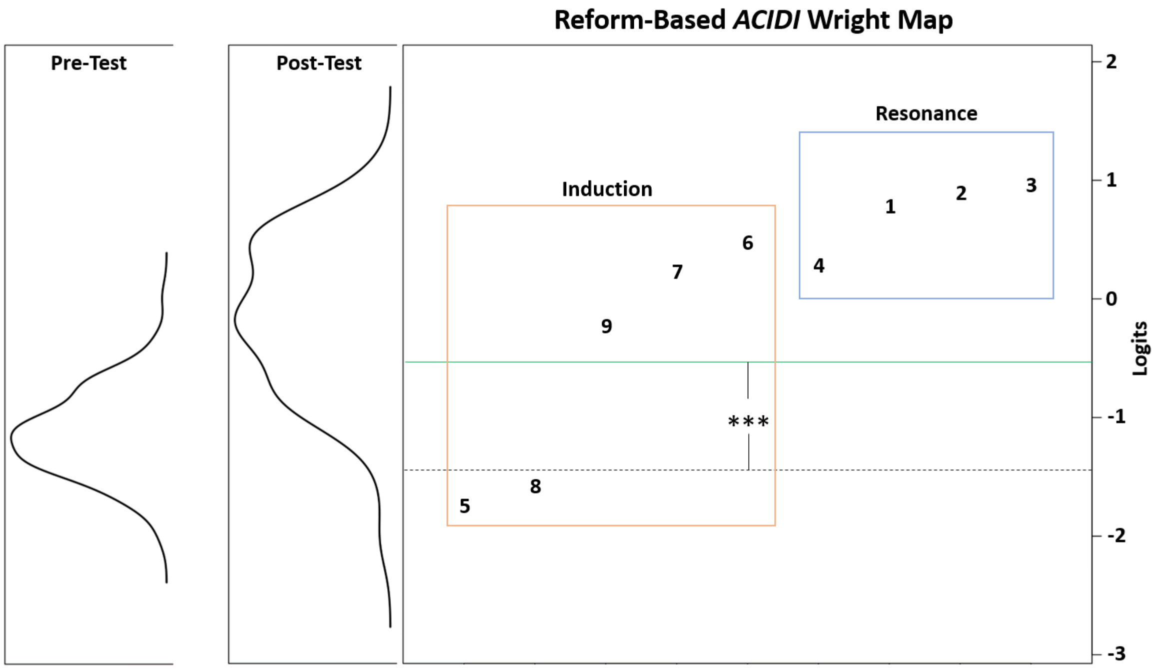


*Figure 4. ACIDI Pre-post student person ability wright map comparison. ***p<0.001. Pre/Post N = 107.*

Student person ability is depicted in the Wright map above (Figure 4). Similar to the wright map presented in the methods (Figure 2), person ability distribution (left half of Figure 4) and item difficulty (right half of Figure 4) can be seen along a logit scale. Pre-test mean student ability significantly increased from -1.1761 to -0.5614 on the post-test (p<0.001). Item difficulty logit scores ranged from -1.7 (easiest question) to 0.93 (most difficult question). Students' probability of answering an average *ACIDI* item (logit = 0) increased from 23.58% (mean pre-test person ability logit = -1.1761) to 36.32% (mean post-test person ability logit = -0.5614). Effect size for pre-post person ability change was 0.61 (medium). The distribution of post-test person ability was larger than the *ACIDI* item difficulty distribution. 4 of the 5 induction questions (5, 8, 9 and 7) lie below all resonance questions on the logit scale, while one induction question (6) rests above the least difficult resonance question (4). All item outfit scores lie within the range of 0.5-1.5. WLE (weighted likelihood estimate) reliability assessment for pre-test person ability was 0.37, and 0.29 for post-test person ability. Item outfit statistics can be found in the appendix.

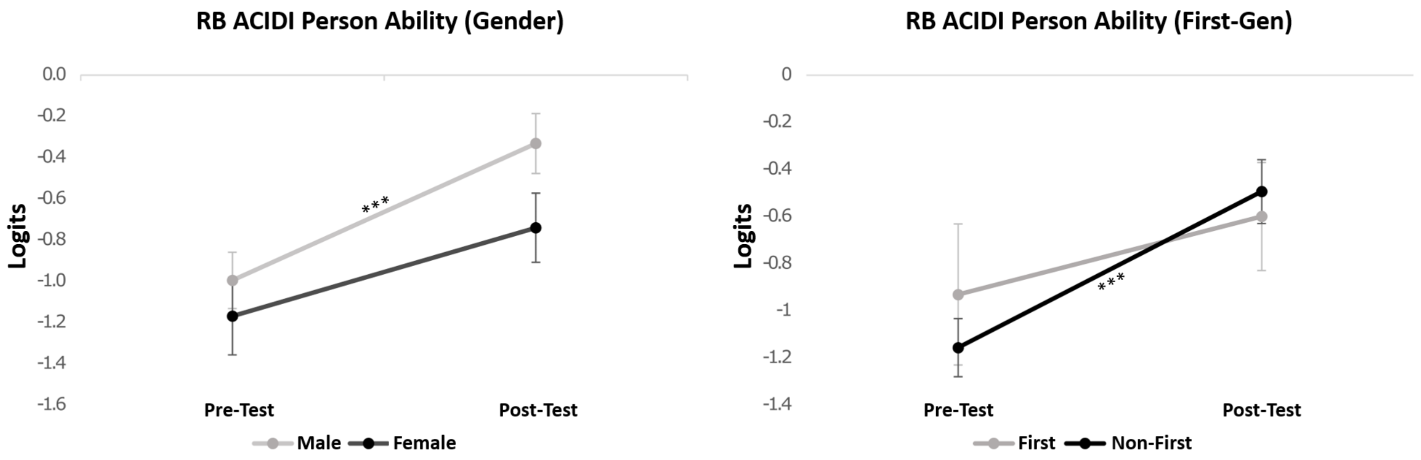


*Figure 5. Student pre-post ability gain comparisons based on gender and student generation. ***p<0.001. N (Male) = 44, N (Female) = 37, N (FG) = 19, N (NFG) = 58.*

Disaggregation of student ability in gender and generation demographics, shown above in figure 5, revealed significant person ability increase from pre to post-test on *ACIDI* in males and non-first-generation students, while females and first-generation students did not significantly improve their person ability scores. Males significantly increased their person ability from -0.9952 (*ACIDI* pre-test) to -0.3317 (*ACIDI* post-test), while non-first-generation students significantly increased their person ability from -1.1559 (*ACIDI* pre-test) to -0.4938 (*ACIDI* post-test). Male *ACIDI* pre and post-test person ability scores were not significantly different than female *ACIDI* pre and post-test person ability scores (p_{pre} = 0.45, p_{post} = 0.07). Similarly, first-generation students' *ACIDI* pre and post-test person ability scores were not significantly different than non-first-generation students' *ACIDI* pre and post-test person ability scores (p_{pre} = 0.41, p_{post} = 0.70).

Implications and Discussion

Validity and Reliability

Dimensionality assessment suggests *ACIDI* is a unidimensional instrument, which Bretz group had previously suggested, claiming organic acidity, and more specifically, conjugate base stability, as the main topic being tested.

All person reliability indicators point towards low reliability of results. However, this may be due to a low sample size of both people and items. Appropriate sample size for persons in Rasch analysis may range from 50 persons to 200 persons, while appropriate sample size for items seems to be greater than 10 (Boone, Stavel & Yale, 2013). While the sample size of this study seems to be appropriate for Rasch analysis, the low number of *ACIDI* items may not be appropriate, and strongly affect reliability calculations, and should be considered when interpreting Rasch reliability markers for *ACIDI*. The nature of this study should also be considered when interpreting reliability with regard to person sample size, as the sample of students for this study is limited to the number of students who enroll and persist through the fall semester, into the following spring semester course.

ACIDI Performance Change

Significant performance increase on questions 1, 8 and 9 (Figure 3) suggest conceptual learning gains, and provide evidence for conceptual reprioritization of dominant, scientifically accepted, induction conceptions. All significant item-level gains were accompanied by large effect sizes (table 1), adding meaning to the conceptual learning gains associated with induction, and suggest strong conceptual reprioritization towards scientifically accepted conceptions of induction. These conceptual learning gains may be attributed to *Chemical Thinking*'s teaching methods regarding induction (i.e. lecture and workshop discussions), as all interviewees had suggested they gained the knowledge to answer the *ACIDI* items from lecture or workshop. Conversely, a significant decline in percent correct on question 2 (see supplemental for *ACIDI* items) suggests a lack of conceptual reprioritization towards scientifically accepted conceptions of resonance stability. Rather, this decrease may suggest conceptual reprioritization towards dominant alternative conceptions regarding resonance stability. An effect size of 0.88 adds insight to this lack of student understanding, suggesting resonance stability as a possible topic of interest that may need to be more carefully presented and thoroughly addressed through discussion. It is also possible that these students needed more time to digest and develop the chemistry skills necessary to correctly answer question 2.

A lack of significant difference between item-level scores on *ACIDI* from post to delayed post-test suggest retention of dominant, scientifically accepted, conceptions of induction. And in some cases, increased conceptual reprioritization towards dominant, scientifically accepted, conceptions of induction with time after instruction. By the same token, a sustained lack of correct answers, and potential presence of dominant alternative conceptions, regarding question 2 is suggested. The significant increases over time at the item level for questions 6, 7 and 9 may be due, although not guaranteed, to the increased amount of time the students had to develop their conceptions regarding induction, allowing for scientifically accepted conceptions to dominate.

Rasch Analysis Results

These conceptual learning gains, or conceptual reprioritizations, can be more meaningfully interpreted via the *ACIDI* wright map of student ability and *ACIDI* item difficulty (figure 4). A comparison of mean person ability logit scores to individual question difficulties on *ACIDI* can be used to gauge what questions the reform-based students have the best chance of answering correctly. Reform-based students had difficulty answering seven of the nine *ACIDI* items, as mean post-person ability was below seven item difficulty scores. The majority of these difficult questions targeted conceptions of resonance (i.e. resonance stability), suggesting resonance as a difficult concept for students to reprioritize their dominant conceptions towards scientifically accepted conceptions of resonance. Similarly, reform-based students had difficulty reprioritizing their dominant conceptions towards scientifically accepted conceptions of induction, as three of the five resonance questions lie above mean post-test person ability. This suggests that concepts tested by these questions as concepts that may need to be discussed more for desired conceptual reprioritization towards dominant, scientifically accepted, conceptions of induction to occur. Using this information, it can be concluded that the *ACIDI* concept inventory was a difficult assessment for the consenting students.

Although *ACIDI* may be difficult for this student population, the distribution of student ability suggests that the inventory is appropriate for these students after instruction, as student post-test ability extends both above and below the distribution of item difficulty, meaning no question on *ACIDI* is too hard or too easy. Furthermore, the *ACIDI* post-test may be best used, as it’s currently constructed, to distinguish high performing students from low performing students, while having difficulty distinguishing those in between. This is supported by the distribution of items across the logit scale, as seven of the nine items on the assessment lie above the average logit score of post-test person ability. Which implies that you must be an above average performer on this assessment to separate your person ability score from your peers by correctly answering questions that are above average in difficulty. The large gap between questions 8 and 9 on the logit scale (logit difference = 1.28), is an area along the scale that is less capable of distinguishing student ability. This inability to distinguish students between the logit values of -1.55 and -0.27 is analogous to a ruler that has two marks of one inch and two inches, no marks in between, and marks below 1 inch and above 2 inches. You may approximate the length of objects that lie in between one and two inches but are unable to accurately determine the length of those objects in the absence of those marks. While you're more capable of measuring objects that are less than one inch and greater than 2 inches, as there are marks present. This translates to an increased ability of appropriately distinguishing students who have logit values less than -1.55 and greater than -0.27.

Item-level and Rasch analysis results provide an answer to the second research question, suggesting *ACIDI* can detect changes in undergraduate general chemistry students' conceptions of organic acidity, and to the degree of *ACIDI*'s capability of distinguishing levels of student performance, as *ACIDI* seems to be capable of distinguishing high and low performing students, while having difficulty distinguishing those in between.

Disaggregated Scores into Gender and Generation

Disaggregation of person ability scores into gender and generation revealed an absence of student achievement gaps between males and females, and first-generation and non-first-generation students. The lack of achievement gaps pleasantly contrasts previously reported gender achievement gaps in science achievement (Nosek et al. 2008; Stephens, Hamedani & Destin, 2014; Warburton, Bugarin & Nunez, 2001). Although there were no significant differences between male and female, and first-generation and non-first-generation students' *ACIDI* pre and post-test student ability, significant increases in person ability from pre to post-test only occurred for males and non-first-generation students. This is an important distinction to note, as *Chemical Thinking* may have played a different role in conceptual reprioritization for males and non-first-generation students compared to females and first-generation students. Furthermore, first-generation and female students may need more discussion, resulting from *Chemical Thinking* instruction, for their person ability scores to significantly increase

Student Interview Insight

All student interview answers regarding where students gained the knowledge to best answer all questions on *ACIDI* suggest that learning gains may likely be attributed to the lecture or workshop portion of the reform-based general chemistry course. This, paired with evidence that suggest conceptual reprioritization towards dominant, scientifically accepted, conceptions of induction and dominant alternative conceptions of resonance, provides an answer to the first research question. It seems that *Chemical Thinking* has helped reform-based general chemistry students adopt dominant, scientifically accepted conceptions of induction, and dominant alternative conceptions of resonance.

Summary

ACIDI can detect changes in undergraduate general chemistry students’ conceptions of organic acidity after *Chemical Thinking* instruction. Furthermore, an association can be made between the *Chemical Thinking* curriculum and conceptual reprioritization of first-year, reform-based general chemistry students’ scientifically accepted conceptions of induction and resonance. This attribution may also provide insight into the effectiveness of *Chemical Thinking*’s impact on reform-based students’ conceptions of resonance stability and suggest that future instruction view this as a sub-topic that may need increased class time for reform-based students to more appropriately gauge and compare the stability of various conjugate base structures. In contrast, an association can be made between *Chemical Thinking* and dominant, scientifically accepted, conceptions regarding induction. Finally, *ACIDI* is appropriate for the tested students, however may be best used to distinguish high and low performing students, while having difficulty to determine those in between. Disaggregation of student ability scores revealed the absence of an *ACIDI* achievement gap between males and females, and first-generation and non-first-generation students.

Limitations

This study was limited to the population of consenting students that took the *ACIDI* concept inventory and are not generalizable due to a lack of representative assessment of participating students and the general student population outside of the university where this study took place. Furthermore, the accelerated student population may play a role in the conceptual reprioritization process, and may not be indicative of conceptual reprioritizations that take place in a non-accelerated student population. The non-randomized sample used in this study limits the representativeness of consenting students to their peers. Interpretation of conceptual reprioritization is limited to the same order in which the *ACIDI* items were presented during each iteration. Lack of item order rearrangement risks conceptual repriortizations to be impacted by problem similarity. The interpretation of conceptual reprioritization is limited to the alignment of students' conceptions with the conceptions represented by the choices within *ACIDI*. Furthermore, the interpretation of student achievement is limited to the ability of *ACIDI* to be able of return results that are appropriately assessing student achievement (i.e. ability of distinguishing high from low performing students).

Future Direction

Future researchers are urged to increase sample size through increasing the number of students who consent, and participate to completion, through aggregation of samples from numerous years or increasing the number of involved chemistry classes. Increasing sample size by increasing the number of classes, both treatment (reform-based curriculum) and control (traditional curriculum), may be possible if propensity score matching, a statistical method used to reduce selection bias, is used to compare student ability at the individual level across classes to assure that the student populations are similar enough to group together for further analysis (Caliendo & Kopeinig, 2008). Collection of student ability indicators, outside of rasch analysis (i.e. undergraduate GPA, high school GPA, placement exam scores, etc.), may also add insight to the strength by which we can attribute conceptual reprioritizations to the course curriculum/instruction.

Modification of the *ACIDI* concept inventory may also be of interest to maximize reliability of the assessment and the results of analyses that use *ACIDI* as a measurement tool. The starting point of modification to the inventory can be reverse engineered based on the goal of the use of the instrument. Part of that goal should be for the students to reach an established standard. For example, if *ACIDI* is intended to measure students' scientifically accepted conceptions, and the score on the instrument is used to be predictive of successful performance (the basis of the standard, i.e. passing/other letter grades) in an organic chemistry class, then it may be best to determine what *ACIDI* scores students who have succeeded in an organic chemistry course received (the desired standard).

The construction, or re-construction of the testing instrument, should be able to distinguish students who are able to meet that standard from those who do not (to have high predictive value). That can be done through continual refinement of questions such that the resulting questions of a predictive assessment have equal, or close to equal, differences in item difficulty (Even marks along a scale). Once those tick marks are established (continuing the organic chemistry example), then provide the assessment to students and proceed to reverse engineer a desired standard, *ACIDI* percent correct or person ability, for tested students to meet. After the establishment of the standard, evaluation of instructional methods or curricula can be performed to determine which is most effective in elevating students to the standard.

The *ACIDI* concept inventory, as it is currently constructed, may not be optimal for reaching an established standard for two reasons. First, *ACIDI* currently has nine items, decreasing its ability to optimally distinguish student ability along a logit scale due to a relatively small number of marks along a scale. Second, as the inventory is currently constructed and previously analyzed, there is only evidence of this study population's item difficulty distribution, or one logit scale that measures undergraduate students' conceptions of organic acidity. That scale contains a gap in measurement, which is not ideal for the person ability separation that may be desired in an instrument designed to predict college students' ability to reach an established standard. These two reasons ought to be addressed before moving forward and using *ACIDI* as a predictive measurement tool.

Chapter 3: Conceptual Reprioritization of Redox Using Different General Chemistry Curricula

Introduction

Background

Undergraduate students' conceptions of oxidation-reduction, have been commonly investigated from a symbolic perspective using chemical equations and less so from a particulate perspective that visualizes electron transfer and dynamics of reaction processes (Brandreit & Bretz, 2014; Rosenthal & Sanger, 2012; Garnett & Treagust, 1992). Previous research has uncovered some misconceptions held by students from both perspectives, as some students have inappropriately assigned oxidation numbers to whole molecules, posited that electrons may move freely, and independent of ions through solution and reliance on surface features of chemical equations when classifying reactions (Stains & Talanquer, 2008; Garnett & Treagust, 1992; Rosenthal & Sanger, 2012). These misconceptions may be due to difficulty in connecting different perspectives (i.e. particulate and symbolic) within chemistry, as students may not be appropriately thinking, or visualizing, about such chemical reactions, especially within the domain of redox (Brandreit & Bretz, 2014).

Cooperative learning has been commonly used to improve student achievement in academic settings (Freeman et al. 2014), however decreasing learning outcomes have also been reported (Bowen, 2000), urging researchers to further investigate the impact of cooperative learning on student outcomes. Reported increases in conceptual understanding, content retention, and student performance in chemistry have been linked with cooperative learning (Bowen, 2000;

Freeman et al. 2014; Shah et al. 2018). Cooperative learning promotes student interaction through the formation of small groups. These small groups offer students to discuss and argue class material, which has been linked to conceptual understanding and may lead to conceptual reprioritization (Nussbaum, Sinatra & Poliquin, 2008; Shah et al. 2018; Bowen, 2000). Argumentation may lead to conceptual reprioritization through student exposure to useful conceptions (Shtulman & Lombrozo, 2016; Lombrozo & Carey, 2006; Lombrozo, Kelemen & Zaitchik, 2007). Conceptual reprioritization, the restructuring of conceptual dominance hierarchies, may take place after useful conceptions compete for dominance within the individual. Diverse learning groups, which may be created in a cooperative learning setting, have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001), and may contribute to higher quality arguments and an increase in exposure to useful conceptions. Erduran's analytical framework for argumentation quality may be used to assess quality of arguments made in cooperative learning environments, and allow for investigation of potential association between argumentation quality and conceptual reprioritization (Erduran, Simon & Osborne, 2004). Eruduran's analytical framework assess quality of vocalized arguments from level 1-5, based on the presence of basic argumentation components (coded via Toulmin's Argumentation Pattern, or TAP). Increasing argumentation quality occurs as students vocalize more claims, data, warrant, and rebuttals. Increased exposure to useful conceptions may occur as you increase argumentation quality from level 1 to 5 (Shah et al. 2018), and perhaps increase odds of conceptual reprioritization. However, there is no literature regarding an impact, or lack thereof, that is made by argumentation on undergraduate students' conceptual reprioritization of of redox conceptions. A potential obstacle towards investigation of argumentation and conceptual reprioritization of undergraduate students' redox conceptions may be the lack of alignment of common general chemistry curricula and instruments designed to measure undergraduate student redox conceptions.

Traditional, college level, general chemistry curricula have been reported to be composed of disconnected theories that emphasize discrete facts and algebraic proficiency of undergraduate students over conceptual understanding (Talanquer, 2013; Van Berkel et al. 2000; Kuhn, 1963; Bulte et al. 2006; Shah et al. 2018). This emphasis on algebraic proficiency does not align with cooperative learning environments that may promote conceptual reprioritization. *Chemical Thinking*, a reform-based general chemistry curriculum, created by Talanquer and co-workers, has shifted the focus from algebraic proficiency to conceptual understanding by organizing curricular topics around fundamental, disciplinary questions (i.e. How do we distinguish substances?). This question-driven focus promotes learning beyond algebraic proficiency and towards conceptual understanding of chemistry material through a discussion-oriented lecture, led by the instructor, that offers students opportunities to voice their conceptions, listen to conceptions of others and evaluate the efficacy of their conceptions of chemistry (Talanquer, 2013; Sevian & Talanquer, 2014; Talanquer, 2018). Peer-discussion and changes in student conceptions have shown to be as valuable as direct instruction, such as lecture (Nussbaum, Sinatra & Poliquin, 2008).

Chemistry education researchers have recently began developing larger quantities of valid instruments that may captures student conceptions called concept inventories (Libarkin, 2008; Brandreit & Bretz, 2014). These concept inventories attempt to capture student conceptions through multiple-choice questions that target many conceptions held about particular topics of chemistry (Libarkin, 2008). These measurement tools have been created with the intent to detect scientifically accepted conceptions, appropriate models of thinking determined by the scientific community, and alternative conceptions, models of thinking that aren't well accepted by the scientific community (Brandreit & Bretz, 2014). Theoretically, these instruments may be used to detect particular student conceptions, determine which is most prevalent among a population of students, and measure the change(s) in conceptions held by students when administered over time (Libarkin, 2008). Previous research regarding concept inventories within the chemistry education community have had difficulty satisfying common reliability analysis, such as Cronbach's alpha (Brandreit & Bretz, 2014). Concept inventory reliability measurement may not align with traditional indicators of reliability (i.e. Cronbach's alpha), as concept inventory items are designed to measure different conceptions and different topics within subject domains (Bretz & McClary, 2015). This may lead to decreased consistency among item responses and reduce item covariance, therefore leading to decreased reliability indicators (Tavakol & Dennick, 2011).

Previous research has not investigated alignment of curricula, educational approach, and instruments that capture student conceptions. Therefore, investigation of concept inventories, when paired with educational approaches and curricula that may promote conceptual reprioritization (i.e. cooperative learning and *Chemical Thinking*), is needed. In addition, this investigation offers instructors an empirical base to determine which educational approaches and curricula are associated with undergraduate students' conceptual reprioritzations towards dominant, scientifically accepted, redox conceptions. This empirical base creation is due to the lack of previous studies comparing association of traditional and reform-based general chemistry curricula with undergraduate conceptual reprioritization of redox conceptions.

Theoretical Framework

Vygotsky’s social constructivism theory lays the theoretical groundwork for cooperative learning, and discussion-oriented instruction, as it posits learning as a social phenomena (Vygotsky, 1962). Lecture and cooperative learning environments may act as platforms for the social construction of knowledge, as they may promote student collaboration (Bowen, 2000; Talanquer, 2013; Talanquer & Pollard, 2010; Sevian & Talanquer, 2014). *Chemical Thinking* may use these platforms for students to socially construct knowledge through its discussion-oriented lecture and promotion of argumentation in its workshop activities (Talanquer & Pollard, 2010). Discussions or arguments that take place in lecture or cooperative learning workshops may increase the likelihood of students being exposed to useful conceptions, which may increase the likelihood of a competition of useful conceptions to take place and theoretically result in a conceptual reprioritization of discussed conceptions (Shtulman & Lombrozo, 2016, Micheli, 2011). *Chemical Thinking* lectures may increase the likelihood of conceptual reprioritization through its “let’s think” activities, that lead to class discussions, or offer students opportunities to present arguments for proposed questions (i.e. Make a list of features, explicit and implicit, that you think are relevant in prediction the directionality of the process). Cooperative learning groups are supported by Vygotsky’s social constructivism through the promotion of social learning in small groups (Vygotsky, 1962). Cooperative learning workshop activities may also increase the likelihood of conceptual reprioritization through their activities (i.e. Justify your choice), that may lead to vocalized arguments, where students may be exposed to useful conceptions (Shtulman & Lombrozo, 2016; Shah et al. 2018; Bowen, 2000; Talanquer, 2013; Sevian & Talanquer, 2014). This study used cooperative learning groups of 3-4 students, where students were encouraged to work together and argue to solve workshop activity problems. The alignment of a reform-based general chemistry curriculum with argumentation and discussion of chemistry concepts in lecture-sized or small groups promotes the social construction of knowledge, and may promote conceptual reprioritization (Vygotsky, 1962; Shtulman & Lombrozo, 2016; Sevian & Talanquer, 2014).

ROXCI

ROXCI, developed by the Bretz group, is a concept inventory that may be practical for relatively large undergraduate courses, due to the ability to use scantrons for the multiple-choice instrument, and capture undergraduate students' redox conceptions. *ROXCI* aims to identify conceptions, held by students, of oxidation-reduction chemistry. The instrument does so through 18 questions, six of which are two-tiered items, and target many themes that arise within the topic of oxidation-reduction (Brandreit & Bretz, 2014). Two-tiered items are extension questions that first ask respondents to answer a question (tier one), then ask for an explanation for that answer (tier two). *ROXCI* targets the themes of: oxidation numbers, electron transfer, surface features of oxidation-reduction reactions, spectator ions, dynamics of reaction processes, and electrostatics and bonding (Brandreit & Bretz, 2014). The inventory uses many symbolic representations of reactions (chemical equations) and pictures as particulate representations for respondents to account for when answering *ROXCI* items. Conceptions are captured by *ROXCI* items, as the answer choices to the items were constructed based on undergraduate student conceptions (Brandreit & Bretz, 2014). Students in this study were expected to use their knowledge and conceptual understanding of oxidation numbers, electron transfer, surface features of redox reactions, spectator ions, dynamics of reaction processes and electrostatics and bonding when answering *ROXCI* items.

Rationale and Research Questions

This study aims to first, increase the record of undergraduate students' conceptions of redox, as few studies have accumulated a record of undergraduate students' conceptions of oxidation-reduction. Second, Evaluate *Chemical Thinking* and a traditional general chemistry curriculum with regard to undergraduate students' redox conceptions, which may provide insight to what chemistry curricula may be associated with conceptual reprioritization. Finally, this study aims to compare the evaluations of *Chemical Thinking* and a traditional general chemistry curriculum with regard to undergraduate students' redox conceptions, as such a comparison has yet to be reported.

The following research questions were created with these goals in mind:

1. What redox conceptions are held by undergraduate general chemistry students after *Chemical Thinking* and traditional chemistry curriculum instruction?
2. Can *ROXCI* detect reprioritizations in undergraduate general chemistry students' redox conceptions? If so, to what degree? Are those changes different depending on general chemistry curricula?
3. How do reform-based general chemistry students' conceptual reprioritizations of redox conceptions compare to traditional general chemistry students?

Methods

Research Design

This study took place in the fall 2017 semester at a large public university in the northeastern region of the United States. The following protocol was executed after receiving IRB approval. Two separate general chemistry classes, one used a reform-based general chemistry curriculum (*Chemical Thinking*) and the other used a traditional curriculum. were given the ROXCI instrument two times. The first iteration of the *ROXCI* assessment took place, before relevant class instruction, during the eighth week of the fall 2017 semester for the traditional curriculum (TC) course, and the 12th week of the fall 2017 semester for the reform-based curriculum (RBC) course. Relevant class instruction took place for two weeks (11 and 12) of the TC course, and following week (13) of the RBC course. Relevant lecture material (i.e. lecture slides or clicker questions) was matched with topics covered by *ROXCI* to determine times of administration for each respective course. The reform-based general chemistry curriculum included discussion-oriented lectures led by the instructor. Discussions were student-centered, mediated by microphone, and offered students the opportunity to be exposed to useful redox conceptions. The traditional general chemistry curriculum included semi-active lecture led by the instructor, and did not involve student-centered discussions. Semi-active lectures are lectures that include very few active learning activities (i.e. a few clicker questions) and their time is dominated by passive learning activities (i.e. didactic lecture). Lecture took place three times per week, for a total of 160 minutes per week, for both classes during the fall 2017 semester. Relevant topics of instruction in both courses include: oxidation numbers, electron transfer, electrostatics and bonding, spectator ions, electrochemistry and balancing of oxidation-reduction reactions.

Cooperative learning workshops were offered in both classes throughout the whole semester, once per week, for 80 minutes per week. Students were placed in small groups of 3-4 students each, designed to maximize diversity based on demographic questionnaire responses. Relevant RBC workshop activities encouraged students to argue their answers to workshop problems (i.e. justify your choice), while relevant TC workshop activities were algebraically focused (i.e. Calculate the cell potential of…) and placed less emphasis on argumentation. Students were expected to apply knowledge gained in lecture, or other relevant means (i.e. office hours, study groups), to the prompts given in the workshop setting. The *ROXCI* post-test assessment was administered, after relevant instruction, during the 13th week of the fall semester for the TC course, and 14th week of the RBC course. Each iteration of the *ROXCI* assessment took 20 minutes, and items remained in the same order.

Five TC students were interviewed, using a semi-structured interview format, for approximately 60 minutes during the 15th and 16th week of the fall 2017 semester to gain insight of how they were thinking about the *ROXCI* instrument and conceptions presented in its answer choices. No RBC students were interviewed due to time commitments.

Consenting students that completed both the pre and post-test, and the demographic questionnaire, were given extra credit equivalent to one full workshop quiz (RBC) or workshop class (TC). Extra credit awarded to fully participating students equated to 0.55% (RBC), or 0.2% (TC), of their overall course grade. A research design summary can be found below in table 1.

Table 1. ROXCI research design summary.

Reform-Based General Chemistry (Week of Fall Semester)	**Traditional General Chemistry (Week of Fall Semester)**	**Action**
12	8	*ROXCI* Pre-Test Administration
13	11-12	Audio and video recording during relevant workshop discussions (80 mins/week)
-	15 & 16	Individual Interviews (~60 mins)
14	13	*ROXCI* Post-Test Administration

Validity and Reliability

Content, construct and substantive validity were assessed to determine if *ROXCI* is truly testing redox conceptions (content), properly testing redox conceptions (construct), and if students were appropriately interpreting (substantive) *ROXCI* items (Libarkin, 2008). Previously, Brandreit & Bretz (2014) have emphasized the importance of semi-structured interviews to be the primary determinant of substantive and secondary determinant of content validity due to the fruitful information that may be gained about the instrument's content, and ways in which students were interpreting *ROXCI* prompts. The primary determinant of content and construct validity was done by course instructors. Course instructors were given the inventory, asked to review it, and determine if the material covered by their respective courses would appropriately prepare their students to answer the questions on the *ROXCI* instrument. This study used semi-structured interviews to assess substantive and content validity.

Reliability of persons and items were assessed using the Rasch dichotomous model in Winsteps version 3.68.2 (Linacre, 2018). The ITEM: dimensionality function, a function of principal components analysis of residuals, was used to determine dimensionality of the instrument, and provides evidence for *ROXCI*'s validity of testing latent construct(s). Rasch dichotomous model was used instead of Cronbach's alpha to determine the internal reliability of *ROXCI* because Chronbach's alpha is more of a confirmatory indicator of internal consistency, rather than determinant of internal consistency (Tavakol & Dennick, 2011). Items 1-6 were paired together as one question due to their two-tiered nature (the second item being an extended question from the first), when assessing dimensionality of *ROXCI*. No student responses were removed from reliability analysis due to a score of 0 or 100 on *ROXCI*. Reliability of person ability was also assessed using the 2pl model for dichotomous data in RStudio version 1.1.447, also due to the dichotomous nature of the analyzed data set.

Data Sample

Consenting students belonged to two separate general chemistry courses. One course was an advanced, three sequence, general chemistry course that used a reform-based curriculum (RBC), and the other was an off-sequence general chemistry II course that used a traditional general chemistry curriculum (TC). The RBC course is expected to be composed of students who have taken at least two years of high school chemistry. Of the 179 total students enrolled in the course, 136 completed both iterations of the *ROXCI* assessment. 100 students (55.87% of the total) consented, took the demographic questionnaire, and completed both iterations of the *ROXCI* assessment. The TC course is expected to be composed of students who have taken general chemistry I, a course equivalent, or students who have not previously succeeded in passing a general chemistry II course. This population may include students who have changed majors and need to take the second half of the general chemistry sequence, transfer students, students who had previously taken general chemistry I in the spring 2017 semester and students who had taken general chemistry II during the spring 2017 semester but did not successfully complete the course. 193 total students were enrolled in the Fall 2017 semester of the TC course, 81 of which completed both iterations of the *ROXCI* assessment. 28 (14.51%) students consented, completed the demographic questionnaire, and completed both iterations of the *ROXCI* assessment. Students in both courses were given a demographic questionnaire at the time of consent to self-report the following demographics: gender, race, ethnicity, first language learned, and student generation. Sufficient demographic responses for statistical analysis only allowed for analysis of gender and student generation demographics in the RBC course.

Sample Representation

Representativeness of the consenting students was analyzed using an independent samples t-test, assuming equal variances, on consenting student pre and post-test scores and whole class pre and post-test scores, in each respective class, using IBM SPSS statistics version 23. The post-test was used as an assessment of representation of students' conceptual understanding of redox chemistry, captured by *ROXCI* after instruction. The pre-test was used as an assessment of representation of students' conceptual understanding of redox chemistry, captured by *ROXCI* before instruction. Results of the independent samples t-test serve as evidence for the representativeness of consenting students to their respective classmates.

Analytical Framework

Item-Response Theory (IRT) was used a the primary form of analysis of student performance on *ROXCI* due to inconsistent item difficulty for *ROXCI* items (Brandreit & Bretz, 2014). Inconsistent item difficulty violates the assumption of parallel items in classical test theory, therefore promoting the use of Item-Response theory, which accounts of unparallel items when calculating student performance in the form of logit (log odds unit) scores of person ability (DeVellis, 2006; Boone, 2016). Cohen's δ was calculated to lend meaning to the item-response theory results. Rasch analysis was the form of IRT used to analyze student responses to *ROXCI*. To compare performance on *ROXCI* among the consenting RBC and TC students, all student responses subject to Rasch analysis were stacked, and post-test item difficulties anchored to the pre-test. Stacking means to place both RBC and TC responses in the same data set to assess equal item difficulty for *ROXCI* items for both classes. Anchoring post-test item difficulties to the pre-test allowed for the appropriate measurement of person ability change for all students along the same logit scale, while keeping item-difficulty constant. Post-test item difficulty was chosen on the theoretical basis that students were prepared to answer *ROXCI* items after appropriate redox instruction, therefore allowing for the most accurate representation of item responses and item difficulties.

A linear mixed-effects model was used to determine significant factors that correlated with person ability scores. The model can be seen below in equation 1.

$$(1) \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + Zu + \varepsilon$$

Where Y, the dependent variable, is person ability. β_0 is the students' baseline person ability, β_1 is the class the student belongs to, β_2 is gender, β_3 is student generation, and β_4 is the number of weeks that had gone past since taking the *ROXCI* pre-test. ε is model error, and Zu, subject ID, is the random effects variable. A linear mixed-effects model was used instead of the more common linear fixed model due to its ability to assume a different baseline person ability for the random effects term, student ID, and appropriately account for the repeated-measures nature of the study (Lindstrom & Bates, 1988). The data set for the model was also stacked, by containing person ability and demographic responses for each respective variable from both courses involved In this study. This model was created and tested using the lmer function, part of the LME4 package version 1.1-17, in RStudio version 1.1.447.

Results

Validity and Reliability

ROXCI construct and content validity was determined by the course instructors for the RBC and TC students. Both course instructors agreed, after reviewing *ROXCI*, that the material covered was appropriate for the students based on course content and the items were appropriately assessing their targeted conceptions. To assess substantive validity, five TC students were interviewed and asked to interpret *ROXCI* items. Four of the five Individual student interviews of TC students offered interpretations of *ROXCI* items. Three of the four students who offered interpretations vocalized interpretations for at least 16 of the 18 questions that were consistent with question wording, the fourth student offered interpretations for six of the 18 *ROXCI* items, all of which were consistent with question wording. The fifth student did not offer interpretations of any *ROXCI* items.

Reliability was assessed using the Rasch dichotomous model in Winsteps version 3.68.2, which returned a person value of 0.67 and an item reliability value of 0.98. Principal component analysis of residuals revealed an eigenvalue less than 2 for the first contrast. This suggests that *ROXCI* is unidimensional and tests one latent construct, serving as evidence for the validity of the instrument testing oxidation-reduction. The 2pl model in RStudio version 1.1.447 returned reliability values of 0.70 (post-test) and 0.71 (post-test).

Sample Representation

An independent samples t-test, assuming equal variances, was performed on consenting students' test level scores and the TC and RBC class test level scores. The results indicated that the TC consenting students' test level scores on the pre and post-test were not significantly different than their peers ($df_{pre} = 107$, $t\text{-}stat_{pre} = -0.305$, $p_{pre} = 0.761$, $df_{post} = 107$, $t\text{-}stat_{post} = -0.269$, $p_{post} = 0.788$). The lack of significant differences between pre and post-test mean scores for consenting students and their TC peers serves as evidence for consenting students' representativenss of their peers before and after instruction. The RBC consenting students' test-level scores were significantly different than the RBC class's test level scores on the pre-test (df = 234, t-stat = -4.952, $p<0.001$), but not significantly different on the post-test (df = 234, t-stat = 0.022, p = 0.982). The presence of a significant difference between consenting RBC students and their peers on the pre-test mean scores serves as evidence for a lack of representativeness of consenting students to their peers before instruction. However, the lack of a significant difference in mean scores between consenting students and their peers on the post-test serve as evidence of consenting students' representativeness of their peers after instruction.

ROXCI Performance

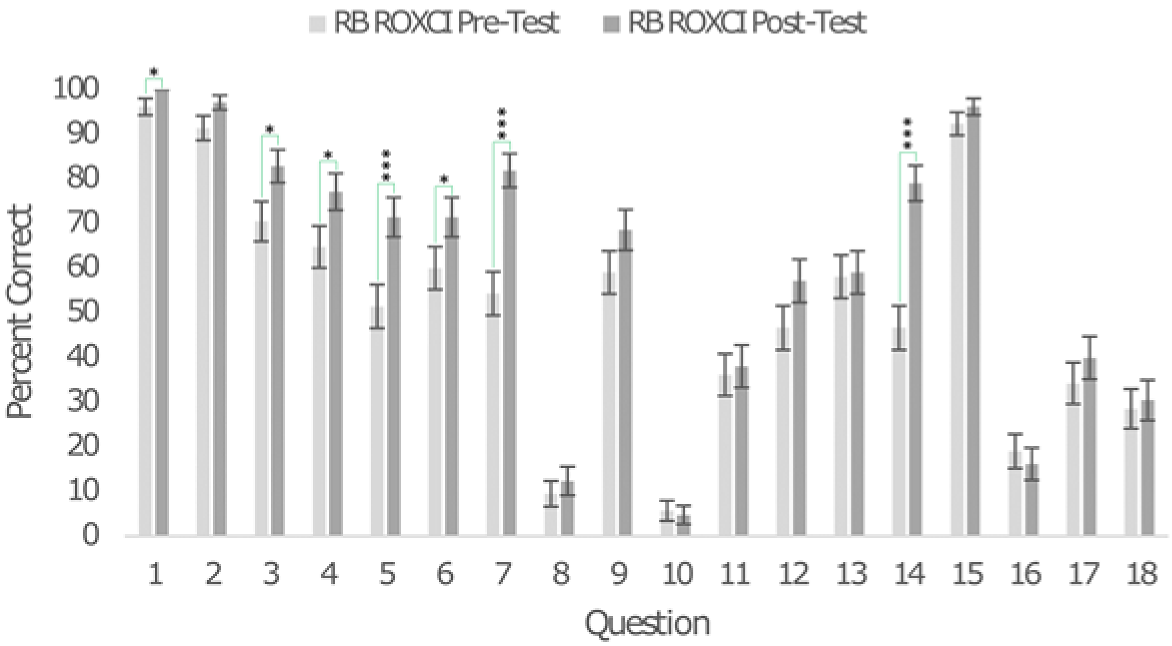


*Figure 1. Reform-based curriculum students' item-level ROXCI performance. *p<0.05, ***p<0.001. Pre/Post N = 100.*

Figure 1 displays RBC consenting students' *ROXCI* item performance (N = 100). Significant increases, from pre to post-test, occurred on seven of the 18 *ROXCI* items without any significant decreases in the remaining questions (see supplemental for *ROXCI* items). Significant gains were made on questions 1, 3-7, and 14. Significant gains were made with regard to the following themes: Oxidation numbers, surface features of redox reactions and electron transfer. Themes that lacked significant improvement (Qs 1, 2, 8-13, 15-17) included spectator ions, electrostatics and dynamics of reaction processes.

*Table 2. ROXCI item-level performance of RBC students. Effect sizes calculated using Cohen's d. *p<0.05, ***p<0.001. Pre/Post N = 100.*

Question	$\bar{X}_{\mathrm{pre}}$ **(%)**	$\bar{X}_{\mathrm{post}}$ **(%)**	**T-Stat**	$P_{\mathrm{pre\text{-}post}}$	$\delta_{\mathrm{pre\text{-}post}}$
1	96.19	100.00	2.029	0.045*	0.20 (S)
2	91.43	97.14	1.922	0.057	0.19
3	70.48	82.86	2.309	0.023*	0.23 (S)
4	64.76	77.14	2.309	0.023*	0.21 (S)
5	51.43	71.43	4.197	<0.001***	0.30 (S)
6	60.00	71.43	2.230	0.028*	0.17
7	54.29	81.90	5.025	<0.001***	0.47 (S)
8	9.52	12.38	0.726	0.469	0.53 (M)
9	59.05	68.57	1.593	0.114	0.17
10	5.71	4.76	-0.446	0.657	0.04
11	36.19	38.10	0.332	0.741	0.03
12	46.67	57.14	1.940	0.055	0.20 (S)
13	58.10	59.05	0.152	0.880	0.05
14	46.67	79.05	5.891	<0.001*	0.53 (M)
15	92.38	96.19	1.157	0.250	0.10
16	19.05	16.19	-0.598	0.551	0.40 (S)
17	34.29	40.00	0.973	0.333	0.07
18	28.57	30.48	0.391	0.697	0.05

Effect sizes, shown above in table 2, for significant gains made by consenting RBC students ranged from small to medium (Cohen's δ 0.2-0.53). Average effect size for significant items that target oxidation numbers (1, 3-7) was 0.26 (S), Surface features of redox reactions (1, 3-6) was 0.22 (S), and electron transfer (3 & 4) was 0.22 (S).

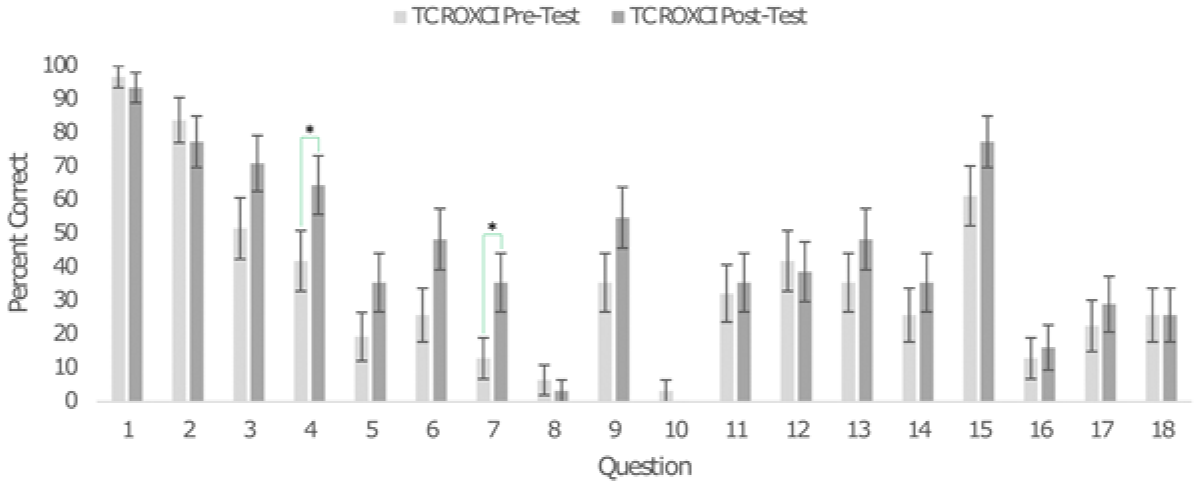


*Figure 2. Traditional curriculum students' item-level ROXCI performance. *p<0.05. Pre/Post N = 28.*

Figure 2 summarizes consenting TC students' item level performance on *ROXCI*. Significant gains were made in questions 4 and 7, without significant decrease in scores on remaining questions. Significant gains were made with regard to the following themes oxidation numbers, surface features of redox reactions, and electron transfer.

*Table 3. ROXCI item-level performance of TC students. Effect sizes calculated using Cohen's d. *p<0.05. Pre/Post N = 28.*

Question	$\bar{X}_{pre}$ (%)	$\bar{X}_{post}$ (%)	**T-Stat**	$P_{pre\text{-}post}$	$\delta_{pre\text{-}post}$
1	96.43	92.86	-0.571	0.572	0.12
2	85.71	82.14	-1.000	0.325	0.08
3	50.00	75.00	1.793	0.083	0.44 (S)
4	42.86	71.43	2.244	0.032*	0.49 (S)
5	17.86	35.71	1.541	0.134	0.32 (S)
6	28.57	50.00	2.038	0.050	0.35 (S)
7	14.29	39.29	2.528	0.017*	0.45 (S)
8	13.57	0.00	-1.000	0.325	0.27 (S)
9	39.29	60.71	1.793	0.083	0.35 (S)
10	3.57	0.00	-1.000	0.325	0.27 (S)
11	35.71	25.71	0.329	0.745	0.00
12	46.43	39.29	-0.329	0.745	0.12
13	28.57	46.43	1.161	0.255	0.30 (S)
14	21.43	32.14	1.139	0.264	0.19
15	64.29	78.57	1.541	0.134	0.26 (S)
16	10.71	14.29	0.373	0.712	0.08
17	21.43	32.14	0.701	0.489	0.19
18	28.57	28.57	0.000	1.000	0.00

Table 3 displays an item level summary of all *ROXCI* items for consenting TC students. Question 4 targeted the themes of oxidation numbers, surface features of redox reactions and electron transfer and had a positive effect size of 0.49. Question 7 targeted the theme of oxidation numbers, and had an effect size of 0.45. Although the TC student did not make significant gains on many questions, 8 of the 18 *ROXCI* item improvements had small effect sizes. Significant learning gains may have been impacted by the small sample size of the consenting TC population (N = 28).

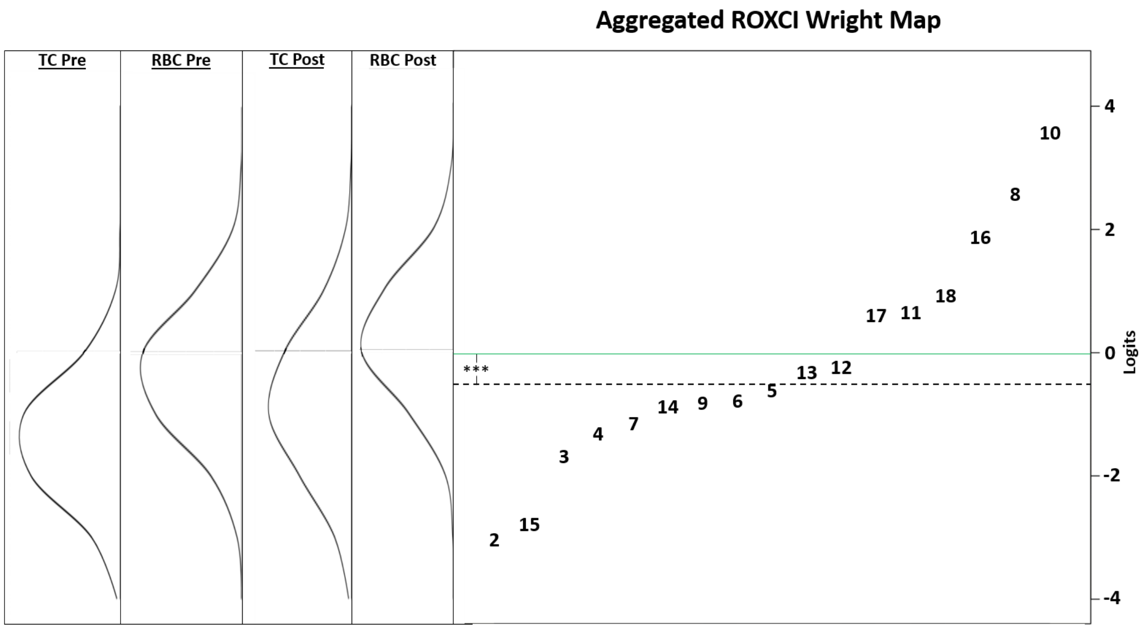


Figure 3. Aggregated general chemistry student ROXCI performance with disaggregated person ability distributions. ***$p<0.001$. Pre/Post N = 128.

Student responses for both classes were stacked and analyzed in RStudio, version 1.1.447, using the 2pl IRT model for dichotomous data. Post-test item difficulties were calculated separately and anchored to the pre-test to establish a consistent scale for pre and post-test ability measurements. Figure 3 exhibits overall consenting student performance for *ROXCI*. Overall, consenting students' average person abilities increased significantly from pre to post-test (Mean_{pre} = -0.5774, Mean_{post} = -0.0124). The overall effect size for increased person ability from pre to post for the aggregated students was small (Cohen's δ = 0.40). Item one was removed from analysis due to it having the only item outfit score to fall outside of the acceptable outfit score range 0.5-1.5 range (Boone, Staver & Yale, 2013), as it had an item outfit score greater than 1.5. The remaining 17 questions falling within the acceptable 0.5-1.5 outfit score range are determined to be appropriate for the tested population (see appendix for *ROXCI* item outfit table). WLE (weighted likelihood estimate) reliability assessment for pre-test person ability was 0.69, and 0.68 for the post-test person ability. In summary, consenting students in both courses made significant person ability gains from pre to post-test with a small effect size, and 17 of the 18 *ROXCI* items were appropriate for the combined population of consenting students.

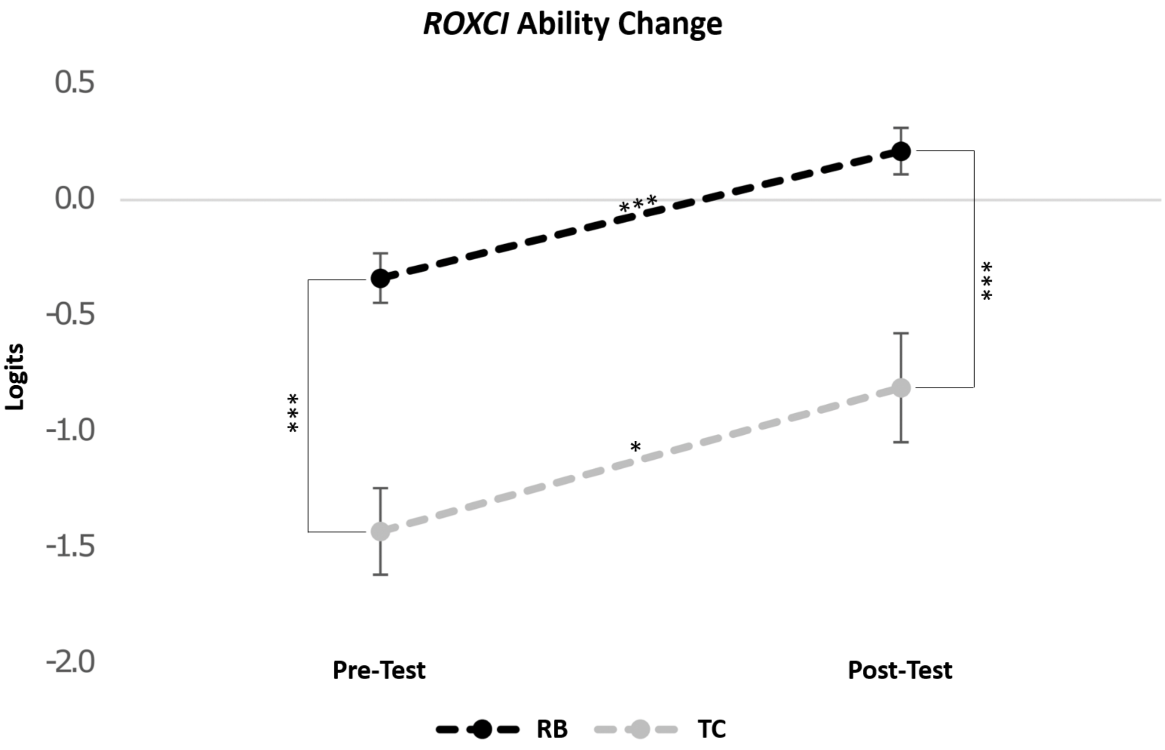


Figure 4. Disaggregated ROXCI person ability by course. *p<0.05, ***p<0.001. RB Pre/Post N = 100. TC Pre/Post N = 28.

Disaggregation of person ability into class is summarized above in figure 4. Consenting RBC students significantly increased their person ability from pre to post-test (mean$_{\text{pre}}$ = -0.3387, mean$_{\text{post}}$ = 0.2109). Consenting TC students also significantly increased their scores from pre to post-test (mean$_{\text{pre}}$ = -1.4299, mean$_{\text{post}}$ = -0.8101). RBC students' pre and post-test logit scores were significantly different than the TC students ($p_{\text{pre \& post}}<0.001$). The overall change in pre to post person ability for both courses was not significantly different (Logit increase$_{\text{RB}}$ = 0.5496, Logit increase$_{\text{TC}}$ = 0.6198, p = 0.749). Both courses had similar effect sizes for pre-post person ability increase (Cohen's δ_{RB} = 0.53, Cohen's δ_{TC} = 0.55). An independent samples t-test performed on pre and post person ability between both consenting populations revealed mean pre and post-test person ability of the RBC population was significantly greater than mean pre and post-test person ability of the TC population ($p_{\text{pre}}<0.001$, $p_{\text{post}}<0.001$). In summary, both courses made significant person ability increases, of about the same size, on *ROXCI*. However, RBC students' person ability scores are significantly greater than TC students' person ability scores for the pre and post *ROXCI* assessments.

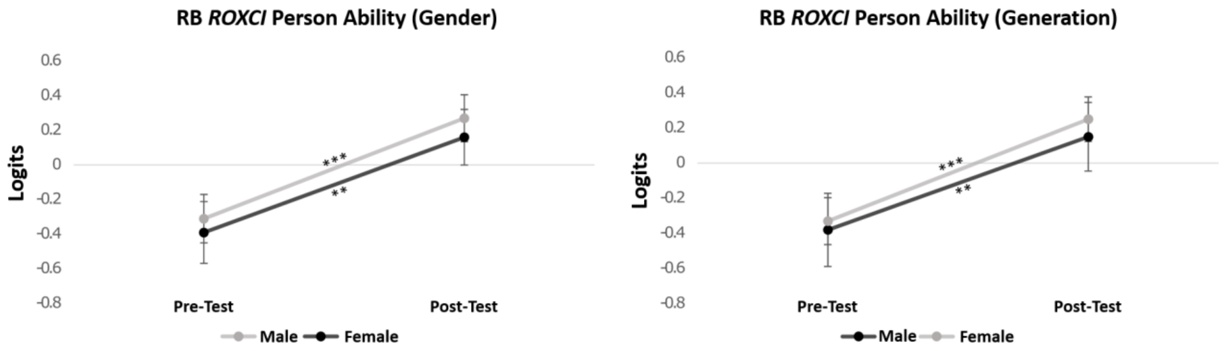


*Figure 5. RBC male, female, first-generation, non-first-generation ROXCI performance. **p<0.01, ***p<0.001. Pre/Post N (Male) = 53. Pre/Post (Female) = 43. Pre/Post N (FG) = 28. Pre/Post N (NFG) = 65.*

Further disaggregation of the reform-based general chemistry course performance into the demographics of gender and student generation is summarized above in Figure 5. Males significantly increased their person ability scores from pre to post-test ($\text{mean}_{\text{pre}} = -0.31$, $\text{mean}_{\text{post}} = 0.27$). Calculated effect size for the increase from pre to post among males was medium (Cohen's $\delta_{\text{Male}} = 0.65$). Females also significantly increased their person ability scores from pre to post-test ($\text{mean}_{\text{pre}} = -0.39$, $\text{mean}_{\text{post}} = 0.16$). Calculated effect size for the increase in scores from pre to post among females was also medium, however lower than the calculated effect size for the increase in male scores. (Cohen's $\delta_{\text{Female}} = 0.51$). Overall person ability scores for males and females were not significantly different on either the pre or post-test ($p_{\text{pre}} = 0.725$, $p_{\text{post}} = 0.573$).

Similarly, first-generation students significantly increased their person ability scores from pre to post-test ($\text{mean}_{\text{pre}} = -0.38$, $\text{mean}_{\text{post}} = 0.15$). Non-first-generation students' person ability scores significantly increased from pre to post-test ($\text{mean}_{\text{pre}} = -0.33$, $\text{mean}_{\text{post}} = 0.25$). Overall person ability scores for first and non-first-generation students were not significantly different on either the pre or post-test ($p_{\text{pre}} = 0.835$, $p_{\text{post}} = 0.648$).

Linear Mixed Effects Model

A linear mixed effects model, eq. 1 below, was created and tested using RStudio version 1.1.447. The purpose of the model was to assess what recorded factors (i.e. time, race, ethnicity) associated most strongly with person ability.

$$(1) \quad Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + Zu + \varepsilon$$

Results revealed the class a student was part of and time of taking the *ROXCI* assessment were significantly correlated with *ROXCI* person ability (p_{class} & $p_{\text{time}} < 0.001$), while having no significant fixed effects correlations. In summary, class and time were significantly influenced *ROXCI* person ability scores without significant interaction with other variables part of the mixed effects model.

Implications and Discussion

Validity and Reliability

The *ROXCI* assessment's content and construct seems to be appropriate for the populations tested due to the post-test person ability distributions covering the item difficulty distributions. Principal components analysis also supports unidimensionality of the *ROXCI* assessment, suggesting it tests one major topic: oxidation-reduction (Brandreit & Bretz, 2014). Interview responses regarding the interpretations of the questions that appear on the assessment support the claim that *ROXCI* is appropriately communicating their questions to respondents.

Person reliability of the *ROXCI* post-test, 0.67, suggests that *ROXCI* is capable of separating students into one or two levels (Boone, Staver & Yale, 2013). However, the 2pl model in RStudio suggests the pre and post-test responses are slightly more capable of separating students into separate levels, as both the pre and post-test reliability values were at or above 0.7.

Sample Representativeness

The results of the independent samples t-tests between the test-level scores of consenting populations compared to each respective class suggest that the post-test scores of the students who consented are representative of their respective classes, due to their post-test scores not being significantly different. This lends support to the claim that the learning outcomes of the consenting student populations is representative of their respective classes. However, analysis of pre and post-test scores may not be representative of a general student population, as both course populations' representativeness of a more general student population has not been assessed. A creation of a database, with appropriate preservation of confidentiality, of student responses to the *ROXCI* instrument would allow for a more appropriate assessment of representation of student population responses towards a more general population.

ROXCI Performance Implications

Item response theory analysis (Figure 4) suggest both populations made significant conceptual learning gains, and potential adoption of dominant, scientifically accepted, redox conceptions regarding oxidation numbers, surface features of redox reactions and electron transfer. The comparable logit score increases and effect sizes in both course adds meaning to the conceptual learning gains made on *ROXCI*. TC and RBC students increased their person ability scores by approximately the same amount (logit increase_{RB} = 0.5496, logit increase_{TC} = 0.6198), suggesting fairly equivalent conceptual learning gains. Both courses seemed to improve their post-test person ability scores by approximately 0.55, or slightly above, standard deviations compared to their pre-test scores ($\delta_{TC} = 0.55$. $\delta_{RB} = 0.53$), suggesting equivalent meaning of conceptual learning gains for each course. Figure 4 allows further interpretation of the particular conceptual learning gains made for the student populations, as the number of questions that both populations had greater than a 50% change of correctly answering (questions below mean pre and post-test person ability) increased from 9 to 11 questions. Furthermore, the item difficulty distribution suggests that questions 6, 9, 14, and 12 and 13 may be redundant and impede on the ability of *ROXCI* to more appropriately separate person ability, as their item difficulties are very similar. However, the decision to remove a question, and which question to remove, must account for what sub-concepts are being tested and student performance among such sub-concepts. The lack of large gaps between item difficulties suggests that *ROXCI* is capable of separating student ability from -3 to 4 logits.

Although these learning gains are comparable among courses that use a different chemistry curriculum, it is important to note that the pre and post-instruction proficiency of the students within these courses are not equivalent. This proficiency difference may be due to the populations of students being different, as the reform-based curriculum population is composed of students who have theoretically performed highly in previous chemistry courses, while the traditional curriculum population being composed of a more diverse population from an academic performance perspective. This implication is supported by the linear mixed effects model, that indicated class and time to be significantly correlated with person ability scores. Class and time were expected to be significant predictors of *ROXCI* person ability based on the theoretical academic makeup of the student populations in each course.

Item-level analysis for learning gains suggest both student populations made significant learning gains with regards to the concepts of oxidation number, electron transfer, and surface features of oxidation-reduction reactions, as the significant gains made by both populations intentionally targeted those themes (Brandreit & Bretz, 2014). A more in depth look at the item-level learning gains for the RBC population (Table 1) suggests that the most significant learning gains were made on questions 7 and 14, as supported by their effect sizes being the largest (Cohen's $\delta_7 = 0.47$, Cohen's $\delta_{14} = 0.53$). Furthermore, these significant gains on questions 7 and 14 suggest that the reform-based student population made their largest conceptual learning gains, and potential dominance of scientifically accepted conceptions, on the themes of oxidation numbers, targeted by question 7, and surface features of oxidation-reduction reactions, targeted by question 14. The TC population seems to have made the most significant conceptual gains, and potential dominance of scientifically accepted conceptions similar to the RBC population, regarding the themes of oxidation numbers, electron transfer and surface features of oxidation-reduction reactions, as suggested by Table 2. These concepts were tested by questions 4 and 7 (Cohen's $\delta_4 = 0.49$, Cohen's $\delta_7 = 0.45$).

A breakdown of the RBC population into gender and student generation reveals some interesting implications with regards to achievement gap differences (Nosek et al. 2008; Stephens, Hamedani & Destin, 2014; Warbuton, Bugarin & Nunez, 2001). The lack of significant differences of pre and post-test person ability among males and females within the RBC population suggest that there is no significant achievement gap between males and females with regard to oxidation-reduction topics assessed by the *ROXCI* inventory. This lack of a gender achievement gap in in contrast to previously reported gender achievement gaps in science and math (Nosek et al. 2008) Although their pre and post-test person ability scores were not significantly different, the effect sizes of learning gains made among males and females were different. Males increased their person ability after instruction by over 0.6 standard deviations compared to pre-instruction ability (Cohen's $\delta = 0.65$), while females increased their person ability after instruction by just over 0.5 standard deviations compared to pre-instruction (Cohen's $\delta = 0.51$). This suggests that although there is not a significant difference between pre and post-test person ability among males and females within the reform-based curriculum population, a difference is still present within the learning gains made between the sub-populations.

A similar suggestion is made in the comparison of first-generation student performance compared to non-first-generation students within the RBC population, as their pre and post-test person ability scores were not significantly different. This lack of a significant difference is also in contrast to previously reported social-class achievement gaps (Stephens, Hamedani & Destin, 2014; Warburton, Bugarin & Nunez, 2001). However, there was a difference in the meaningfulness of their learning gains, although not as large as the difference among males and females, as the non-first-generation students' learning gain effect size (Cohen's δ = 0.61) was greater than the first-generations students (Cohen's δ = 0.52) within the RBC population.

Summary

ROXCI is capable of detecting changes in undergraduate general chemistry students' redox conceptions. Furthermore, students within both a traditional general chemistry and reform-based general chemistry curriculum course made significant, and similar, scientifically accepted conceptual learning gains regarding redox concepts. However, these gains are not exactly equivalent, as the reform-based general chemistry curriculum populations' pre and post-test person abilities were significantly greater than their traditional general chemistry curriculum counterparts. A likely reason for these differences was indicated by the linear mixed effects regression analysis, revealing class and time as significant predictors of *ROXCI* person ability. There is also a mixed presence of dominant, scientifically accepted, and alternative conceptions in this study's population.

Item-level gains suggest that students with the traditional general chemistry and reform-based general chemistry curriculum populations made their most significant conceptual learning gains regarding the concepts of oxidation numbers, electron transfer and surface features of redox reactions, as supported by the items that the respective student populations made their most meaningful learning gains.

ROXCI is capable of detecting changes in undergraduate general chemistry students' redox conceptions past the class level. Disaggregation of the reform-based general chemistry curriculum population into the sub-populations of gender and student generation revealed statistically similar, yet slightly different, conceptual learning gains on the overall *ROXCI* assessment. Males within the reform-based population increased their average person ability from pre to post-test more than their female peers. Similarly, non-first-generation students and first-generation students made statistically similar, yet slightly different, learning gains on the *ROXCI* assessment, as non-first-generation students increased their average person ability scores from pre to post-test more than their first-generation peers.

The *ROXCI* assessment is capable of distinguishing student ability from -3 to 4 logits with regard to concepts within oxidation-reduction, however some items may be redundant based on similar item difficulties.

Limitations

This study was limited to the student populations, and their conceptual learning outcomes, tested at the respective university where this study took place, as a comparison of learning outcomes to a more general student population was not able to be assessed. The consenting student samples were not randomized, limiting their representativenss of their peers. Student conceptions captured by the *ROXCI* instrument are limited to the conceptions conveyed by the presented choices on the instrument, and therefore may not be representative of all conceptions held by the student populations involved in this study. Data samples within the reform-based and traditional general chemistry curriculum courses were limited to the students who successfully completed both iterations of the *ROXCI* assessment to completion, therefore limiting statistical power of results. These samples may have been influenced by student attendance and other factors that were not able to be captured by this study. Conceptual reprioritizations of both classes may be impacted by problem similarity, as each iteration of *ROXCI* followed the same item order. Comparison of TC and RBC learning gains should be understood with caution, as their learning gains were statistically similar, however their proficiencies with regard to redox conceptions are not equivalent. Similarly, student interviews were limited to the traditional general chemistry curriculum students, which limits insight into attribution of reform-based curriculum instruction and reform-based student thought processes when taking *ROXCI*.

Future Direction

The *ROXCI* assessment may be subject to refinement, depending on what its purpose is to be regarding undergraduate education and how scientifically accepted conceptions ties into a meaningful chemistry progression. Question may be removed, their prompts re-worded, and the conceptions captured by the multiple-choice options may be subject to change to reflect more accurate conceptions held by undergraduate students. An increase in sample size may also be very beneficial to increase the statistical power for claims that stem from results of *ROXCI* performance. To increase generalizability of scores and appropriately compare students, a larger sample size would be necessary, and propensity score matching among students within different institutions, or classroom treatments (i.e. different curricula, pedagogy, etc.), may be appropriate to accurately compare students of similar characteristics (academic ability, demographics, etc.) to increase the robustness of the *ROXCI* assessment as a diagnostic or predictive tool. Propensity score matching may be appropriate because it attempts to compare similarities among populations within different treatment groups (Caliendo & Kopeinig, 2008). However, there are many factors that impact the outcome variable (i.e. academic performance), which may increase the dimensionality of independent variables that may be used to predict learning outcomes and increase the difficulty of creating a large sample size after utilizing propensity score matching (Morgan et al. 2008). However, if there is conditional independence among factors, then the outcomes due to those factors may be independent of their treatments once those factors are accounted for in the propensity scores (Dawin, 1979). Propensity score matching may be able to overcome such an obstacle by reverse engineering from matched propensity scores to similar factors held between such scores to determine what factors are important and ultimately reduce dimensionality of reported factors to make the assessment as simple as possible. Propensity scores for students in different institutions, or other treatment variables, can be estimated and matched according to a logit probability model (i.e. a Rasch model), then the scores within different populations can be matched to each other, stratified to appropriately increase representativeness of matched scores and then compare the means of those populations to assess generalizability of learning outcomes from different treatment variables (Caliendo & Kopeinig, 2008).

A creation of a database, with appropriate preservation of confidentiality, may create an opportunity to assess representativeness of learning gains for future student populations tested by *ROXCI* for oxidation-reduction chemistry at the undergraduate level. A more meaningful assessment of scores on the *ROXCI* instrument to performance in a general chemistry course is currently absent in the literature. The assessment of *ROXCI* performance regarding student performance in an undergraduate general chemistry course would lend support for *ROXCI* being a predictor of general chemistry course performance at the undergraduate level and may support the attainment of scientifically accepted conceptions to be more worthwhile to students and instructors. However, a community standard of undergraduate chemistry performance must be set to determine what scores ought to be considered meaningful and possible predictors of general chemistry course performance. This standard may be reverse engineered, as successful student *ROXCI* scores (i.e. passing letter grades, benchmark course performance, etc.) may be investigated and determined to be the standard that instructors would strive to have their students meet. An obstacle to the creation of this standard is the adoption of such a standard at a national level. The standard must be determined at a national level and be robust for it to be considered as worthwhile to instructors and undergraduate institutions across the nation. It may also be unclear as to who ought to determine a national standard of undergraduate chemistry performance, and its criteria, at the current time. However, the creation of a community standard may pave the way for the encouragement of widespread concept inventory use.

Chapter 4: Association of Argumentation Quality and Conceptual Reprioritization of Organic Acidity

Introduction

Background

Acid-base chemistry has been a common chemistry unit that has been quite difficult for undergraduate and graduate students as students have turned to quick, surface-level, routes of understanding chemistry material through memorization and heuristics, had difficulty with qualitative application of mental models of organic acidity, and difficulty understanding nucleophiles and electrophiles (Bretz & McClary, 2015; Bhattacharyya, 2006; Cartrette & Mayo, 2011; Bradley & Mosimege, 1998). Dependency on heuristics and memorization has led to test-taking mistakes (i.e. most acidic hydrogen being explicitly drawn, inappropriate interpretation of organic structures), fragmented mental models and reduced long-term retention of chemistry concepts (Bhattacharyya, 2006). A potential cause of these difficulties have been thought to be that traditional chemistry curricula don't offer students enough opportunities to more deeply conceptualize course material due to their emphasis on algebraic proficiency (Sevian & Talanquer, 2014; Talanquer, 2018; Talanquer, 2013).

Traditional chemistry curricula, usually paired with a didactic lecture, tend to rely on students' mathematical proficiency, rather than their conceptual understanding of chemistry material (Talanquer & Pollard, 2010; Talanquer, 2013). A recently developed general chemistry curriculum, *Chemical Thinking*, developed by Talanquer and co-workers encourages student discussion through a discussion-oriented lecture and cooperative workshop activities. *Chemical Thinking* is centered around fundamental questions of chemistry that are relevant to chemical practice (Talanquer & Pollard, 2010). This focus on discussion in lecture and cooperative learning workshop activities provides students the opportunity to voice, listen to, and reprioritize dominant conceptions regarding the topic of interest. Argumentation, a form of discussion, may also be promoted by a curriculum that encourages student discussion and lead to deeper understanding of learnt material (Shah et al. 2018; Shtulman & Lombrozo, 2016; Potvin, 2017; Talanquer & Pollard, 2010). Argumentation may also be promoted by cooperative learning, as it places students in small groups to collaborate and solve problems (Bowen, 2000; Freeman et al. 2014). Diversity of small learning groups has been linked to increased critical thinking and problem-solving skills, which may lead to increased argumentation quality (Hurtado, 2001).

Theoretical Framework

Vygotsky's social constructivism lays the groundwork for learning that occurs in cooperative learning and discussion-oriented instruction through its support of socially constructed knowledge, executed by student-student and student-instructor collaboration (Vygotsky, 1962). Cooperative learning creates an environment that encourages the social construction of knowledge by promoting student collaboration in small groups. Furthermore, cooperative learning in small, diverse groups, may lead to increased argumentation, a form of discussion, as diverse learning groups has been linked to increased critical thinking and problem-solving (Hurtado, 2001; Shah et al. 2018). Argumentation, previously defined as "a verbal activity oriented towards the realization of a goal," (Micheli, 2011) may expose students to many conceptions, and potentially increase the probability of a conceptual reprioritization occuring

(Shah et al. 2018; Shtulman & Lombrozo, 2016; Potvin, 2017). *Chemical Thinking*, a recently developed general chemistry curriculum, uses cooperative learning workshop activities that promote student-student discussion, and potential argumentation, and a discussion-oriented lecture, which promotes instructor-student and student-student discussion (Talanquer & Pollard, 2010). Based on the social construction of knowledge, the promotion of discussion and argumentation by *Chemical Thinking* may in turn promote conceptual reprioritization.

Discussion & Argumentation

Argumentation and discussion are linguistic mechanisms that offer students the opportunities to voice, collect, and reprioritize conceptions. Argumentation, "a verbal activity oriented towards the realization of a goal," (Micheli, 2011) is an important form of communication and may impact the social construction of knowledge, as stronger arguments may lead to conceptual reprioritization towards dominant, scientifically accepted conceptions (Vygotsky, 1962, Shtulman & Lombrozo, 2016; Potvin, 2017; Shah et al. 2018; Nussbaum Sinatra & Poliquin, 2008). Argumentation is common within the scientific community, as scientists must argue on their behalves for the stories that their research may tell. Recently, the increased use of argumentation to teach the practice of science has arisen (Talanquer, 2013). This increased use opposes its less frequent use in the past, which may have impacted students' ability to argue and reprioritize their conceptions (McNeil & Pimentel, 2010; Cohen, 1994; Sampson & Clark, 2009). However, Increasing student participation in discussion and argumentation has been tough and varying degrees of participation my impact the benefits one can receive (Cohen, 1994; Sampson & Clark, 2009). Active participation in discussion or argumentation may not be necessary to collect exposed conceptions by discussion mates, which in turn implies active participation in discussion may not be required to undergo conceptual reprioritization. Although, active participation in discussion and argumentation has been linked to increased student performance when applying discussed ideas in problem-solving questions (Cohen, 1994), which may imply an increased likelihood of undergoing conceptual reprioritization if one actively participates in discussion. Previous research has shown increased learning gains in undergraduate physics classrooms with argumentation having a central role in problem solving activities and stronger scientific understanding (Nussbaum, Sinatra & Poliquin, 2008; Mason, 1998). However, no research has been performed to investigate potential associations of argumentation with undergraduate students' conceptual reprioritization(s) of oxidation-reduction.

Toulmin's Argumentation Pattern & Argumentation quality

Toulmin's argumentation pattern (TAP) is a simple methodological tool for identifying components of arguments (Toulmin, 1958). TAP breaks down arguments into the following six components: claim, data, warrant, rebuttal, qualifier, and backings (Toulmin, 1958). A basic argument (BA) requires three of these components: claim, data and warrant (Kulatunga et al. 2014). The claim is used to put forth an idea, data is used to support that claim, and a warrant is utilized to connect referenced data to the claim (Kulatunga et al. 2014). Definitions of the components, and basic argument, are listed below in table 1.

Table 1. Argumentation components adapted from Toulmin (1958) & Kulatunga et al. (2014).

Component	**Definition**
Claim	An assertion put forth to the public regarding the topic/question of interest.
Data	Facts or information used to support a claim.
Warrant	A justified connection between data and a claim.
Backing	Assumptions under which the warrant holds power.
Qualifier	Conditions under which a claim is true.
Rebuttal	Refutations that may undermine a previous claim.
Basic Argument (BA)	A verbal utterance that contains a claim, data and warrant connecting the data to the claim.

TAP can be used to identify these components in transcripts from student-student arguments to understand how students may be arguing for, or against, claims made by themselves or others. However, TAP only assesses the presence of argumentation components in discussions and does not assess quality or correctness of arguments (Erduran, Simon & Osborne, 2004).

Erduran and co-workers have created an analytical framework that uses TAP to assess argumentation quality (Erduran, Simon & Osborne, 2004), shown below in table 2. Cooperative learning settings and discussion-oriented instruction may increase the likelihood of arguments taking place, providing students increased opportunities to be exposed to conceptions, therefore increasing the likelihood of those participating in discussion or argumentation to undergo conceptual reprioritization (Shtulman & Lombrozo, 2016; Bell & linn, 2000; McNeil & Pimentel, 2010; Shah et al. 2018). However, there has been no research done to investigate associations of argumentation quality and undergraduate students' conceptions of oxidation-reduction, therefore making the investigation potential association(s) an area of interest.

Table 2. Argumentation quality framework adapted from Erduran, Simon & Osborne (2004).

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

Rationale and Research Questions

There is evidence for the association of argumentation and discussion with student conceptions in science (Bell & Linn, 2000; McNeil & Pimentel, 2010; Shah et al. 2018; Nussbaum, Sinatra & Poliquin, 2008). The promotion of argumentation through cooperative learning activities and discussion-oriented lectures (Talanquer & Pollard, 2010; Talanquer, 2013; Sevian & Talanquer, 2014) by a reform-based general chemistry curriculum may allow for investigation of an association between argumentation quality and general chemistry students' conceptions of organic acidity, as such an investigation has yet to be reported.

The main objectives of this study was to gain information that may provide insight/explanation to the quantitative results captured by the *ACIDI* concept inventory and investigate potential association between argumentation quality and dominant, scientifically accepted, conceptions. With two objectives in mind, the following research questions were formulated.

1. What insight can individual interviews provide regarding organic acidity conceptions held by reform-based general chemistry curriculum students?
2. Is there any evidence for the association of argumentation quality and dominant, scientifically accepted, conceptions held by undergraduate general chemistry students?

Methods

Research Design & Data Sample

This study took place at a large pubic university in the northeastern region of the United States. The following protocol was executed after receiving IRB approval. A first year, reform-based general chemistry course was given the *ACIDI* concept inventory three times, for 15 minutes each, during the Fall 2017 semester. Initial collection of organic acidity conceptions present within the study population using *ACIDI* was done once before relevant course instruction. A second collection of organic acidity conceptions present in the study population was done after relevant course instruction and allowed for the tracking of reprioritized, or lack thereof, conceptions. The final collection of organic acidity conceptions present within the tested population was done the following semester to measure persistence of dominant organic acidity conceptions well after relevant instruction.

The reform-based course offered a discussion-oriented lecture three times per week for 160 minutes/week, and one cooperative learning workshop (two separate sections) per week for 80 minutes/week, for 14 weeks during the Fall 2017 semester. Lecture and workshop activities were based on the *Chemical Thinking* general chemistry curriculum. Lecture and workshop activities were matched with *ACIDI* material to determine timing of *ACIDI* assessments. Relevant course instruction included information and problems with regard to the following organic acidity topics: identification of the most acidic proton, resonance (de)stabilization and induction. . Relevant lecture material and instruction was determined by matching lecture material (i.e. slides, clicker questions) with *ACIDI* material. Lecture and workshop activities offered students opportunities to voice, collect and reprioritize their conceptions regarding organic acidity. Lecture discussions were led by the instructor, and workshop activities were facilitated by (under)graduate TAs. For the first half of workshop, students were allowed to work in their workshop groups for approximately 30 minutes before a 10 minute period of reviewing workshop questions. A second cycle was followed for the second half of workshop. Cooperative learning workshops placed students in small, diverse (i.e. race, gender, ethnicity), learning groups of 3-4 students. All students were given a demographic questionnaire to collect the demographic information that was used to create the workshop groups. Three workshop groups per section, totaling six groups, were audio and video recorded for the whole semester. Two workshops of interest, those covering material relevant to *ACIDI*, were determined based on workshop activities from the *Chemical Thinking* curriculum. All participants in this study were expected to complete the *ACIDI* pre, post and delayed post-test. A summary of the *ACIDI* research design can be found below, in table 3.

Table 3. ACIDI research design summary.

Reform-Based General Chemistry (Week of Fall Semester)	**Action**
6	*ACIDI* Pre-Test
7 & 9	Audio and Video Recording relevant workshop discussions (80 mins/week)
10	*ACIDI* Post-Test
15 & 16	Individual Interviews (30 mins)
2 (Spring)	*ACIDI* Delayed Post-Test

Representativeness of Sample

ACIDI pre and post-test scores for sixteen of the eighteen individuals that were audio and video recorded and completed both administrations of ACIDI were compared to the whole class' *ACIDI* pre and post-test scores via an independent samples t-test, assuming equal variances, in IBM SPSS Statistics version 23. Results of the independent samples t-tests serve as evidence to suggest audio and video recorded students' representativeness of their peers.

Interviews

Thirteen of the eighteen students that were audio and video recorded were interviewed during the final week of the semester to gain insight into conceptions held by reform-based general chemistry curriculum students. Individuals were interviewed in a semi-structured manner for about 30 minutes. Interview protocol can be found in the appendix. Interviews were then analyzed to extract information regarding how students were interpreting *ACIDI* question prompts and thinking about topics covered by questions the study participants significantly improved from pre to post-test. Questions 1, 2, 8 and 9 were prioritized during analysis and extraction of interview information due to the class significantly improving on those questions from pre to post-test. Questions 1 and 2 targeted resonance stabilization and questions 8 and 9 targeted induction (Bretz & McClary, 2015; Shah et al. 2018).

Discourse Analysis

Workshop discourse of the students of interest, those who were audio and video recorded, was collected, transcribed, coded and analyzed to gain insight into how students were thinking about *ACIDI* related material and assign argumentation quality. Workshop transcripts were coded according to Toulmin's argumentation pattern (Toulmin, 1958). Two independent researchers separately coded all relevant *ACIDI* workshop transcripts, then compared coding and resolved coding conflicts to at least 90% agreement. Transcribed workshops and interviews that were relevant to *ACIDI* can be found in the supplemental. Argumentation components are defined below in table 1 (Toulmin, 1958; Kulatunga et al. 2014).

Table 1. Argumentation components adapted from Toulmin (1958) & Kulatunga et al. (2014).

Component	**Definition**
Claim	An assertion put forth to the public regarding the topic/question of interest.
Data	Facts or information used to support a claim.
Warrant	A justified connection between data and a claim.
Backing	Assumptions under which the warrant holds power.
Qualifier	Conditions under which a claim is true.
Rebuttal	Refutations that may undermine a previous claim.
Basic Argument (CBA)	A verbal utterance that contains a claim, data and warrant connecting the data to the claim.

Quality of arguments created during workshop discourse were assessed using Erduran's analytical framework (Erduran, Simon & Osborne, 2004), shown below in Table 2.

Table 2. Argumentation quality framework adapted from Erduran, Simon & Osborne (2004).

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

Further clarification of how particular language used in the criteria of level 3, 4 and 5 must be addressed. A "weak rebuttal," part of a level 3 argument, was understood by the coders of this study as a simple refutation of a previous claim and do not go further beyond a simple "No, I don't think so"-like verbalization. A "clearly identifiable rebuttal," part of a level 4 argument, was understood as a verbalization that involves an "I don't think so"-like aspect and goes a step further with additional data or information that would further weaken a previous claim. An "extended argument," part of a level 5 argument, was understood as an argument with numerous claims and at least two rebuttals. Examples of argument quality can be found in the appendix.

A cumulation of quality arguments, or total argumentation quality, was calculated by adding each argument quality level for all quality arguments made during relevant workshop periods. Total argumentation quality was then compared to *ACIDI* post-test responses to investigate possible association of total argumentation quality and dominant, scientifically accepted, conceptions held by students after instruction. Strength of association was investigated using the correlation function in Microsoft Excel version 1805.

Results

Representativeness of Sample

An independent samples t-test was performed on the sixteen audio and video recorded individuals' *ACIDI* pre and post-test score in comparison to the pre and post-test scores for the whole class. Results of the t-tests indicated that the *ACIDI* pre (df = 164, t-stat = 0.541, p = 0.59) and post-test scores (df = 164, t-stat = 0.721, p = 0.472) for the groups were not significantly different than the whole class's pre and post-test scores. The lack of significant differences of mean scores serve as evidence for audio and video recorded students' representativeness of their peers.

Interview Insight

Thirteen students were interviewed to gain insight into what conceptions were held by the reform-based students and provide evidence for substantive validity of *ACIDI*. When individuals were asked to interpret *ACIDI* prompts, six of the thirteen students did not successfully offer an interpretation of all questions and only re-read the questions. The remaining seven interviewees that offered interpretations vocalized interpretations that were consistent with the *ACIDI* prompt wording for at least 5 of the 9 *ACIDI* items. Of the items that lacked vocalized interpretations, students offered either incomplete or partially incorrect interpretations of *ACIDI* items.

Interview information regarding questions 1 and 2 revealed that students had scientifically accurate conceptions of resonance stabilization and its impact on conjugate base stability. Representative quotes regarding questions 1 and 2 were extracted and can be seen below.

...That the carboxyl group allows for resonance when there is a double bond next to it. Which de-localizes the negative charge, and makes it like a stronger acid...

...I'm going to go with choice four, saying compound A has the most stable conjugate base, because the hydrogen that would probably be deprotonated is the one in the OH group at the top; and that would cause a resonance in the Benzene ring. The resonance that spreads to the Benzene ring would be an ideal position for the negative charge...

...Well, it is a carboxylic acid but it has the COOH and the O has the resonance...

... But then B has the resonance, which is why it's better able to stabilize the conjugate base...

...I knew that the resonance within that carboxylic acid is much stronger than it would be in the benzene ring...

...because I was thinking that the hydrogen is most acidic in C because of the two oxygen's and their induction. And also that resonance, but then if I say that then, but then, resonance has more of an effect on the acidity than induction. So that kind of makes B more acidic...

A few students had inappropriately evaluated the presence of resonance structures in acetylacetone compared to phenol, shown below (Question 2). These students claimed acetylacetone did not have any resonance structures due to inability to identify an acidic hydrogen. Which would leave behind a negative charge, after deprotonation, to participate in resonance. However, these students successfully identified resonance structures within phenol.

... But now I would construct the argument that B has the resonance through the Benzene ring, while A has no resonance, it just has induction for the hydrogens on carbons relative to the oxygens. But that affect isn't as strong as resonance...

...I would say that B is more acidic than A, because B has the resonance in the ring and A does not...

...For the post test I chose B, that B is stronger than A. Because when we look at OH, B, which the ring, the organic ring with the OH, we can form a resonance structure. It has a lone pair that can participate in a resonance structure, which delocalizes electron density around the hydrogen. So it will be acidic, and A I don't see any hydrogens, So that's why I chose choice one...

One student, below, was aware of both resonance structures and successfully compared their stability and determined that the resonance stabilization within acetylacetone, where you can resonate a negative charge from one carbon to two separate oxygens, was stronger than the resonance stabilization in phenol, where you can resonate a negative charge from one oxygen to three separate carbons in phenol's benzene ring.

...So for that one, I looked at the resonance for structure A, where it has the two double bonds to the oxygen, plus oxygen is pretty electronegative, as opposed to B. Where it does have resonance within like the benzene ring, but I felt like that wasn't quite as strong as A, yeah...

Similarly, students tended to correctly interpret and apply induction to questions 1, 8 and 9 in the form of electron donating and withdrawing groups to the structures of acetic acid and methylphenol.

...Now, I'm looking at the fourth answer, compound C has the most positive acidic hydrogen, because there's the oxygen that acts like through induction is drawing charge away and making it more positive and more willing to dissociate. I think that's probably a better answer than one. Just because saying it's a carboxylic acid doesn't say anything about how acids and bases work, it's just kind of if you remember a trend...

...So C it's more acidic because it's less electron rich, which makes sense because then you don't have the electron donating group...

...I still think C is more acidic than B. Because, B has the methyl group, which is the electron doing group, but C doesn't. So C would just have the resonance with the ring, which will make that hydrogen more positive than B's hydrogen...

...Because to determine the acidity of a compound, and C, is the electron density and compound C has like two negative oxygens. So it will delocalize the electron density around hydrogens, so it will deprotonate easier...

...You would always, any type of inductive effect, you would want it to be pulling away to stabilize it [student referring to conjugate base stability and acidity]...

...Because it is a carboxylic acid, and there is an electronegative double bond here, so the electronegative inductive effect will be there so hydrogen will leave easily...

All students that were interviewed had claimed that they had guessed on the *ACIDI* pre-test, prior to instruction, with some students claiming they had used information gained from high school AP chemistry to formulate pre-test answers. Similarly, all interviewees had determined that they had gained the knowledge to answer post-test questions on the *ACIDI* concept inventory from the lecture or workshop portion of the class. While a few students had also attributed the workshop periods to being settings that they had applied knowledge, gained in lecture, to questions in workshop. Which helped them answer the post-test more appropriately.

Interviewees had discussed several obstacles to choosing answers on *ACIDI*. Several students' decision-making skills for the *ACIDI* concept inventory seemed to be influenced primarily by what they perceived to have been emphasized in lecture, or explicitly told that the concept may appear on a test.

> *…Yeah I don't know, I just saw one, and I knew that, that was the reason. Like that was the reason that they taught us, they didn't really put an emphasis, they didn't really say much about positive acidic hydrogens* [regarding question 1]…
>
> *...For the pretest, I chose one. Which has to do with induction, because that's what we were exposed to. But for the post-test, after we learned about conjugation stuff, I chose two, because that was what they taught us was better than induction* [regarding question 2]…
>
> *…The posttest, i chose two* [C, acetaldehyde, is more acidic than B, acetone, because the conjugate base of C is more stable]. *And I think that one* [C is more acidic than B because C has a hydrogen atom instead of another methyl group] *is better than one, because we learned that..., if you have wait, hold on, sorry. So I chose that two is correct, the correct statement, because in C it has a hydrogen attached to it. As opposed to the methyl, basically that's like the only difference between B & C. And I thought that was correct, because we learned that in class that you know, carbon groups are electron donating. And they're more likely, if there's more substitutions, it's less acidic. So, yeah I thought that was fair answer considering what I learned* [regarding question 6]…
>
> *...Exam one, would be a 3, because I thought it was the most fair. Everything that we went over in class, was on the exam…*

Others had discussed their inability to identify an implicit acidic hydrogen atom, which influenced their comparison of acidity of separate organic structures.

…The only thing, because I've never done organic chemistry before this, so these structures... originally I don't think they stood out to me right away, that the... like when you just draw the line to another carbon, and now I know that there is always three hydrogens attached…

…so I think the answer is one, because of the induction in compound B. And also the resonance structure in A, there's only, I mean there is no hydrogen there. So there's no way for it to be deprotonated… [referring to acetylacetone vs phenol in question 2].

Two interviewees had discussed their difficulties in choosing their answers, as they were unsure if the reason or explanation for acidity ought to be chosen as the best answer to question 4.

… And then, but at this point, I would be kind of like conflicted between option 4 and option 1 because option 4 says it has the more stable conjugate base, which is the reason why it's a good acid because the base can exist on its own. But resonance structure is like the explanation for why it has a more stable conjugate base. So at this point I would be conflicted now, versus in the past. I made the wrong decision, but I was very sure of it…

…At the time, I picked option 3. Basically just because I said the electronegativity. Now, I might pick I might pick 2 or 4. So it has the most stable conjugate base, which is the reason why it's the most acidic, but option 2 explains why it has the more stable conjugate base. So it's like the explanation and the other one is the answer….

…So... I would say that four could also be an answer, because that's why something is acidic. Because it has the most stable conjugate base. But ,I would say that it's the most stable because, it has two electrons withdrawing groups, with his answer two….

…And in terms of one and four, I don't know which is better, because resonance is obviously important. Having a stable conjugate base is obviously important. Maybe I would just say one, because resonance makes it a more stable conjugate base. So it's going to little bit more to the why it's a more stable conjugate base…

Argumentation Analysis

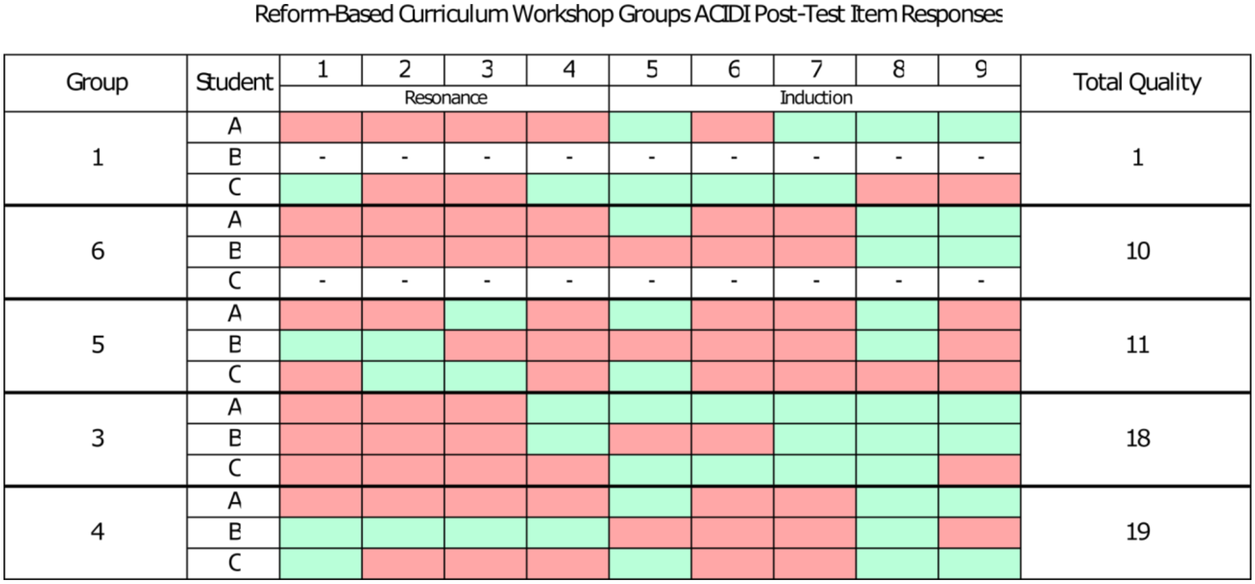
Table 4. Reform-based curriculum workshop group's total argumentation quality compared to ACIDI person ability scores.

Reform-Based General Chemistry *ACIDI* Argumentation Quality

Workshop Group	Level 1	Level 2	Level 3	Level 4	Level 5	Total Quality	Mean Person Ability$_{\text{Pre}}$	Mean Person Ability$_{\text{Post}}$	Person Ability Change$_{\text{Pre-Post}}$
1	1	0	0	0	0	1	-0.178	0.067	0.245
6	0	1	0	1	1	11	-1.332	-1.021	0.311
5	0	2	1	1	0	11	-1.125	-0.710	0.415
4	1	1	2	3	1	26	-1.121	-0.192	0.929
3*	2	0	0	0	4	22	-1.847	0.149	1.996

*Technical difficulty prevented audio recording for one workshop class

Total argumentation quality, shown above in table 4, was determined by summing all quality arguments, and their levels, together. There were two workshop periods that were relevant to *ACIDI*, and arguments created by group 3 were not captured for 1 (50%) workshop period. Group 2 did not vocalize an argument that at least met level 1 requirements (Table 2), therefore they were removed from this analysis. Increasing total argumentation quality seemed to be associated with increasing *ACIDI* pre-post person ability score change, an indicator of conceptual reprioritization towards dominant, scientifically accepted, conceptions of organic acidity (R = 0.72). There were no associations between total argumentation quality and mean group pre or post-test ability scores.



Reform-Based Curriculum Workshop Groups ACIDI Post-Test Item Responses

Group	Student	1	2	3	4	5	6	7	8	9	Total Quality
		Resonance				Induction					
1	A	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟩	1
	B	-	-	-	-	-	-	-	-	-	
	C	🟩	🟥	🟥	🟩	🟩	🟩	🟩	🟥	🟥	
6	A	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	10
	B	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	
	C	-	-	-	-	-	-	-	-	-	
5	A	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟩	🟥	11
	B	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟥	
	C	🟥	🟩	🟩	🟥	🟩	🟥	🟥	🟥	🟥	
3	A	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟩	🟩	18
	B	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟩	
	C	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟩	🟥	
4	A	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	19
	B	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟩	🟥	
	C	🟩	🟥	🟥	🟥	🟩	🟥	🟥	🟩	🟩	

Figure 1. Reform-based curriculum workshop groups' ACIDI post-test item responses compared to total argumentation quality.

Figure 1, above, displays each group's total quality of arguments made during relevant cooperative learning workshops. Group 2 did not vocalize a quality argument, therefore they were removed from analysis. Of the five remaining groups that had created at least one quality argument, an increasing trend in correct *ACIDI* responses appears to be present alongside increasing total quality of arguments made. Furthermore, there seems to be a positive trend of increasingly correct answers to induction questions and total quality of arguments made.

Implications and discussion

Representativeness of Sample

The lack of significant difference between *ACIDI* pre and post-test scores for 16 of the 18 individuals that were audio and video recorded and the whole reform-based general chemistry class suggests that their scores are representative of the whole class.

Interviews & Workshop Discourse

Interviewed students seemed to have an appropriate understanding, and application, of induction and its effect on organic acidity for the molecules presented in questions 1, 8 and 9 on *ACIDI*. Interviewees had shown an ability to connect electron donating and withdrawing effects to increasing/decreasing organic acidity. This connection was evident when interviewees had discussed charge (de)stabilization and how that relates to conjugate base stability, causing a species to be more or less acidic than another. More specifically, these tools were used appropriately for the organic structures of acetic acid (electron withdrawing) and methylphenol (electron donating). This lends support to the claim that the *ACIDI* concept inventory significant score improvements for questions 1, 8 and 9 from pre to post-test may be representative of participating students having dominant, scientifically accepted, induction conceptions and appropriate application of such conceptions.

Furthermore, interviewees had expressed multiple appropriate conceptions regarding resonance (de)stabilization, and its application, for questions 1 and 2 on *ACIDI*. The appropriate application of resonance stabilization had been expressed by interviewees for the structures of acetic acid, acetylacetone and methylphenol. Multiple interviewees were able to apply resonance stabilization to stability of the structure's conjugate base through negative charge delocalization. However, there were numerous interviewees that had revealed a major obstacle to appropriate application of resonance stabilization to acetylacetone as they were unable to identify an acidic hydrogen within acetylacetone to be deprotonated. This suggests that rather than inappropriately answering the question due to incorrect application of conceptions of resonance, interviewees were unable to identify the appropriate structural features that would lead to an opportunity to appropriately apply of conceptions of resonance stabilization. This evidence can then be applied to the increases, and lack thereof, in scores from pre to post-test on *ACIDI*, as the obstacle provides insight into reasoning for incorrect answers being chosen for question 2 (comparing acidity of phenol and acetylacetone), and appropriate application of scientifically accepted resonance stabilization conceptions towards both phenol and acetylacetone. However, more frequent application towards phenol as students were able to identify its most acidic hydrogen.

Interviews provided insight into where students gained knowledge, and how that gain impacted their decision making on *ACIDI*. Interviewees had determined that students had gained the knowledge to answer *ACIDI* items from discussion-oriented lecture, or the cooperative learning workshop, and applied that knowledge in the workshop period. This attribution, paired with the insight gained during individual interviews, suggests two important ideas. First, learning gains made from pre to post-test on *ACIDI*, due to knowledge gained during the discussion-oriented lecture or cooperative learning workshop, allows association of the *Chemical Thinking* curriculum with increased dominant, scientifically accepted, conceptions of resonance stabilization and induction for the reform-based general chemistry curriculum students that were part of this study. However, representativeness of the sample of this study to a more general population of students was not performed due to inability to perform such a test. Therefore we cannot determine the generality of conceptual learning gains, or conceptual reprioritization, on *ACIDI* due to *Chemical Thinking*'s curriculum. Second, *Chemical Thinking* may need to more explicitly address implicit hydrogens that are part of organic structure representations, as some interviewees were unable to appropriately interpret acetylacetone's organic structure.

Two potential obstacles to developing deep conceptual understandings of organic acidity were present within the interviews, as some interviewees were focusing on material that was explicitly covered in lecture/workshop and seemed to tailor their studying habits to what they were told was going to be on their exams, rather than tailoring studying habits to all presented material.

Interviews also provided important insight into how students may be choosing their answers on *ACIDI*, as multiple students referenced discomfort in choosing a "best" answer due to perception of two separate choices as an explanation and the other as a reason. Leading to inability to determine which one would be considered the correct choice. This may be better understood as students having difficulty choosing a major evaluating factor that makes one organic structure more acidic than another (i.e. induction, resonance), or impact of that factor on conjugate base stability, as this issue was encountered when answering question 4. This suggests that the prompt or correct answer choices for question 4 (see supplemental for *ACIDI*, Q4), and their wording, may need to be reviewed to reduce the likelihood of other students encountering this issue.

Argumentation Quality and ACIDI Post-Test Responses

A positive association was evident between total argumentation quality and *ACIDI* pre-post person ability change, which suggests that increasing argumentation quality may be linked to conceptual reprioritization of dominant, scientifically accepted, conceptions of organic acidity. This may be due to increased useful conception exposure as argumentation quality increases, as higher quality arguments are thought to have stronger argument components embedded in them (Erduran, Simon & Osborne, 2004). However, it is not clear that Erduran and co-workers' analytical framework captures useful conceptions.

There appeared to be a positive association between total group argumentation quality and correct responses to *ACIDI* items that targeted induction. This may serve as evidence for an association between argumentation quality and dominant, scientifically accepted, conceptions of induction. Furthermore, it may be possible that increasing conceptions that students are exposed to, and perhaps increasingly useful conceptions, as total argumentation quality increases may be due to the increased number of claims, data, warrants, and rebuttals that are accounted for as argumentation quality increases. However, just the presence of claims, data, warrants, and rebuttals may not be enough to increase the likelihood of conceptual reprioritization to the threshold of it occurring. Perhaps the usefulness of exposed conceptions may not be accounted for by TAP or Erduran's analytical framework, as they do not inherently assess the usefulness of exposed conceptions (Toulmin, 1958; Erduran, Simon & Osborne, 2004).

Summary

Audio and video recorded students' scores were representative of the whole reform-based curriculum class's scores. Interviews of thirteen of the eighteen students of interest provided valuable insight into how they were thinking about questions that the class significantly improved from pre to post-test of *ACIDI*. Interviewees seemed to be appropriately thinking about induction and resonance stabilization when correctly interpreting organic structures in the *ACIDI* concept inventory, particularly methylphenol, acetic acid and acetylacetone. Interviews unveiled an obstacle to appropriately interpreting organic structures was the implicit hydrogens, specifically for acetylacetone. An additional obstacle to appropriately answering the *ACIDI* concept inventory was students' inability to determine if they ought to choose a major evaluating factor (characteristics of structures that lead to a stable conjugate base) when determining acidity of organic structures or the result of that major evaluating factor (i.e. stable conjugate base). Interviewees also seemed to be tailoring their studying habits to what material was explicitly discussed in lecture or explicitly told to be on their exam. Argumentation quality, in the form of accumulated quality of arguments made, seemed to be positively associated with conceptual reprioritization towards dominant, scientifically accepted, conceptions of organic acidity. Finally, increasing argumentation quality was positively associated with dominant, scientifically accepted, conceptions of induction.

Limitations

Sample size of audio and video recorded groups and students was limited to the number of separate workshop sections involved in this study's general chemistry course and more importantly limited to the number of cameras and microphones available to record workshop groups. Discourse and argumentation analysis was limited to audible verbalizations in the workshop portion of the reform-based course. Coding of workshop transcripts are limited to identification of components and not the correctness of components. All insight in this study is limited to the study's population, as representativeness of the study sample's conceptions to a more general student population was unable to be assessed. Audio and video recorded groups were not randomly selected, limiting their representativeness of their peers.

Future Direction

Several modifications can be considered to improve future experiments that are similar to this study. First, increasing sample size of students who were audio and video recorded in a workshop, or workshop equivalent, setting. This can be done by increasing the number of workshop sections that are part of the course or increasing the amount of audio and video equipment to be able to capture student discussions and arguments in the workshop setting. Second, increasing the number of workshops with relevant *ACIDI* material would allow for increased collection of conceptions regarding organic acidity. Third, increasing the number of individual interviews can provide more diverse insight into conceptions of organic acidity within the tested student population, and increase the explanatory power given by interviewees. Fourth, continual modification of the *ACIDI* concept inventory choices, and increasing the number of items may increase the reliability of the instrument and more accurately reflect student conceptions of organic acidity. Fifth, creation of a database of student responses to the *ACIDI* concept inventory would allow for assessment of representativeness of tested student populations. Finally, the creation of an analytical framework that is capable, or best suited, of capturing useful conceptions may provide more insight into the association of argumentation quality and increasing dominant, scientifically accepted, conceptions.

Chapter 5: Association of Argumentation Quality and Scientifically Accepted Redox Conceptions

Introduction

Background

Chemistry student conceptions of redox have been primarily investigated from a symbolic perspective, use of chemical equations, rather than a particulate perspective such as visual depictions of electron transfer (Brandreit & Bretz, 2014; Rosenthal & Sanger, 2012; Garnett & Treagust, 1992). Misconceptions have been reported from both conceptions, as a reported symbolic misconception is inappropriate assignment of oxidation numbers to whole molecules, rather than the atoms that make up the molecules (Rosenthal & Sanger, 2012). While a particulate misconception is one of free electron movement through solution, independent of ions (Garnett & Treagust, 1992). The previously mentioned misconceptions, as well as others, may have been created, and continue to exist, due to the difficulty of connecting different chemical perspectives (Brandreit & Bretz, 2014). This may increase the difficulty of scientifically accepted conceptions dominating within students minds.

Cooperative learning practices have been commonly adopted in undergraduate chemistry courses, as they have been associated with increased undergraduate student performance in chemistry (Bowen, 2000; Freeman et al. 2014). Cooperative learning offers students the opportunities to be exposed to conceptions held by themselves and their peers, which may increase the likelihood of conceptual reprioritization occuring (Bowen, 2000; Freeman et al.

2014; Shah et al. 2018; Shtulman & Lombrozo, 2016). Conceptual reprioritization is the restructuring of conceptual dominance hierarchies that may occur after a competition between useful conceptions. However, there is no literature regarding associations of undergraduate student argumentation and their conceptions of oxidation-reduction. Cooperative learning environments may increase the likelihood of conceptual reprioritization as they promote student-student collaboration, which may take form as discussions regarding course material (Shah et al. 2018; Nussbaum, Sinatra & Poliquin, 2008; Shtulman & Lombrozo, 2016). Argumentation, a form of discussion, may take place during the discussions that occur in cooperative learning environments and increase the likelihood of students being exposed to useful conceptions (Shah et al. 2018; Shtulman & Lomborozo, 2016). Diversification of learning groups may impact exposure to conceptions, and perhaps useful conceptions, as more diverse groups have been linked to increased problem-solving and critical thinking skills (Hurtado, 2001).

Traditional, college-level, general chemistry curricula have been reported to be composed of an inconsistent string of theories that focus on algebraic proficiency of undergraduate students rather than conceptual understanding of chemistry (Talanquer, 2013; Van Berkel et al. 2000; Kuhn, 1963; Shah et al. 2018). This decreased focus on conceptual understanding may decrease the likelihood of conceptual reprioritization occurring, as students may not be exposed to as many chemistry conceptions (Shtulman & Lomboroz, 2016). *Chemical Thinking*, a reform-based general chemistry curriculum created by Talanquer and associates (Talanquer & Pollard, 2010), has shifted the focus from algebraic proficiency to fundamental questions of chemical practice (i.e. How do we distinguish substances?). This emphasis on fundamental questions of chemical practice, may promote a deeper conceptual understanding of chemistry, as it may increase exposure to chemistry conceptions (Talanquer & Pollard, 201; Talanquer, 2013; Sevian &

Talanquer, 2014; Shtulman & Lombrozo, 2016). *Chemical Thinking* promotes conceptual understanding of chemistry through a discussion-oriented lecture, led by the instructor, that offers students opportunities to be exposed to their own conceptions, as well as conceptions of others (Talanquer & Pollard, 2010). Increased exposure to conceptions may offer students more appropriate chances restructure their conceptual dominance hierarchies (Shtulman & Lombrozo, 2016). *Chemical Thinking* also utilizes cooperative learning activities that place students into small groups and encourage argumentation through its promotion of student-student collaboration to solve activity problems (Talanquer & Pollard, 2010). Argumentation quality may also play a role in the restructuring of conceptual dominance hierarchies, as a current analytical framework of argumentation quality takes account for argumentation components (i.e. rebuttal) that are thought to be part of stronger arguments (Erduran, Simon & Osborne, 2004). However, possible association(s) of argumentation quality and undergraduate student redox conceptions has yet to be investigated. A possible obstacle that may interrupt the investigation of association(s) between argumentation quality and undergraduate students' conceptions of redox chemistry is the lack of alignment of college-level curricula and measurement tools that are capable of capturing student conceptions (Shah et al. 2018).

Chemistry education researchers have recently begun developing valid instruments, concept inventories, that are designed capture student conceptions of chemistry material (Libarkin, 2008; Brandreit & Bretz, 2014). These instruments attempt to capture both scientifically accepted and alternative conceptions held by students using multiple choice questions that target previously reported conceptions of chemistry (Libarkin, 2008). Repeated use of concept inventories allow researchers to track dominant conceptions held by students at different points in time, which renders these instruments useful for capturing conceptual reprioritization (Shtulman & Lomboroz, 2016; Libarkin, 2008). Concept inventories, when paired with argumentation quality assessment, allow for investigation of a possible association between argumentation quality and conceptual reprioritization. Furthermore, this investigation may be more appropriately executed if general chemistry curricula align with the promotion of conceptual understanding.

Theoretical Framework

Vygotsky's social constructivism theory supports the use of cooperative learning, and discussion-based instruction to promote conceptual reprioritization, as social constructivism posits the social construction of knowledge (Vygotsky, 1962). Cooperative learning environments and discussion-based lectures provide a platform for the social construction of knowledge through student-student and student-instructor collaboration (Vygotsky, 1962; Shah et al. 2018; Sevian & Talanquer, 2014; Talanquer & Pollard, 2010). Discussions or arguments that take place in discussion-based lectures of cooperative learning workshop activities, may increase student exposure to conceptions, and perhaps useful conceptions, as students may verbalize their conceptions when solving lecture questions or workshop activities. *Chemical Thinking* uses "let's think" activities, to create class discussions, and may promote student discussion and argumentation through its workshop activities (i.e. justify your choice), which in turn may increase the odds of conceptual reprioritization occuring (Talanquer & Pollard, 2010; Shtulman & Lombrozo, 2016). The alignment of social constructivism with *Chemical Thinking* and its lecture and workshop activities, compared to traditional general chemistry curricula, allows a strong opportunity to capture conceptual reprioritization in undergraduate students' redox conceptions and investigate a potential association between argumentation quality and dominant, scientifically accepted, redox conceptions held by undergraduates (Shtulman & Lombrozo, 2016; Talanquer & Pollard, 2010; Shah et al. 2018).

Discussion & Argumentation

Argumentation, a form of discussion, has been previously defined as “A verbal activity oriented towards the realization of a goal” (Micheli, 2011). Argumentation may aid in the social construction of knowledge and may even be the primary form of communication and social construction of knowledge within scientific communities, as scientists tend to argue on behalf of the stories they try to tell through publications (Shtulman & Lombrozo, 2016; Vygotsky, 1962; Potvin, 2017; Shah et al. 2018). The use of argumentation in science courses has been on the rise, implying its lesser use in the past, which may have impacted students’ ability to argue and ultimately reprioritze their conceptions (Talanquer, 2013; McNeil & Pimentel, 2010; Cohen, 1994; Sampson & Clark, 2009). Student participation in discussion and argumentation has been tough an shown to be done to varying degrees, which may impact benefits that students may take away from arguments (Cohen, 1994; Sampson & Clark, 2009). Student participation in argumentation may be active, vocalize arguments or argument components, or passive, strictly listening to vocalized arguments or argument components. Active participation in discussion and argumentation has been associated with increased student ability to apply discussed ideas to questions, which may lead to stronger conceptual understanding (Cohen, 1994; Mason, 1998). Increasing argumentation quality may also lead to stronger conceptual understanding, and perhaps conceptual reprioritization, as previously structured frameworks that asses argumentation quality take argumentation components, that are thought to be part of stronger arguments, into account (Erduran, Simon & Osborne, 2004). However, No research has been performed to investigate associations of argumentation with undergraduate students' conceptions, and potential reprioritization of such conceptions, of oxidation-reduction.

Toulmin's Argumentation Pattern (TAP) & Argumentation Quality

Toulmin's argumentation pattern (TAP) is capable of identifying argumentation components that are vocalized in arguments (Toulmin, 1958). TAP breaks down arguments into six components: claim, data, warrant, rebuttal, qualifier and backings. Kulatunga and associates combined TAP components to create a basic argument (BA), which requires three components: claim, data and warrant (Kulatunga et al. 2014). The claims is used to put forth an idea, data is referenced to support the claim, and a warrant functions to connect the data to the claim (Kulatunga et al. 2014). Definitions of the components, and a basic argument, are below in table 1.

Table 1. Argumentation components adapted from Toulmin (1958) & Kulatunga et al. (2014).

Component	Definition
Claim	An assertion put forth to the public regarding the topic/question of interest.
Data	Facts or information used to support a claim.
Warrant	A justified connection between data and a claim.
Backing	Assumptions under which the warrant holds power.
Qualifier	Conditions under which a claim is true.
Rebuttal	Refutations that may undermine a previous claim.
Basic Argument (BA)	A verbal utterance that contains a claim, data and warrant connecting the data to the claim.

TAP can be used to identify these components in transcripts of student-student discussions or arguments. However, TAP alone does not assess argumentation quality. Rather, Argumentation components, and their combinations, may be utilized to assess argumentation quality. Erduran and associates have created an analytical framework (Erduran, Simon & Osborne, 2004) based on TAP to assess argumentation quality, which can be seen below in Table 2. Cooperative learning activities and discussion-oriented instruction may encourage students to argue. Those arguments may increase the likelihood of participating students undergoing conceptual reprioritization, and their quality may also be assessed (Shtulman & Lombrozo, 2016;

Shah et al. 2018; McNeil & Pimentel, 2010). This allows for potential investigation of association between argumentation quality and conceptual reprioritization. However, There has been no previous research investigating association(s) of argumentation quality and undergraduate general chemistry students' conceptual reprioritization of redox concepts. Therefore making such an investigation necessary to gain insight into potential association(s), or lack thereof.

Table 2. Argumentation quality framework adapted from Erduran, Simon & Osborne (2004).

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

Rationale & Research Questions

Previous reports have associated argumentation and discussion with student conceptions of science (Bell & Linn, 2000; McNeil & Pimentel, 2010; Shah et al. 2018; Nussbaum, Sinatra & Poliquin, 2009). However, no investigation has been performed to assess argumentation quality and its association to undergraduate students' redox conceptions. The alignment of social constructivism with a reform-based general chemistry curriculum, compared to traditional general chemistry curricula, through its discussion-oriented lecture and promotion of argumentation in cooperative learning workshop activities provides an opportunity to investigate such association(s), or lack thereof.

The main objectives of such an investigation include: provide insight to quantitative results captured by *ROXCI* through interview analysis and analyzing argumentation quality with regard to dominant, scientifically accepted, redox conceptions held by undergraduate general chemistry students. The following research questions were formulated with these objectives in mind:

1. What insight can individual interviews provide regarding redox conceptions held by traditional general chemistry curriculum students?
2. Is there any evidence for association of argumentation quality and dominant, scientifically accepted, redox conceptions held by undergraduate general chemistry students?

Methods

Research Design

This study was conducted during the Fall 2017 semester at a large public university in the northeastern region of the united states. The following protocol was executed after IRB approval. Two general chemistry classes were given the *ROXCI* concept inventory twice, for 20 minutes per assessment. One general chemistry course was part of a first-year, reform-based chemistry-organic chemistry sequence, and used the reform-based general chemistry curriculum (RBC), *Chemical Thinking*. The other general chemistry course was an off-sequence general chemistry II course that used a traditional general chemistry curriculum (TC).

Each course offered a lecture class three times per week for a total of 160 minutes/week, and a cooperative learning workshop once per week for a total of 80 minutes per week. The first administration, collection of initial redox conceptions, of *ROXCI* was done before relevant course instruction during the eighth week of the fall 2017 semester for the TC course, and the twelfth week of the RBC course. Relevant class instruction then began for four weeks in the TC course, and one week in the RBC course. Relevant topic of instruction include: oxidation numbers, electron transfer, electrostatics and bonding, spectator ions, electrochemistry and balancing redox reactions.

The RBC course included discussion-oriented lectures that were led by the instructor, while the TC course lecture was mostly didactic, with a few clicker questions per class. Both courses included cooperative learning workshop activities, however the RBC placed more emphasis on argumentation through its activities (i.e. Justify your choice) compared to the TC course, which emphasized algebraic proficiency (i.e. Calculate the cell potential of...). The RBC cooperative learning workshops were facilitated by undergraduate and graduate teaching assistants and followed a repeated cycle of 30 minutes group work/problem solving, then 10 minutes of class discussion of half of the workshop activity problems. The TC course cooperative learning workshops were facilitated by graduate TAs and all class time was dedicated to group work/problem-solving. Both class' cooperative learning groups were created with emphasis on group diversity (i.e. race, ethnicity), as diverse learning groups have been linked to increased critical thinking and problem-solving skills (Hurtado, 2001). The second administration, collection of redox conceptions after relevant instruction, was done during the thirteenth week of the TC course, and the fourteenth week of the RBC course.

18 students per general chemistry course, for a total of 36 students, were audio and video recorded during all weeks of the fall semester. Relevant cooperative learning workshops were transcribed, coded using TAP, and analyzed using Erduran's analytical framework for argument quality assessment. The RBC course included one relevant workshop activity during week thirteen of the fall semester, while the TC course included two relevant workshop activities during the eleventh and twelfth week of the fall semester.

Five TC students were interviewed in a semi-structured format for approximately 60 minutes during the fifteenth and sixteenth week of the semester. Interviews were then analyzed to gain insight of how TC students were thinking of *ROXCI* items and answer choices. No reform-based curriculum students were interviewed due to time commitments. A summary of the *ROXCI* research design can be found below in table 3.

Table 3. ROXCI research design summary.

Reform-Based General Chemistry (Week of Fall Semester)	Traditional General Chemistry (Week of Fall Semester)	Action
12	8	*ROXCI* Pre-Test Administration
13	11-12	Audio and video recording relevant workshop discussions (80 mins/week)
-	15-16	Individual Interviews (60 mins)
14	13	*ROXCI* Post-Test Administration

Data Sample

Two separate general chemistry courses were involved in this study. One course was a first-year, reform-based general chemistry course that used a reform-based general chemistry curriculum (RBC). The RBC class was composed of students who completed at least two years of high school chemistry (i.e. Intro Chemistry and IB/AP Chemistry). Of the 179 total students enrolled in the course, 18 consenting students, or six groups, were audio and video recorded during the workshop portion of their course. All groups created at least one level 1 quality argument with at least two group members present and two group members completed the *ROXCI* post-test. No RBC students were interviewed in this study due to time conflicts.

The other course was an off-sequence general chemistry II course that used a traditional general chemistry curriculum (TC). The TC course was composed of students who passed general chemistry I in the previous spring semester or earlier, failed general chemistry II in the previous spring semester, and transfer students. Of the 193 students enrolled in the TC course, 18 were audio and video recorded during the workshop portion of their course. Three of the six (1, 2 and 6) workshop groups had created at least one level 1 quality argument with at least two group members present and completed the ROXCI pre and post-tests, while three groups (2, 5 and 6) has created at least one level 1 argument with at least two group members and completed the *ROXCI* post-test. Five of the 18 audio and video recorded TC students were individually interviewed to provide insight of how TC students were thinking about *ROXCI* items and answer choices.

Representativeness of Sample

Audio and video recorded group *ROXCI* scores were compared to scores of their respective classes to determine if their scores were representative of their peers. This comparison was done using an independent samples t-test, assuming equal variances, in IBM SPSS Statistics version 23. The results of the independent samples t-test serve as evidence for the representativeness of audio and video recorded groups to their respective peers.

Interviews

Five of the eighteen traditional curriculum students that were audio and video recorded during the workshop period of their course were individually interviewed for about an hour to gain insight into how students were thinking about *ROXCI* prompts and answer choices, and topics covered by the *ROXCI* assessment. Interviews were performed in a semi-structured manner and asked students several questions regarding *ROXCI* items and their respective course (see appendix for interview protocol). Interviews took place during the final weeks of the fall 2017 semester. Interviews were analyzed to extract information regarding how students were interpreting *ROXCI* prompts and thinking about *ROXCI* answer choices. Questions 4 and 7 were prioritized during analysis and extraction of interview information because they were the two questions the TC students significantly improved on from pre to post-test. Questions 4 and 7 targeted the topics of oxidation numbers, surface features of oxidation-reduction reactions, and electron transfer.

Reform-based curriculum students were unable to be individually interviewed about their thought processes regarding oxidation-reduction and the *ROXCI* assessment due to time constraints.

Argumentation Analysis

Workshop discourse of the students of interest, those who were audio and video recorded, was collected, transcribed, coded and analyzed to gain insight into how students were arguing *ROXCI* related material. Workshop transcripts were coded according to Toulmin's argumentation pattern, or TAP (Toulmin, 1958). Two independent researchers coded all relevant *ROXCI* workshop transcripts, then compared and resolved coding conflicts with at least 90% agreement. Coded argumentation components are defined below in Table 1.

Table 1. Argumentation components adapted from Toulmin (1958) & Kulatunga et al. (2014).

Component	Definition
Claim	An assertion put forth to the public regarding the topic/question of interest.
Data	Facts or information used to support a claim.
Warrant	A justified connection between data and a claim.
Backing	Assumptions under which the warrant holds power.
Qualifier	Conditions under which a claim is true.
Rebuttal	Refutations that may undermine a previous claim.
Basic Argument (BA)	A verbal utterance that contains a claim, data and warrant connecting the data to the claim.

Quality of arguments created during workshop were assessed using Erduran's analytical framework (Erduran, Simon & Osborne, 2004), shown below in Table 2.

Table 2. Argumentation quality framework adapted from Erduran, Simon & Osborne (2004).

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

Further clarification of how particular language used in the criteria of level 3, 4, and 5 arguments must be addressed. A “weak rebuttal,” part of a level 3 argument, was understood by the coders of this study as a simple refutation of a previous claim and do not go further than a simple “No, I don’t think so”-like verbalization. A “clearly identifiable rebuttal,” part of a level 4 argument, was understood as a verbalization that involves an “I don’t think so”-like aspect and goes a step further with additional data or information that would further weaken a previous claim. An “extended argument,” part of a level 5 argument, was understood as an argument with numerous claims and at least two rebuttals. Examples of argumentation quality can be found in the appendix.

Argumentation quality was analyzed by calculating total argumentation quality, the sum of all quality arguments made during relevant workshops for both TC and RBC samples. Total argumentation quality was then compared to *ROXCI* person ability scores to investigate association between argumentation quality and dominant, scientifically accepted, redox conceptions. Strength of association was investigated using the correlation function in Microsoft Excel version 1805.

Results

Representativeness of Sample

Representativeness of groups of interest in both courses were assessed using an independent samples t-test comparing *ROXCI* pre and post-test scores between individuals within the groups of interest and their respective classes. Reform-based curriculum groups of interest *ROXCI* pre-test scores were representative of their class (df = 147, t-stat = 2.790, p = 0..133). Similarly, their post-test scores were also representative of their class (df = 147, t-stat = 0.811, p = 0.0.53). Traditional curriculum groups of interest *ROXCI* pre and post test scores were representative of their class (df_{pre} = 86, $t\text{-}stat_{pre}$ = 1.317, p_{pre} = 0.600, df_{post} = 86, $t\text{-}stat_{post}$ = 1.098, p_{post} = 0.507). Lack of significant difference between mean pre and post-test scores for both samples of audio and video recorded students serves as evidence for their representativeness of their peers.

Interview Insight

Five traditional curriculum (TC) students were interviewed to gain insight into how students were thinking about *ROXCI* items that TC students significantly improved on from pre to post-test, how students interpreted *ROXCI* prompts and answer choices, and where students gained the knowledge to answer the *ROXCI* items.

Four of the five students who were interviewed offered interpretations to *ROXCI* item prompts. One student did not vocalize interpretations of the *ROXCI* prompts. Of the four interviewees who interpreted *ROXCI* item prompts, one interpreted six prompts appropriately (interpretation consistent with wording of prompt), while the other 12 item prompt interpretations were not verbalized. The other three students appropriately interpreted at least 16 of the 18 item prompts. Interviewees had also discussed various methods of how they had gained knowledge to appropriately answer the questions posed in *ROXCI*, as two claimed lecture/workshop were the primary modes of gaining knowledge, one claimed the homework program, Aleks, and two others cited knowledge from previous general chemistry courses in high school or their general chemistry I class.

TC students expressed ideas about oxidation numbers when discussing questions 3, 4 and 7 during the individual interviews. Questions 3 and 4 are two-tiered (question 3 is a question, and 4 is the reason for the answer to question 3). Representative interview quotes of how interviewees were discussing oxidation numbers are listed below.

> *…The charge on the aluminum changes and oxygen changes also. So, aluminum, on the reactant side, the charge is zero and the same thing for oxygen on the reactant side. And then in the product side aluminum the charge is +3 and in oxygen it is -2. So aluminum is being oxidized and oxygen is being reduced…*
>
> *...And N is being reduced because in the reactants, N is +5 and in the products it is +4, so it is being reduced…*
>
> *...The charge on Al changes which is the common definition of oxidation reduction. And the charge on O to also changes. That is kind of the point of an oxidation reduction reaction…*
>
> *...So in fact nitrogen is going from plus five to plus four it looks like. So it means it's actually B for this one. Nitrogen is reduced because H is actually staying the same…*
>
> *...Cause like, I don't think compound have an oxidation number. But like I could see how you would think that like the* NO_3*, it looks like it becomes more positive, even though like you're losing an oxygen. If you've not seen that…*

A combination reaction of aluminum and oxygen occurs to form aluminum oxide in questions 3 and 4. Interviewees seem to have associated increasing or decreasing oxidation numbers with identification of this combination reaction, and translated that to an oxidation-reduction reaction.

One interviewee had even addressed a common misconception of assigning oxidation numbers to whole compounds, although this misconception (seen below) seemed to be partially present in another interviewee's thoughts on question 7 (a redox reaction involving copper and nitrate).

> *…I think it is C because NO_3^-, so the overall charge is minus one and then it becomes NO_2 so the overall charge is zero. So a negative number to zero would mean it is being oxidized…*

This student seems to have has associated NO_3^-'s charge with the charge of NO_2, even though NO_3 and NO_2 are different molecules. Furthermore, the student seems to imply that the charges of the molecules are oxidation numbers, and since they change in the positive direction, then oxidation has occurred from NO_3^- to NO_2.

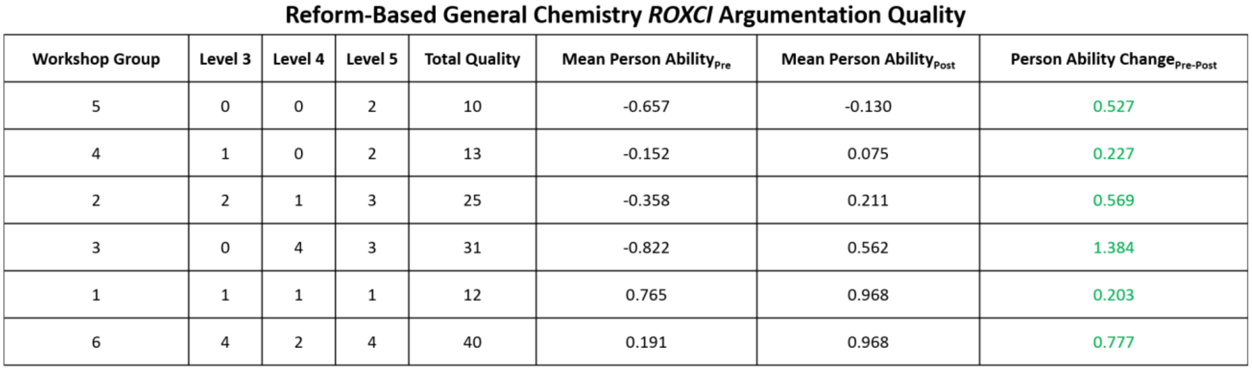
Interviewees also provided insight into identification of electron transfer occurring in questions 3 and 4.

> *…choose B as the answer because in the equation you see that O_2 gives electrons to Al in order to form a bond…*
>
> *…O_2 gives electrons to Al to form a bond…*
>
> *…I think O_2 gets reduced and it becomes more negative. So it doesn't give electrons…*

Two interviewees have claimed that O_2 gives electrons to Al in the combination reaction that occurs in the formation of aluminum oxide and is referenced by questions 3 and 4, while another interviewee has claimed the opposite, that O_2 does not give electrons in the combination reaction.

Argumentation Analysis

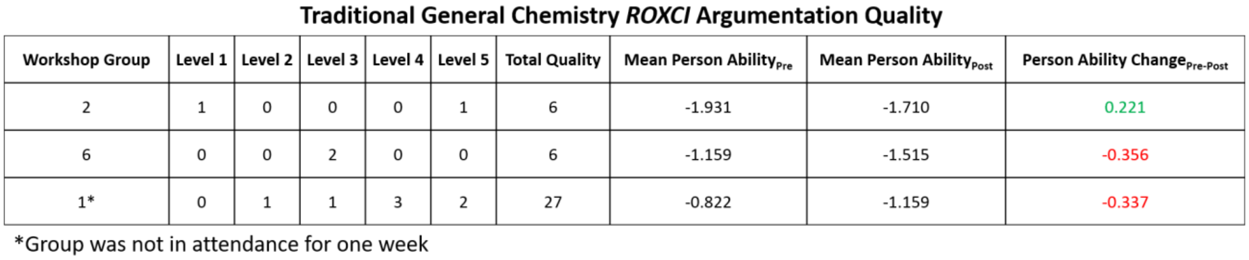
Table 4. Reform-based curriculum students' total argumentation quality compared to ROXCI person ability and change in person ability scores.



Reform-Based General Chemistry *ROXCI* Argumentation Quality

Workshop Group	Level 3	Level 4	Level 5	Total Quality	Mean Person Ability$_{Pre}$	Mean Person Ability$_{Post}$	Person Ability Change$_{Pre\text{-}Post}$
5	0	0	2	10	-0.657	-0.130	0.527
4	1	0	2	13	-0.152	0.075	0.227
2	2	1	3	25	-0.358	0.211	0.569
3	0	4	3	31	-0.822	0.562	1.384
1	1	1	1	12	0.765	0.968	0.203
6	4	2	4	40	0.191	0.968	0.777

Table 4. Traditional curriculum workshop group argumentation quality compared to ROXCI person ability and change in person ability scores.



Traditional General Chemistry *ROXCI* Argumentation Quality

Workshop Group	Level 1	Level 2	Level 3	Level 4	Level 5	Total Quality	Mean Person Ability$_{Pre}$	Mean Person Ability$_{Post}$	Person Ability Change$_{Pre\text{-}Post}$
2	1	0	0	0	1	6	-1.931	-1.710	0.221
6	0	0	2	0	0	6	-1.159	-1.515	-0.356
1*	0	1	1	3	2	27	-0.822	-1.159	-0.337

*Group was not in attendance for one week

Total argumentation quality of RBC students' arguments was assessed by summing all quality arguments together, shown above in Table 4. When total argumentation quality was compared to *ROXCI* person ability scores, a indicator of dominant, scientifically accepted, conceptions; a positive associative trend can be seen between total argumentation quality and increasing mean group post-test person ability (R = 0.52). An associative trend between total argumentation quality and person ability change, a score indicative of conceptual reprioritization, did not seem to be present. No RBC group created an argument below level 3 quality during their one relevant workshop period.

Total argumentation quality of TC workshop group arguments, shown above in table 5, compared to mean group person ability revealed differing results compared to the RBC workshop groups. Not enough evidence for an association between argumentation quality and increasing dominant, scientifically accepted, redox conceptions was collected. However, TC students had created arguments that ranged from level 1-5, as opposed to RBC students creating arguments from level 3-5.

Implications and discussion

Representativeness of Sample

The lack of significant differences between the reform-based curriculum class and groups of interest for the ROXCI pre-test suggest that groups' pre and post-test scores are representative of their class.

The traditional curriculum groups' *ROXCI* pre and post-test scores were not significantly different than their class, suggesting that the *ROXCI* pre and post-test scores of the groups are representative of their class.

Interviews Insight

Interpretations of *ROXCI* items by traditional curriculum students being consistent with the wording of the item prompts suggest that interviewed students were appropriately interpreting the *ROXCI* item prompts, which supports the instrument's substantive validity. TC interviewees provided insight to where they gained knowledge to answer the *ROXCI* items, suggesting that conceptual gains made on *ROXCI* may be due to various portions of the traditional general chemistry curriculum or not due to the curriculum used, as two students suggested lecture/workshop, one suggested the hw program and the final two suggested prior chemistry knowledge from high school chemistry or general chemistry I. This warrants caution as to how the TC impacted their redox conceptions, and potentially conceptual reprioritization.

Interviews of the traditional curriculum students provided insight how they were thinking of oxidation numbers and electron transfer. Vocalized conceptions of oxidation numbers in the interviews suggests that scientifically accepted and alternative conceptions are present in the traditional curriculum population, as interviewees had appropriately associated changes in oxidation numbers, in the combination reaction of aluminum and oxygen to form aluminum oxide, with an oxidation-reduction reaction. However, the misconception of compounds having an oxidation number may still be present within the population. Furthermore, as indicated by one interviewee, misinterpretation of reactants and their products may have led to the propagation of the misconception.

Conceptions regarding electron transfer that were exposed in the interviews suggest that there are scientifically accepted and alternative conceptions regarding how electrons transfer, in the combination reaction of aluminum and oxygen to form aluminum oxide, as two interviewees had inappropriately expressed electrons transfer from oxygen to aluminum, while one interviewee had appropriately expressed electrons transferred from aluminum to oxygen, using the reduction in oxygen's oxidation number in the products as support for their claim.

The presence of these oxidation number and electron transfer conceptions may suggest that the traditional general chemistry curriculum may not be sensitive enough to the ways in which the traditional curriculum students interpret reactants and products of an oxidation-reduction reaction. It is possible that traditional curriculum students may require increased discussion to be exposed to conceptions that align with their particular learning style to more appropriately interpret reactants and products of an oxidation-reduction reaction. However, the traditional curriculum may be capable of providing students the opportunity to appropriately apply oxidation numbers to identify an oxidation-reduction reaction. Although, mixed results were present regarding students' conceptions of electron transfer in a combination reaction.

Argumentation Quality Analysis

Reform-based (RBC) students increased their mean group post-test person ability as total argumentation quality increased. This may be due to the number of useful conceptions that students are exposed to, which may increase as total argumentation quality increases, as argumentation quality levels account for stronger argumentation components (i.e. rebuttal) being present in higher quality arguments (Erduran, Simon & Osborne, 2004; Shtulman & Lombrozo, 2016; Potvin, 2017). However, there did not seem to be an associative trend between mean person ability change for either RBC or TC students, which would suggest that conceptual reprioritization may not be connected to argumentation quality. Similarly, no associative trend was seen between TC students' group argumentation quality and mean post-test *ROXCI* ability scores. Although this is only one study, and it may not be the last to investigate such an association. Therefore, the lack of evident association between conceptual reprioritization towards dominant, scientifically accepted, redox conceptions, indicated by mean person ability change, and argumentation quality ought to encourage continued investigation. This would create a stronger record of data that suggest an association, or lack thereof, between argumentation quality and desired conceptual reprioritization.

Summary

Scores of the individuals that made up the workshop groups of interest, groups that were audio and video recorded during workshop, were representative of their respective classes. Interviews revealed that traditional curriculum students may be appropriately interpreting *ROXCI* item prompts, suggesting the *ROXCI* assessment prompts are appropriate for the traditional curriculum students.

Furthermore, traditional curriculum interviewees indicated that they had gained their knowledge from various aspects of their course (i.e. homework program, lecture, workshop), or from their previous high school or general chemistry I course. Traditional curriculum student interviews indicated a mix of scientifically accepted and alternative redox conceptions regarding oxidation numbers and electron transfer being present in their traditional curriculum student population.

Workshop argumentation analysis of only reform-based general chemistry curriculum workshop groups uncovered a positive association between total argumentation quality and mean *ROXCI* post-test person ability. This may also suggest a positive association between increasing total argumentation quality and dominant, scientifically accepted, redox conceptions. No association between conceptual reprioritization towards dominant, scientifically accepted, redox conceptions and increasing total argumentation quality was evident. However, more investigative studies must be done to form a stronger conclusion regarding an association, or lack thereof, between argumentation quality and conceptual reprioritization.

Limitations

Discourse and argumentation analysis was limited to audible verbalizations in the workshop periods of each course. Coding of workshop transcripts were limited to identification of components and not the correctness of components. Furthermore, the statistical analyses, and their power, were limited to the number of individuals and groups in the workshop setting and the number of groups that had members who successfully completed the *ROXCI* pre and post-test. Interview insight was limited to the traditional curriculum students due to time constraints preventing additional interview of the reform-based curriculum students. All insight gained by this study is limited to this study's populations, as representativeness of the study's samples conceptions to a general student population was not able to be assessed. Audio and video recorded students were not randomly chosen, limiting their representativeness to their peers.

Future Direction

This study may provide insight into numerous modifications for future studies related to *ROXCI* and argumentation. First, to increase statistical power of the results of argumentation analysis, more groups and individuals ought to be recorded in the workshop, or workshop-like, setting. This can be done by increasing the number of sections of a general chemistry course that are audio and video recorded, increasing the amount of audio and video equipment to be capable of recording more students' oxidation-reduction discussions, and offer stronger incentives to participants to increase the probability of the participants fulfilling the study expectations.

Second, increasing the amount of relevant workshop material to be captured by audio and video equipment may increase the number of captured student conceptions. Third, increasing the quantity of individual interviewees would allow for more meaningful insight of how students may be thinking about oxidation-reduction. Fourth, the creation of a database of *ROXCI* responses would allow for the assessment, and possible generalization, of conceptual learning gains made by students that are captured by *ROXCI*. It may also provide insight to instructors for what oxidation-reduction material requires more careful teaching or more time for students to develop a deep conceptual understanding.

Finally, to convince instructors to adopt curricula that emphasize the value of conceptual understanding of *ROXCI* material, a proficiency standard, that is viewed as meaningful and achievable, must be created to assess associations between conceptual understanding and that standard. If there are positive associations between conceptual understanding of oxidation-reduction material and a proficiency standard, adoption of curricula that emphasize conceptual understanding may be viewed as ideal.

An increased number of studies regarding the *ROXCI* assessment and general chemistry students' conceptions of oxidation-reduction must be done to continue to update what conceptions are accurately held, as there have been no other studies (aside from Brandreit & Bretz, 2014) investigating general chemistry students' conceptions of oxidation-reduction using the *ROXCI* assessment. Furthermore, more chemistry education studies must be done to more fully understand associations, or lack thereof, between argumentation quality and conceptual reprioritization. An increased number of such studies would allow instructors to identify argument quality levels that may be associated with conceptual reprioritization. More importantly, the identification of prompts that elicit quality arguments may provide insight into how to structure workshop activities for students to gain the most meaningful conceptual learning experience regarding oxidation-reduction material. However, the previous suggestion is not strictly limited to the topic of oxidation-reduction, as other difficult general chemistry topics may be investigated, or the subject of chemistry, as these questions have also not been investigated across other STEM contexts.

Chapter 6: Conclusion

Summary

There are several conclusions and implications that can be inferred from the results of this study. Valuable insight has been gained of undergraduate general chemistry students' conceptions of organic acidity, and oxidation-reduction from student interviews. Increased validity and reliability statistics of the *ROXCI* and *ACIDI* instruments were recorded. While other insight of the *ROXCI* and *ACIDI* instruments has been uncovered, such as their ability to separate student ability, and suggestions towards future modifications and purposes of their use. Evidence for association between argumentation quality undergraduate students' dominant, scientifically accepted, conceptions of reprioritization of conceptions of oxidation-reduction was recorded. Similarly, evidence for association of argumentation quality and undergraduate general chemistry students' conceptual reprioritization towards dominant, scientifically accepted, conceptions of organic acidity was reported.

ACIDI

The *ACIDI* -concept inventory was revealed to be unidimensional and appropriately difficult for the reform-based (RBC) general chemistry curriculum students after instruction. *ACIDI* item prompts were commonly interpreted in an appropriate manner, while some answer choices' wording interrupted RBC interviewee's ability to determine the best answer to *ACIDI* items. RBC interviewees had also suggested that they had gained the knowledge to answer the *ACIDI* prompts from the lecture or workshop portion of their course.

Rasch analysis provided a more nuanced interpretation of *ACIDI* scores, which had been previously absent in the chemistry education literature. Rasch analysis indicated that *ACIDI* seems to be more capable of distinguishing high from low performing students, while having difficulty distinguishing those in between. This may be due to the low number of questions asked by *ACIDI*. Furthermore, Rasch analysis indicated of significant conceptual learning gains with evidence of significant person ability gains.

A positive association was seen between the *Chemical Thinking* curriculum and reform-based general chemistry curriculum students' conceptions of induction and resonance. Although *Chemical Thinking* may need to improve its instruction of organic structure representations, as some had trouble identifying implicit hydrogens within acetylacetone. This difficulty of implicit hydrogen identification seemed to impair RBC students' ability to appropriately compare the acidity of acetylacetone to phenol.

RBC student interviews and workshop discussions revealed dominant, scientifically accepted, conceptions of induction and resonance stabilization for *p*-methylphenol and acetic acid. Analysis of RBC students' workshop arguments revealed increasing total argumentation quality seemed to be associated (R = 0.72) with increasing *ACIDI* pre-post person ability change, which is indicative of conceptual reprioritization. Evidence of conceptual reprioritization towards dominant, scientifically accepted, conceptions of induction and resonance was also evident in *ACIDI* item-level (items 1, 8 & 9) changes and persistence of that change.

ROXCI

The *ROXCI* assessment was determined to be unidimensional, and appropriately difficult for both reform-based (RBC) and traditional (TC) general chemistry curriculum students. *ROXCI* item prompts were appropriately interpreted by the majority of TC interviewees, while no RBC students' *ROXCI* item interpretations were collected. Providing support for the *ROXCI* instrument's substantive validity to TC students. TC interviewees had suggested that they had gained the knowledge to answer *ROXCI* items from various aspects of the traditional curriculum class (i.e. lecture, workshop, homework program) or previous chemistry knowledge gained from previous chemistry courses at the high school and college-level. This cautions the connection of conceptual learning gains to the traditional curriculum., as students had used outside/prior knowledge to answer *ROXCI* items.

Rasch analysis had revealed that person ability before and after instruction in the tested populations was significantly different, as the RBC students outperformed the TC students. Both RBC and TC students made about the same significant gains in person ability on the *ROXCI* assessment after course instruction. Results of a mixed effects multiple linear regression had suggested that time and class were significantly correlated with *ROXCI* person ability scores, suggesting that the class a student was in and instruction may associated with their change in person ability score. Disaggregation of person ability gains into gender and generation had revealed that RBC males and females, and first-generation and non-first-generation students, had made statistically similar person ability gains on *ROXCI*. Finally, Rasch analysis revealed that the *ROXCI* assessment was capable of distinguishing student ability of the respective courses along a logit scale of -3 to 4. However, some *ROXCI* items may be redundant, as several were of the same logit score difficulty.

Item-level gains suggested that RBC and TC made significant conceptual learning gains on the concepts of oxidation numbers, electron transfer and surface features of oxidation-reduction reactions. These gains were supported by TC student interviews, as there were dominant, scientifically accepted, redox conceptions presented. However, alternative conceptions were still present.

Workshop argumentation analysis of reform-based curriculum students' arguments had revealed that total argumentation quality of RBC students seemed to be positively associated with *ROXCI* post-test person ability scores (R = 0.52).

It is important to note the inability to include three TC workshop group's argumentation data, as three groups had not successfully completed both the *ROXCI* pre and post-test.

‘

Limitations

This study was limited to the student population in the reform-based and traditional general chemistry courses at the university where this study took place. Consenting student samples were not randomly selected, limiting their representativeness of their peers. Furthermore, aggregated quantitative analysis was limited to the number of consenting students who had completed the *ACIDI* pre, post and delayed post-test, and consenting students who had completed the *ROXCI* pre and post-test. Quantitative analysis that was disaggregated into demographics was limited to the number of students who had successfully completed the demographic questionnaire and the sample sizes of self-reported demographics. All quantitative indicators of conceptual reprioritization may be impacted by problem similarity, as both inventory item orders remained the same. Qualitative workshop argumentation analysis was limited to the number of students capable of being audio and video recorded during the workshop portion of their respective courses, student workshop attendance, and number of relevant workshops, while interview analysis was limited to the number of group members of interest who were able to attend their interviews. Furthermore, assessment of argumentation quality was limited to the criteria of argumentation levels in Erduran and co-workers’ analytical framework.

Future Direction

Although this study contributes to the progression of the chemistry education community’s understanding of undergraduate general chemistry students’ conceptions of organic acidity and oxidation-reduction, much work must continue to be done to understand potential conceptual reprioritization of organic acidity and oxidation-reduction conceptions. *Explanatory coexistence* may also provide an explanation for the persistence of scientifically inaccurate conceptions present within chemistry students, as students may hold onto such conceptions due to their usefulness in a chemistry course. However, investigation into what makes such conceptions useful has not been attempted, which necessitates such an investigation to better understand the persistence of alternative conceptions. Furthermore, potential uses of *ACIDI* and *ROXCI* as predictive tools and their relation to the creation of a community standard, of scientifically accepted conceptions of organic acidity and oxidation-reduction, were suggested. The creation of a database of undergraduate students’ conceptions of organic acidity and oxidation-reduction to more appropriately assess generalizability of results from different study populations was also recommended. This database may be created by collecting all student response data for *ACIDI* and *ROXCI*, across all institutions that have such data, in a running file that preserves participant confidentiality. Student responses can then be compared to each other to determine representation. Lastly, more research must be done to further assess the role of argumentation, and argumentation quality, in conceptual reprioritization of undergraduate general chemistry students’ conceptions of organic acidity and oxidation-reduction.

Thesis References

AERA, A. (2014). NCME.(2014). *Standards for educational and psychological testing*, 11-31.

Adadan, E., & Savasci, F. (2012). An analysis of 16–17-year-old students' understanding of solution chemistry concepts using a two-tier diagnostic instrument. *International Journal of Science Education*, *34*(4), 513-544.

Barbera, J. (2013). A psychometric analysis of the chemical concepts inventory. *Journal of Chemical Education*, *90*(5), 546-553.

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the web with KIE. *International Journal of Science Education*, *22*(8), 797-817.

Bhattacharyya, G. (2006). Practitioner development in organic chemistry: how graduate students conceptualize organic acids. *Chemistry Education Research and Practice*, *7*(4), 240-247.

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer Science & Business Media.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How?. *CBE—Life Sciences Education*, *15*(4), rm4.

Bowen, C. W. (2000). A quantitative literature review of cooperative learning effects on high school and college chemistry achievement. *Journal of Chemical education*, *77*(1), 116.

Bradley, J. D., & Mosimege, M. D. (1998). Misconceptions in acids and bases: A comparative study of student teachers with different chemistry backgrounds. *South African Journal of Chemistry*, *51*, 137-145.

Brandriet, A. R., & Bretz, S. L. (2014). The development of the redox concept inventory as a measure of students' symbolic and particulate redox understandings and confidence. *Journal of Chemical Education*, *91*(8), 1132-1144.

Bretz, S. L., & McClary, L. (2014). Students' Understandings of Acid Strength: How Meaningful Is Reliability When Measuring Alternative Conceptions?. *Journal of Chemical Education*, *92*(2), 212-219.

Bretz, S. L., & Linenberger, K. J. (2012). Development of the enzyme–substrate interactions concept inventory. *Biochemistry and Molecular Biology Education*, *40*(4), 229-233.

Bulte, A. M., Westbroek, H. B., de Jong, O., & Pilot, A. (2006). A research approach to designing chemistry education using authentic practices as contexts. *International Journal of Science Education*, *28*(9), 1063-1086.

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, *22*(1), 31-72.

Cartrette, D. P., & Mayo, P. M. (2011). Students' understanding of acids/bases in organic chemistry contexts. *Chemistry Education Research and Practice*, *12*(1), 29-39.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of educational research*, *64*(1), 1-35.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1-31.

Dawson, V., & Venville, G. J. (2009). High-school Students' Informal Reasoning and Argumentation about Biotechnology: An indicator of scientific literacy?. *International Journal of Science Education*, *31*(11), 1421-1445.

DeVellis, R. F. (2006). Classical test theory. *Medical care*, S50-S59.

De Jong, O., Acampo, J., & Verdonk, A. (1995). Problems in teaching the topic of redox reactions: actions and conceptions of chemistry teachers. *Journal of Research in Science Teaching*, *32*(10), 1097-1110.

Erduran, S., Simon, S., & Osborne, J. (2004). TAPping into argumentation: Developments in the application of Toulmin's argument pattern for studying science discourse. *Science education*, *88*(6), 915-933.

Foisy, L. M. B., Potvin, P., Riopel, M., & Masson, S. (2015). Is inhibition involved in overcoming a common physics misconception in mechanics?. *Trends in Neuroscience and Education*, *4*(1-2), 26-36.

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences*, *111*(23), 8410-8415.

Garnett, P. J., & Treagust, D. F. (1992). Conceptual difficulties experienced by senior high school students of electrochemistry: Electric circuits and oxidation-reduction equations. *Journal of Research in Science Teaching*, *29*(2), 121-142.

Heng, L. L., Surif, J., & Seng, C. H. (2014). Individual versus Group Argumentation: Student's Performance in a Malaysian Context. *International Education Studies*, *7*(7), 109-124.

Hurtado, S. (2001). Linking Diversity and Educational Purpose: How Diversity Affects the Classroom Environment and Student Development.

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1991). Active learning. *Cooperation in the College Classroom*, 1998.

Kousathana, M., Demerouti, M., & Tsaparlis, G. (2005). Instructional misconceptions in acid-base equilibria: An analysis from a history and philosophy of science perspective. *Science & Education*, *14*(2), 173-193.

Kuhn, T. S. (1963). *The function of dogma in scientific research* (pp. 347-369). na.

Kulatunga, U., Moog, R. S., & Lewis, J. E. (2014). Use of Toulmin's argumentation scheme for student discourse to gain insight about guided inquiry activities in college chemistry. *Journal of College Science Teaching*, *43*(5), 78-86.

Libarkin, J. (2008). Concept inventories in higher education science. In *BOSE Conf*.

Linacre, John Michael. “Reliability and Separation of Measures.” Table 23.99 Largest Residual Correlations for Items, Winsteps, 2018. www.winsteps.com/winman/reliability.htm.

Linder, C. J. (1993). A challenge to conceptual change. *Science Education*, *77*(3), 293-300.

Lindstrom, M. J., & Bates, D. M. (1988). Newton—Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014-1022.

Lombrozo, T., & Carey, S. (2006). Functional explanation and the function of explanation. *Cognition*, *99*(2), 167-204.

Lombrozo, T., Kelemen, D., & Zaitchik, D. (2007). Inferring design: Evidence of a preference for teleological explanations in patients with Alzheimer's disease. *Psychological Science*, *18*(11), 999-1006

Luxford, C. J., & Bretz, S. L. (2014). Development of the bonding representations inventory to identify student misconceptions about covalent and ionic bonding representations. *Journal of Chemical Education*, *91*(3), 312-320.

Marcus, M., & Minc, H. (1965). *Introduction to linear algebra*. Courier Corporation.

Mason, L. (1998). Sharing cognition to construct scientific knowledge in school context: The role of oral and written discourse. *Instructional science*, *26*(5), 359-389.

Masson, S., Potvin, P., Riopel, M., & Foisy, L. M. B. (2014). Differences in brain activation between novices and experts in science during a task involving a common misconception in electricity. *Mind, Brain, and Education*, *8*(1), 44-55.

McNeill, K. L., & Pimentel, D. S. (2010). Scientific discourse in three urban classrooms: The role of the teacher in engaging high school students in argumentation. *Science Education*, *94*(2), 203-229.

Micheli, R. (2012). Arguing without trying to persuade? Elements for a non-persuasive definition of argumentation. *Argumentation*, *26*(1), 115-126.

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of special education*, *43*(4), 236-254.

Nakhleh, M. B. (1994). Student's models of matter in the context of acid-base chemistry. *Journal of chemical education*, *71*(6), 495.

Nussbaum, E. M., Sinatra, G. M., & Poliquin, A. (2008). Role of epistemic beliefs and scientific argumentation in science learning. *International Journal of Science Education*, *30*(15), 1977-1999.

Nosek, B. A., Smyth, F. L., Sriram, N., Lindner, N. M., Devos, T., Ayala, A., & Kesebir, S. (2009). National differences in gender–science stereotypes predict national sex differences in science and math achievement. *Proceedings of the National Academy of Sciences*, *106*(26), 10593-10597.

Nyachwaya, J. M., Mohamed, A. R., Roehrig, G. H., Wood, N. B., Kern, A. L., & Schneider, J. L. (2011). The development of an open-ended drawing tool: an alternative diagnostic tool for assessing students' understanding of the particulate nature of matter. *Chemistry Education Research and Practice*, *12*(2), 121-132.

O'Connor, C., Michaels, S., Chapin, S., & Harbaugh, A. G. (2017). The silent and the vocal: participation and learning in whole-class discussion. *Learning and Instruction*, *48*, 5-13.

Özmen, H. (2008). Determination of students' alternative conceptions about chemical equilibrium: a review of research and the case of Turkey. *Chemistry Education Research and Practice*, *9*(3), 225-233.

Pentecost, T. C., & Barbera, J. (2013). Measuring learning gains in chemical education: a comparison of two methods. *Journal of Chemical Education*, *90*(7), 839-845.

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

Potvin, P. (2017). The Coexistence Claim and Its Possible Implications for Success in Teaching for Conceptual" Change". *European Journal of Science and Mathematics Education*, *5*(1), 55-66.

Powell, K. C., & Kalina, C. J. (2009). COGNITIVE AND SOCIAL CONSTRUCTIVISM: DEVELOPING TOOLS FOR AN i EFFECTIVE CLASSROOM. *Education*, *130*(2).

Prince, M. (2004). Does active learning work? A review of the research. *Journal of engineering education*, *93*(3), 223-231.

Rahayu, S., Chandrasegaran, A. L., Treagust, D. F., Kita, M., & Ibnu, S. (2011). Understanding acid–base concepts: Evaluating the efficacy of a senior high school student-centred instructional program in Indonesia. *International Journal of Science and Mathematics Education*, *9*(6), 1439-1458.

Rosenthal, D. P., & Sanger, M. J. (2012). Student misinterpretations and misconceptions based on their explanations of two computer animations of varying complexity depicting the same oxidation–reduction reaction. *Chemistry Education Research and Practice*, *13*(4), 471-483.

Sampson, V., & Clark, D. (2009). The impact of collaboration on the outcomes of scientific argumentation. *Science education*, *93*(3), 448-484.

Sevian, H., & Talanquer, V. (2014). Rethinking chemistry: A learning progression on chemical thinking. *Chemistry Education Research and Practice*, *15*(1), 10-23.

Shah, L., Rodriguez, C. A., Bartoli, M., & Rushton, G. T. (2018). Analysing the impact of a discussion-oriented curriculum on first-year general chemistry students' conceptions of relative acidity. *Chemistry Education Research and Practice*, *19*(2), 543-557.

Shtulman, A., & Lombrozo, T. (2016). Bundles of contradiction: A coexistence view of conceptual change. *Core knowledge and conceptual change*, 49-67.

Sia, D. T., Treagust, D. F., & Chandrasegaran, A. L. (2012). High School Students’ Proficiency and Confidence Levels in Displaying Their Understanding of Basic Electrolysis Concepts. *International Journal of Science and Mathematics Education*, *10*(6), 1325-1345.

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary educational psychology*, *21*(1), 43-69.

Stains, M., & Talanquer, V. (2008). Classification of chemical reactions: Stages of expertise. *Journal of Research in Science Teaching*, *45*(7), 771-793.

Stains, M., Escriu-Sune, M., Molina Alvarez de Santizo, M. L., & Sevian, H. (2011). Assessing secondary and college students’ implicit assumptions about the particulate nature of matter: Development and validation of the structure and motion of matter survey. *Journal of Chemical Education*, *88*(10), 1359-1365.

Steif, P. S., & Hansen, M. A. (2007). New practices for administering and analyzing the results of concept inventories. *Journal of Engineering Education*, *96*(3), 205-212.

Stephens, N. M., Hamedani, M. G., & Destin, M. (2014). Closing the social-class achievement gap: A difference-education intervention improves first-generation students’ academic performance and all students’ college transition. *Psychological science*, *25*(4), 943-953.

Strauss, M. E., & Smith, G. T. (2009). Construct validity: Advances in theory and methodology. *Annual review of clinical psychology*, *5*, 1-25.

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.

Talanquer, V. (2013). School chemistry: the need for transgression. *Science & Education*, *22*(7), 1757-1773.

Talanquer, V. (2018). Progressions in reasoning about structure–property relationships. *Chemistry Education Research and Practice*.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, *2*, 53.

Toulmin, S. (2003). The Uses of Argument. 1958. *Cambridge: Cambridge UP*.

Treagust, D. F., Chandrasegaran, A. L., Crowley, J., Yung, B. H., Cheong, I. P. A., & Othman, J. (2010). Evaluating Students' Understanding of Kinetic Particle Theory Concepts Relating to the States of Matter, Changes of State and Diffusion. *International Journal of Science and Mathematics Education*, *8*(1), 141-164.

Van Berkel, B., De Vos, W., Verdonk, A. H., & Pilot, A. (2000). Normal science education and its dangers: The case of school chemistry. *Science & Education*, *9*(1-2), 123-159.

Voska, K. W., & Heikkinen, H. W. (2000). Identification and analysis of student conceptions used to solve chemical equilibrium problems. *Journal of Research in Science Teaching: The Official Journal of the National Association for Research in Science Teaching*, *37*(2), 160-176.

Vygotsky, L. S. (1962). Language and thought. *Massachusetts Institute of Technology Press, Ontario, Canada*.

Wang, J., & Bao, L. (2010). Analyzing force concept inventory with item response theory. *American Journal of Physics*, *78*(10), 1064-1070.

Warburton, E. C., Bugarin, R., & Nunez, A. M. (2001). Bridging the Gap: Academic Preparation and Postsecondary Success of First-Generation Students. Statistical Analysis Report. Postsecondary Education Descriptive Analysis Reports.

Wei, S., Liu, X., Wang, Z., & Wang, X. (2012). Using Rasch measurement to develop a computer modeling-based instrument to assess students' conceptual understanding of matter. *Journal of Chemical Education*, *89*(3), 335-345.

Wold, S., Esbensen, K., & Geladi, P. (1987). Principal component analysis. *Chemometrics and intelligent laboratory systems*, *2*(1-3), 37-52.

Appendix

Interview Protocol

Please take a minute before each round of questioning to re-read the question and choices and identify any terms that you are unfamiliar with or are confusing.

1. Can you please re-read the question and tell me what you think the question is asking?
2. Can you please construct an argument for why the choice you think best answers the question is better than the other choices that are listed and why each of those choices are not as good as the 'best' choice?
3. Can you construct an argument against your idea for which answer you think is best?
4. Did you guess when answering this question?
5. Can you describe to me how you came to understand the question? Where did you gain the knowledge to best answer the question? (Lecture, workshop, office hours, study group, studying alone, private tutor, online search etc).
6. Can you think of any other things that can be added to the question that may provide clarification for anything that may be confusing? Things like pictures, definitions, etc.

Workshop Questions

1. Do you remember talking about these ideas in lecture or workshop?
2. How did you come to understand the content?
3. Can you describe your group dynamic in workshop? (positive, negative, how do you all get along?)
    a. How does that relate to how you and your groupmates construct arguments?

Rating Questions (Scale from 1-10, 1 being easiest and 10 being the most difficult). Ask why after they rate things.

1. Having conversations with group mates
2. Interpreting & Answering workshop prompts as a group
3. How difficult lecture conversations have been
4. How difficult contributing to lecture conversations have been
5. How difficult are the pre-class quizzes? Workshop Quizzes (Weekly HW)? Exams?

ACIDI Post-Test Item Outfit Table

Item	Outfit MNSQ
1	1.2283382
2	0.9543216
3	0.9804298
4	0.9699622
5	1.0025761
6	1.1073538
7	1.0323052
8	0.5660240
9	(Constraint)

***ROXCI* Item Outfit Table**

Item	Outfit MNSQ
2	1.4094242
3	0.7332997
4	0.8861356
5	0.8480966
6	0.8847421
7	0.8846000
8	1.0758999
9	1.0387355
10	0.7069144
11	1.1189802
12	0.9097548
13	1.1419177
14	0.9367491
15	0.9720052
16	0.9113872
17	1.1773374
18	1.1732047

Workshop Argumentation Quality Examples (using Argumentation Quality Framework, below, adapted from Erduran, Simon & Osborne, 2004)

Argument Quality (level)	**Criteria**
Level 1	Claim versus claim/counter-claim
Level 2	Claim versus claim with either data, warrants, or backings, but no rebuttals.
Level 3	Series of claims versus claims/counter-claims with either data, warrants, or backing with the occasional weak rebuttal.
Level 4	Claim with a clearly identifiable rebuttal. Argument may have several claims/counter-claims.
Level 5	Extended argument with more than one rebuttal.

ACIDI

Level 1: claim v counter-claim.

Student A: Isn't the other one sp3 as well?
TA: well this is what I was getting to with the resonance.
Student B: oh wait, this is sp2.

Level 2: Series of claims and a piece of data was included.

Student C: the nitrogen [inaudible].
Student B: but why why why why? That is the question.
Student A: because it's more electro…
Student B: positive.
Student A: no. It's more dense...I don't know what I'm saying.
Student B: if it's more dense, then it won't want to...it's not basic it's acidic.

Level 3: Series of claims with data and a weak rebuttal.

Student A: It's not that one.
Student B: This one?
Student A: Cuz it's right next to the double bonded O.
Student B: Yeah, but…
Student A: It'll be more acidic because it has more resonance. It can have more resonance.
Student C: I'll say this one.
Student B: This one?
Student C: Yeah.
Student A: Well it is next to… yeah.
Student C: It has resonance. Also this one.
Student A: That one is next to a CH3.
Student C: Yeah.

Level 4: Series of claims w/ clear rebuttal and data.

Student B: acids are accepting...wait bases are acceptors.
Student A: so this one would be the weakest right because essentially here it has two places to [inaudible]
Student B: but can you make the argument that this is more filled, so this would be filled like that. So I think it's between these two.
Student A: yeah that's what I'm saying.
Student B: and I think this one would be most basic.
Student A: oh you think it's [inaudible].
Student B: before we were thinking of electronegativity, now think less electronegative. This has the least electronegativity [the rest of his explanation is inaudible].

Level 5: Numerous rebuttals, claims and a basic argument is presented.

Student B: Yeah, that’s the alcohol. Unless we count the ones on the carbon, but then I don’t know if we count the ones in the carbon.
Student C: That’s what I was talking about, yeah.
Student A: They wouldn’t even participate because all the…
Student B: Yeah, it’s the least likely to get protonated though.
Student A: I know, but you can’t really say which specific one though.
Student B: You would say probably this one? Or actually you would say this one.
Student C: No, because it [inaudible].
Student A: I think what we’re overthinking it, I think it is just this at least.
Student B: No, it would be this one because the [inaudible] groups attract negative charge, so if you put it here. This would distribute the negative charge between all of this. Where this it would just clump it up over here.
Student C: I’m saying [inaudible] because then it’s further away from the OH and the carboxyl group.
Student B: But if one were to be protonated, it wouldn’t be this one.
Student C: Deprotonated, yeah.
Student A: I think we’re overthinking this by a long shot.

ROXCI Reform-Based Curriculum Students (RBC)

Level 1: None

Level 2: None

Level 3: Series of claims with data, warrants and a weak rebuttal.

> Student A: It's oxidized.
> Student B: Wait, this is carbon. So this is 0, wait why would this be 0? Cuz the electrons are being taken away from here. So that's, it actually goes to a -2, it goes to a -2.
> Student B: So this goes from a -1 to a -2. -1 to -2 means it's reduced. TA! Does that make sense?
> Student A: Why is this 0 here?
> TA: I think it's +2.
> Student B: I keep on…. It's +1 to +2. Cuz this takes away. +1 to +2, so it's oxidized, but not for the reason you said.
> Student C: I think it's [inaudible]. The carbon bonds on oxygen, it's gonna take away. To gain and this is oxidized.
> Student B: So it's +2.
> Student C: And here, it's bonded to a hydrogen. Carbon is more electronegative, so it's gonna gain, reduced.

Level 4: Series of claims with data, warrants and a clear rebuttal.

> Student C: We'll say the carbon.
> Student B: No, but the electron goes back to this, so, and then it just goes there. So I don't think
> anything happens to this carbon. Maybe…
> Student C: Maybe there's no… that's +, that's -2, that's +
> Student B: This is -2
> Student A: Wait, I'm confused.
> Student B: What's oxidized?
> Student C: Reduced, yeah.
> Student A: Gain of electrons.
> Student B: Probably the carbon, I mean the same thing happens here except it's oxidized. Which means it's gained something.
> Student C: Yeah.
> Student B: Might as well be the -O. It already gave a proton.
> Student C: Yeah.

Level 5: Extended argument with more than one rebuttal.

Student B: Do we think it’s oxidation or reduction that we’re getting? Probably oxidation, right? You’re adding protons.
Student C: Yeah, it would be oxidation because its CH… yeah.
Student C: -3? That doesn’t seem right.
Student A: Yeah, it’s not. It’s not that.
Student B: It’s not neutral. It would be -2.
Student B: I think, wait it has to be oxidized, right? I think… that’s not being oxidized, that’s being reduced.
Student C: That’s being reduced.

ROXCI Traditional Curriculum Students (TC)

Level 1: Claim v claim.

Student A: wouldn’t it be 3 moles? It’s 3 moles.
Student C: It’s not moles, it’s number of electrons.
Student A: Well it would be 3.
Student C: Yeah, that’s why I got 3.

Level 2: Claim v claim with data.

Student A: what side are we adding the waters to balance the oxygens? let's see, this side has more oxygens. So I'm going to guess we're going to add it to this side.
Student B: this side.
Student A: yeah. both of them.
Student B: yeah.

Level 3: Claim v claim with a weak rebuttal.

Student B: I'm not quite sure, I think it's this. Like PbO2 + S + 6 = uh, whatever.
Student A: what's [inaudible]. Well [inaudible] electron, is it just 1 Pb?
Student B: Uh, no. It's Pb +4 because it says lead (IV) oxide.
Student A: well yea, but it's like…
Student B: Yea.
Student A: Well no, isn't this separate here?
Student B: oh, yea it is.
Student A: just of Pb and then + Pb whatever...
Student B: yea, you're right.

Level 4: Claim v claim with clear rebuttal.

TA: Yes, do you remember what I said about... I don't know if I told you guys, so where does the electrons go? what is the indication of where your electrons should go?
Student B: Wherever the positive side is.
TA: Wherever the protons are, so you add it to the side that has the protons.
Student A: so you take this and add it over there.
Student B: that makes no sense to me.
Student A: does that make it better?
TA: So if you...
Student B: Why though? Cuz that's already negative.
TA: so if you think about it, so how did you determine how many electrons.
Student A: that makes sense because this is 2 - and this is 3 - so it's going to cancel out the 5 +.

Level 5: Extended argument with multiple rebuttals

Student A: OHs to that side.

Student B: Where did this one come from? oh it's coming from this one, ok. So what they did, which makes no sense, though because they... they made the whatever it's called. They made the water into an OH which cancels that. We don't have water on this side. Alright, let's see we got a 4, 5, 6 hydrogen's on the side. We have 5 hydrogens on the side.

Student A: Wait, I Got 5 hydrogens on this side, how did you get 6?

Student B: I have four H2, right? cuz 2 and nevermind I don't have 5H. You have 5H?

Student A: No. I have 2 and, woah! how'd you get 7 on this side? Let's see.

Student B: because I have the.

Student A: No look, see, it's balanced, 5 and 5. right?

Student B: What do you have? H2 2 and then I have the 3 so that's 5.

Student A: so how'd you get 7?

Student B: I have no idea

Student A: so 5 and 5?

Student B: yes. where the hell did that 7 come from?

Student A: I don't know. but now this is 3.

Student B: it came from somewhere, hold on. 2H2O and then we did that okay 5 so that's three H.

Student A: So 5 and 5 on this side. So we're good on Hs with that one, 5 Os to 5 Os. So, we're good on hydrogens and oxygens.

Student B: are we? 1, 2. 1, 4, 5. I have five oxygen on this side?

Student A: yep.

Student B: and then 2, yeah you're right. We have 5.

Student A: 5. Okay, we're good. everything is balanced, great.