Advancing Equity Amongst General Chemistry Students with Variable Preparations in Mathematics

Vanessa R. Ralph
University of South Florida

Follow this and additional works at: https://scholarcommons.usf.edu/etd

Part of the Chemistry Commons, and the Education Commons

Scholar Commons Citation
Ralph, Vanessa R., "Advancing Equity Amongst General Chemistry Students with Variable Preparations in Mathematics" (2020). Graduate Theses and Dissertations.
https://scholarcommons.usf.edu/etd/8284

This Dissertation is brought to you for free and open access by the Graduate School at Scholar Commons. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Scholar Commons. For more information, please contact scholarcommons@usf.edu.
Advancing Equity Amongst General Chemistry Students with Variable Preparations in Mathematics

by

Vanessa R. Ralph

A dissertation submitted in the partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of Chemistry
College of Arts and Sciences
University of South Florida

Major Professor: Scott E. Lewis, Ph.D.
Jeffrey R. Raker, Ph.D.
James W. Leahy, Ph.D.
Leia K. Cain, Ph.D.

Date of Approval: February 21st, 2020

Keywords: science education reform, program evaluation, assessments

Copyright © 2020, Vanessa R. Ralph
DEDICATION

I dedicate this dissertation to the family, teachers, and students who inspired my enrollment, progress, and completion of this doctoral degree. Olivia provided me boundless motivation to ensure our futures are not limited in opportunity and to demonstrate the role Hispanic women can play in the science workforce. The countless evenings we spent coloring, singing, and playing soccer served to reorient me to what mattered most throughout my experiences in graduate school and provided respites from research. Adam has been selfless in supporting my career ambitions and so willingly took on my share of the meals, playdates, housekeeping, and care. Thank you for preventing my minimization of the growth I experienced throughout this process. Steve, Antonia, and Steven, you all contributed to my resilience and humor. These tools were instrumental in achieving this milestone, as they have been throughout the whole of my life.

Throughout secondary school, several teachers helped me to combat the fallacy placed by other societal influences that I would be unable to participate in the sciences, given my ethnicity and gender. Mr. Michael Otero was the voice I still lean on when I doubt myself and my ability to excel in the world of STEM. This work is dedicated to all teachers who provide a light to students of marginalized groups to pursue all they are capable of without concern of what others say deem possible and hope to adopt this role as I progress in my career. Finally, I dedicate this work to my students at Cardinal Mooney Catholic High School and the State College of Florida. Thank you for inspiring so much of what I value in my career and for enduring my early lessons as a teacher.
ACKNOWLEDGEMENTS

My advisor, Dr. Scott Lewis, has generously served as a dedicated resource for guidance and constructive feedback while demonstrating leadership grounded in humor and understanding. I recognize the gift of his leadership and hope to pay tribute to his efforts through the students I will advise going forward. We, at the University of South Florida, are incredibly fortunate to have many phenomenal mentors. Drs. Jeffrey Raker and Jennifer Lewis dedicated their time and feedback and have had a strong influence on the research, perspectives, and methods I have sought to apply in these works. Drs. James Leahy and Leia Cain provided much-appreciated guidance throughout the preparation of these works as members of my committee. Dr. Bernard Batson was instrumental toward earning support by the Florida-Georgia Louis Stokes Alliance for Minority Participation Bridge to the Doctorate award. Dr. John Ferron provided instruction in statistics and single-subjects research designs that informed the works presented. Dr. Jenay Robert, the Center for Excellence in Science Education at PennState, contributed efforts in training me in the use of qualitative data analysis software and has continued to serve as an excellent source of professional support and guidance.

Mentorship was also offered by a talented cohort of graduate students who supported the advancement of these works, including James Kingsepp, M.D. Tawabur Rahman, Ying Wang, Ayesha Farheen, Stephanie Feola, Guizella Rocabado, Jacob McAlpin, Rebecca Gibbons, Amber Dood, Aaron Clark, and Brandon Yik. Undergraduate researchers contributed considerably to the advancement of these works, including Yvonne Nguyen, Francesca Gallucci, Gustavo A. Velasquez, Rebeca Gonzalez Jauregui, and Cassandra Ly. Appreciation is also extended to the faculty, staff, and students at the University of South Florida, without whom this research would not have been possible.
# TABLE OF CONTENTS

List of Tables viii

List of Figures vi

List of Boxes xii

List of Equations xiii

Abstract xiv

Chapter 1: Understanding the Relationship Between Math Test Score and Chemistry Performance 1
  Introduction 1
  Precollege Mathematics Preparation Predicts Academic Chemistry Performance 2
  Implications for Supporting At-Risk Chemistry Students 3
    Remedial Instruction in Mathematics 4
    Remedial Instruction in Chemistry 6
    Interventions of Supplemental Instruction 6
    Multi-Faceted Interventions for At-Risk Chemistry Students 7
    Incorporating Collaborative Learning Environments 8
  Operationalizing Equity in the Context of Science Education Reform 10
    Past Works and Lynch’s Framework for Evaluating Science Education Reforms 12
  Directions for this Work 13
  References 13

Chapter 2: Methodologies Used to Evaluate Academic Performance 18
  Note to Reader 18
  Research Setting 18
    Class Sizes 19
    Content and Prerequisites 19
    Coordination and Grading Schemes 19
    Pedagogy 20
    Retention and Pass Rates 20
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Instruments Assessing Chemistry Knowledge</td>
<td>21</td>
</tr>
<tr>
<td>Ethics</td>
<td>22</td>
</tr>
<tr>
<td>Identifying Disparities Amongst General Chemistry Students</td>
<td>23</td>
</tr>
<tr>
<td>Concerns Regarding Equity and Diversity</td>
<td>24</td>
</tr>
<tr>
<td>Examining the Extent of Inequity Observed at the Setting</td>
<td>25</td>
</tr>
<tr>
<td>References</td>
<td>28</td>
</tr>
<tr>
<td>Chapter 3: An Evaluation of Peer-Led Team Learning and Equity</td>
<td>30</td>
</tr>
<tr>
<td>Note to Reader</td>
<td>30</td>
</tr>
<tr>
<td>The Evidence in Evidence-Based Practices</td>
<td>30</td>
</tr>
<tr>
<td>A Shift in the “Gold Standard”</td>
<td>31</td>
</tr>
<tr>
<td>Challenges to Enacting Randomized Control Trials in Educational Settings</td>
<td>31</td>
</tr>
<tr>
<td>Objectives of this Study</td>
<td>32</td>
</tr>
<tr>
<td>Introducing Randomization Tests</td>
<td>33</td>
</tr>
<tr>
<td>A Shift in Randomization</td>
<td>33</td>
</tr>
<tr>
<td>How Randomization Tests Work</td>
<td>37</td>
</tr>
<tr>
<td>Methods</td>
<td>39</td>
</tr>
<tr>
<td>Instructors and Survey Items</td>
<td>39</td>
</tr>
<tr>
<td>Effectiveness and Equity as Outcome Measures</td>
<td>41</td>
</tr>
<tr>
<td>Problematizing Effectiveness: Mean Standardized Residual (MSR)</td>
<td>41</td>
</tr>
<tr>
<td>Problematizing Equity: Differential Standardized Residual (DSR)</td>
<td>43</td>
</tr>
<tr>
<td>Interpreting Effect Sizes</td>
<td>45</td>
</tr>
<tr>
<td>Non-Overlap Indices</td>
<td>46</td>
</tr>
<tr>
<td>Results</td>
<td>48</td>
</tr>
<tr>
<td>Instructor-Level Differences in the Impact of PLTL on Overall Student Performance</td>
<td>48</td>
</tr>
<tr>
<td>Professor Parker (BA)</td>
<td>48</td>
</tr>
<tr>
<td>Professor Davis (AB)</td>
<td>50</td>
</tr>
<tr>
<td>Professor Morales (AB and ABA)</td>
<td>52</td>
</tr>
<tr>
<td>Professor Stacy (ABAB or BABA)</td>
<td>55</td>
</tr>
<tr>
<td>Summary of PLTL Impact on Student Outcomes</td>
<td>56</td>
</tr>
<tr>
<td>Instructor-Level Differences also Moderate the Impact of PLTL on Equity</td>
<td>57</td>
</tr>
<tr>
<td>Equity of Individuality</td>
<td>57</td>
</tr>
<tr>
<td>Equity of Parity</td>
<td>58</td>
</tr>
<tr>
<td>The Impact of PLTL on Equity</td>
<td>60</td>
</tr>
<tr>
<td>Discussion</td>
<td>60</td>
</tr>
<tr>
<td>An Evaluation of Peer-Led Team Learning in the Context of Prior Literature</td>
<td>61</td>
</tr>
<tr>
<td>Methods Addressing Instructor-Level Differences</td>
<td>63</td>
</tr>
<tr>
<td>Limitations</td>
<td>67</td>
</tr>
<tr>
<td>Chapter</td>
<td>Title</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
</tr>
<tr>
<td>4</td>
<td>Chemistry Topics Posing Incommensurate Difficulty to Students with Low Math Aptitude Scores</td>
</tr>
<tr>
<td></td>
<td>Note to Reader</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
</tr>
<tr>
<td></td>
<td>Identifying At-Risk Students in Chemistry</td>
</tr>
<tr>
<td></td>
<td>Assisting At-Risk Students in Chemistry</td>
</tr>
<tr>
<td></td>
<td>Purpose of the Study</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
</tr>
<tr>
<td></td>
<td>Treatment of Missing Data</td>
</tr>
<tr>
<td></td>
<td>Assessment Items Scored and Mapped onto Major Topics Results</td>
</tr>
<tr>
<td></td>
<td>Results</td>
</tr>
<tr>
<td></td>
<td>At-Risk Students are Disproportionately Challenged on Mole Concept and Stoichiometry Assessment items</td>
</tr>
<tr>
<td></td>
<td>Mole Concept and Stoichiometry are Essential to Student Success</td>
</tr>
<tr>
<td></td>
<td>At-Risk Students Achieving Proficiency in Mole Concept and Stoichiometry Performed Equitably or Surpassed their Not At-Risk Peers</td>
</tr>
<tr>
<td></td>
<td>Explorations of Missing Data</td>
</tr>
<tr>
<td></td>
<td>Discussion and Implications</td>
</tr>
<tr>
<td></td>
<td>Conclusion</td>
</tr>
<tr>
<td></td>
<td>References</td>
</tr>
<tr>
<td>5</td>
<td>An Explanative Basis for the Differential Performance of Students with Low Math Aptitude</td>
</tr>
<tr>
<td></td>
<td>Note to Reader</td>
</tr>
<tr>
<td></td>
<td>Introduction</td>
</tr>
<tr>
<td></td>
<td>Prediction for the Sake of Intervention</td>
</tr>
<tr>
<td></td>
<td>Challenges Chemistry Students Face with the Mole Concept and Stoichiometry</td>
</tr>
<tr>
<td></td>
<td>Purpose</td>
</tr>
<tr>
<td></td>
<td>Research Objectives</td>
</tr>
<tr>
<td></td>
<td>Methods</td>
</tr>
<tr>
<td></td>
<td>Development and Design of Post-Assessment Survey Prompts</td>
</tr>
<tr>
<td></td>
<td>Coding Students’ Responses</td>
</tr>
<tr>
<td></td>
<td>Results</td>
</tr>
<tr>
<td></td>
<td>Challenges of At-Risk Chemistry Students</td>
</tr>
<tr>
<td></td>
<td>Interchangeable Chemical Identity</td>
</tr>
</tbody>
</table>
Chapter 6: The Impact of Representations in Assessments on Student Performance and Equity

Note to Reader

Introduction

Differentiating Representations (Symbolic vs. Particulate) from Task (Conceptual vs. Algorithmic)

Investigating the Role of Representations in Chemistry Assessments

Advancing Research Concerning Equitable Assessment Design Practices

Purpose

Methods

Research Setting

Data Collection

Substituting Representations

Supplementing Representations

Identifying At-Risk Students

Measures and Evaluation

Results

Substituting Symbolic with Particulate Representation had Negative Impacts on Student Performance and Equity

Supplementing Items with Tables had a Positive Impact on Student Performance and Equity, whereas, Supplementing with Particulate Representations Presented Variable Outcomes

Representations in Assessments had a Considerable Impact on the Distractors Students Selected

Substituted Representations

Atomic vs. Molecular

The Role of Molar Mass

“In Excess” as a Source of Confusion

A Comparison of the Challenges Observed between Not and At-Risk Chemistry Students

The Mole Concept Item

The Stoichiometry Item

The Gas Laws Item

The Changes in Energy Item

Discussion

Implications for More Equitable Assessment Design

Other Implications for Improved Assessment Design

Pedagogical Implications for the Instruction of Stoichiometry

Limitations and Future Works

Conclusion

References
Supplemented Representations 173
Discussion 175
The Representations used in an Assessment Task have a Significant Impact on the Distractors to which Students Attend 175
Supplementing Assessment Tasks with Tables can Improve At-Risk Student Performance 176
Modeling and Assessing Translations between Representations of Chemical Phenomena 177
Limitations and Future Works 178
Conclusion 179
References 180

Chapter 7: Conclusions and Future Directions 184
Relevance of the Work Presented to Chemistry Education Research 184
Addressing the “Describe” Gap in the Literature 184
Implications for Research and Practice 185
Randomization Tests (Chapter 3) 185
Incommensurate Topics (Chapter 4) 186
An Explanative Basis (Chapter 5) 187
Representations in Assessment (Chapter 6) 188
A Dissertation, Synthesized 189

Impacts Beyond Chemistry Education Research 190
Diversity and Inclusion in the STEM Workforce 190
Prerequisites and Remedial Coursework are Insufficient Strategies, Calls for Interdisciplinary Cooperation 190
Advancing Understandings as to How Tables Impact Students’ Conceptualization of Algorithms Applied throughout the Physical Sciences 191

References 192

Appendix A: Institutional Review 193
Appendix B: Copyright Permissions 198
Appendix C: Topics Coding Scheme for First-Semester General Chemistry (GCI) 200
Appendix D: Effect Sizes of Topic Scores by Semester 204
Appendix E: Results of Additional Missing Data Analyses 206

About the Author End Page
LIST OF TABLES

Table 1: Sample sizes, medians, standard deviations (SD), and quartiles for math composite scores. 23
Table 2: Race or ethnicity disaggregated by risk for students enrolled in GCI at the research setting. 24
Table 3: The percentage of students from each group (not-at-risk or NAR and at-risk or AR) who achieved each final course outcome in first- and second-semester general chemistry courses (GCI and GCII, respectively). 27
Table 4: The number of questions (Q’s) posited by an instructor to students, observed over 20 lessons with at least 4 observations per phase (possible starting points highlighted). 37
Table 5: Absolute values of differences in the mean number of questions asked in each lesson (actual starting point in blue). 38
Table 6: Students’ GCI Final exam scores regressed onto each of their GCII exams. 42
Table 7: Interpreting NAP effects. 47
Table 8: A summary of overall student performance as an outcome measure used to evaluate the reformed (PLTL) instructional pedagogy and its impact relative to the reference (didactic) approach. 57
Table 9: A summary of differential student performance (DSR) as an outcome measure used to evaluate the reformed (PLTL) instructional pedagogy and its impact relative to the reference (didactic) approach. 59
Table 10: A summary of the methods used in studies evaluating the impact of PLTL in science courses. 62
Table 11: Topics in order of teaching sequence followed by two-letter codes and frequencies. 81
Table 12: Pearson’s correlation values for topic scores, the SATM, and assessments 85
Table 13: A priori coding scheme comprised of the challenges by which students have been observed to struggle with the presented mole concept and stoichiometry assessment items in past works. 108
Table 14: Codes (organized by theme) with prevalence measured as the percentage each code comprises of the total number of inaccuracies identified in students' solution processes ($n_{NAR} = 393$, $n_{AR} = 234$).

Table 15: Code prevalence for students who arrived at accurate numerical values for the assessment items implementing an inaccurate process.

Table 16: Example tabular representation of worked problem.

Table 17: Each student was randomly assigned to a version of a stoichiometry item with each representation.

Table 18: Of students who completed all four exams and had a math aptitude score ($n = 1,469$), the differential performance observed between those at-risk (Scoring in the bottom quartile of math aptitude scores), and not-at-risk is evident.

Table 19: Chi-square tests comparing the number of students who selected the correct answer for assessment tasks substituting representation.

Table 20: Chi-square tests comparing the number of students who selected the correct answer for assessments supplementing symbolic representation (S) with either a particulate diagram (P) or table (T).

Table 21: Chi-square tests comparing the number of students who selected each of 5 distractors for items of substituted representation.

Table 22: Chi-square tests comparing the number of students who selected each of 5 distractors for items of supplemented representation.

Table A1: Frequencies of students within each subgroup of the overall cohort.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A visual summary of the cycles by which an equality of outcomes can be evaluated.</td>
</tr>
<tr>
<td>2</td>
<td>Citations related to each stage of the equity scheme demonstrating a progression from investigate to adapt and evaluate with limited research describing the nature of the disparities observed.</td>
</tr>
<tr>
<td>3</td>
<td>Sankey diagram where the number of not-at-risk and at-risk students (left) are represented by the width of the bands connecting students to the final grades achieved (sorted in descending likelihood) in first- and second-semester general chemistry courses (GCI and GCII, respectively; middle and right).</td>
</tr>
<tr>
<td>4</td>
<td>Illustration of the differences between two methodologies that enact randomization in the assignment of students (randomized control trial) versus the assignment of pedagogy over a period of time (randomization tests).</td>
</tr>
<tr>
<td>5</td>
<td>Graphical representation of hypothetical data used in randomization tests design for students of one or more instructors.</td>
</tr>
<tr>
<td>6</td>
<td>Example case for calculating nonoverlap of all pairs (NAP) – numbers over data points in phase A indicate overlap with phase B.</td>
</tr>
<tr>
<td>7</td>
<td>Graphic display of Professor Parker’s performance feedback from semester 1 to 7.</td>
</tr>
<tr>
<td>8</td>
<td>Graphic display of Professor Davis’ performance feedback from semester 3 to 7.</td>
</tr>
<tr>
<td>9</td>
<td>Graphic display of Professor Morales’s performance feedback from semester 2 to 7 for an AB (upper section) and ABA (lower section) design.</td>
</tr>
<tr>
<td>10</td>
<td>Graphic display of Professor Stacy’s performance feedback from semester 1 to 7.</td>
</tr>
<tr>
<td>11</td>
<td>Bar graphs depicting the average MSRs (mean-centered by each case or instructor) for students of the not-at-risk (NAR) and at-risk (AR) group receiving didactic-instruction and peer-led team learning (PLTL).</td>
</tr>
<tr>
<td>12</td>
<td>An illustration representative of the various methodological approaches toward addressing instructor-level differences.</td>
</tr>
</tbody>
</table>
Figure 13: A wide range of mean topics scores were observed on interim and ACS exams overall.

Figure 14: Mole concept (MC) and stoichiometry (ST) are topics posing incommensurate difficulty to students with low math aptitude scores.

Figure 15: Interim assessment item on mole concept in gas laws.

Figure 16: Interim assessment item on stoichiometry in molarity.

Figure 17: Interim assessment item stoichiometry in gas laws.

Figure 18: Interim assessment item on stoichiometry in changes in enthalpy.

Figure 19: At-risk students proficient in mole concept and stoichiometry (Risers) perform similarity to the not at-risk cohort.

Figure 20: Topic performance of at-risk students proficient in mole concept and stoichiometry (Risers).

Figure 21: Trends related to disproportionate topic difficulty for at-risk students are consistent with students of lower ACTM scores.

Figure 22: The mole concept interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

Figure 23: Varieties by which students communicated interchangeability between numbers sharing the unit of mass but differing in chemical identity.

Figure 24: Misapplying mass percentages to the moles or molar masses of oxygen.

Figure 25: The stoichiometry interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

Figure 26: Returning to the mole concept assessment item, the code “atomic vs. molecular” was prevalent.

Figure 27: Students were applying coefficients representing stoichiometric proportions from balanced chemical reactions to the molar masses of chemicals.

Figure 28: The gas laws interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

Figure 29: Misapplications of molar mass in response to the gas law assessment item.

Figure 30: The changes in energy interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.
Figure 31: Misapplications of molar mass in response to the changes in energy assessment item.

Figure 32: Interchangeable chemical identities manifested via uncharacteristic solution processes involving excess reactant to the gas laws (first two responses) and changes in energy (last three responses) assessment items.

Figure 33: Incorporating a limiting reactants algorithm for stoichiometry (Clark and Robert), the gas laws (Meredith and Sam), and changes in energy (Katy and Oscar) assessment items describing all other reactants as in excess.

Figure 34: Common and chemically implausible solution strategies along with representational competence challenges observed amongst the not-at-risk chemistry students.

Figure 35: Inaccurate conceptions of algorithms communicated by not-at-risk chemistry students that result in accurate responses to the stoichiometry assessment item.

Figure 36: Inaccurate conceptions of algorithms communicated by not-at-risk chemistry students that result in accurate responses to the gas laws assessment item.

Figure 37: Inaccurate conceptions of algorithms communicated by not-at-risk chemistry students that result in accurate responses to the changes in energy assessment item.

Figure 38: Potential adaptation of distractors for a product-oriented assessment item to that of a process-oriented assessment item.

Figure 39: Suggestions for adapting assessment items in Figures 25, 28, and 30 toward items to avoid eliciting a limiting reactants algorithm from students as a result of reading the term “excess”.

Figure 40: Examples of representations for use in the assessment of stoichiometry.

Figure 41: Examples of chemically implausible solution processes presented by not-at-risk students in response to a task involving stoichiometry.

Figure 42: Three versions of a stoichiometry item involving percent yield (correct answer in bold).

Figure 43: Three versions of a stoichiometry item involving the molarity of a solution (correct answer in bold).

Figure 44: Three versions of a stoichiometry item in the context of gaseous solutions (correct answer in bold).
Figure 45: A histogram of the distribution of math test scores for students enrolled in first semester, general chemistry course at the research setting during data collection.

Figure 46: An example of paired representations for modeling and assessing translations.

Figure 47: Citations related to each stage of Lynch’s equity schema demonstrating a progression from investigate to adapt and evaluate with limited research describing the nature of the disparities observed and the papers published of these works contributing to the literature base.

Figure A1: Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Spring 2016.

Figure A2: Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Fall 2016.

Figure A3: Figure A3. Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Spring 2017.

Figure A4: Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Fall 2017.

Figure A5: Mean effect sizes between students with complete data and students missing tests.
# LIST OF BOXES

| Box 1: | Survey items administered to second-semester general chemistry instructors. | 40 |
| Box 2: | Survey prompts for each stoichiometric task. | 107 |
| Box 3: | Substituted representations for the dissolution of ionic compounds. | 161 |
| Box 4: | Substituted representations in the calculation of a theoretical yield. | 162 |
### LIST OF EQUATIONS

<table>
<thead>
<tr>
<th>Equation</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equation 1</td>
<td>Permutations formula for randomization tests.</td>
<td>39</td>
</tr>
<tr>
<td>Equation 2</td>
<td>Cohen’s $d$ formula where $MD$ is the difference in mean outcome measures in the B (reformed) and A (reference) phases and $SD$ is the pooled standard deviation across phases</td>
<td>45</td>
</tr>
<tr>
<td>Equation 3</td>
<td>Nonoverlap of all pairs formula where the number of pairs ($N$) and number of overlapping pairs ($n_o$) are used to calculate the index.</td>
<td>46</td>
</tr>
<tr>
<td>Equation 4</td>
<td>Cohen’s $w$ formula where $N$ is equal to the total sample size or the number of students who answered the items for which proportions are being compared and $\chi^2$ is equal to the chi-square statistic.</td>
<td>169</td>
</tr>
</tbody>
</table>
ABSTRACT

Reducing the inequities propagating the lack of representation in Science, Technology, Engineering, and Mathematics (STEM) careers has become a national imperative made increasingly complex given the inequitable preparations in science and mathematics students experience before college. A common institutional-level approach to promoting equity is the use of math and science prerequisites to prepare students who seek to enroll in lower-level STEM coursework. However, the literature does not support the efficacy of this approach, finding no substantive improvement to student performance despite considerable financial and time costs to students, instructors, and universities alike. Set in first-semester General Chemistry courses taken early in STEM degree programs, these works sought to identify and address the systems in place that contribute to inequitable STEM education and provide a cost-effective and research-based alternative to the prerequisite approach.

Students were identified as at-risk by having scored in the bottom-quartile of math composite SAT or ACT scores (or 25% of the overall cohort). Often representative of students scoring near or below national averages on the SAT and ACT, at-risk students achieved 43.8% of failing grades (D, F, W) in chemistry courses and comprised only 9.9% of the A’s awarded. Of these students at-risk, a disproportionate number self-identify as Black/African American or Hispanic/Latino suggesting that preparation in secondary-school mathematics, using SAT and ACT as a proxy, is not equitably achieved (or measured) by students of underrepresented minorities in the sciences.

First, the impact of a collaborative pedagogy (peer-lead team learning) on student performance and equity was examined. Instructor-level differences were observed wherein student performance improved in 3 of the 4 instructors’ classrooms when enacting peer-led
team learning. Significant reductions in the equities observed between at-risk chemistry students and not-at-risk chemistry students were observed in classrooms of 2 of the 4 instructors.

Then, student responses to 32 assessments spanning eight semesters were analyzed in first-semester General Chemistry classes at a large, doctorate-granting institution in the Southeastern United States. Students’ performance on assessment items coded by topic presented a consistent pattern of differential performance on items designed to measure proficiencies with the mole concept and stoichiometry, a topic fundamental to student success in General Chemistry I and II in addition to Analytical Chemistry.

In an analysis of students’ solution processes to assessment items of these topics, differential performance could not be attributed to differences in conceptual understanding nor mathematical prowess, but rather the frequency in which students with higher math composite scores arrived at a correct answer using a chemically flawed or inaccurate solution process. Based on these results, we concluded that assessments of these topics may reward rote memorization without the conceptualization of procedural approaches to chemistry while simultaneously exacerbating the inequity observed between students of differing preparations in mathematics.

Finally, to evaluate our recommendation of assessments that measure students’ understanding of the topic beyond application of procedural knowledge, heuristics, and algorithms, items were designed to encourage students’ attention to and organization of their solution processes. Student performance (overall) improved by an average of 8.5% while reducing the differential performance observed of students at-risk by as much as 12.4% (more than one letter grade).

These works contribute a variety of perspectives to science education, not the least of which is the importance of understanding the source of inequity before enacting intervention for a group of students to prevent unintended negative consequences. Implications include
practical examples of assessment items reoriented from promoting a series of decontextualized algorithmic procedures toward those promoting the conceptualization of foundational science concepts. Future works seek to investigate the impact of these changes in assessment on student performance, retention, and equity when modeled in instruction and assessed throughout the semester.
CHAPTER 1: UNDERSTANDING THE RELATIONSHIP BETWEEN MATH TEST SCORE AND CHEMISTRY PERFORMANCE

Introduction

The promotion of safer, healthier, more-educated communities with strong economic growth is strongly correlated to an investment in science, technology, engineering, and mathematics (STEM) education.¹ Stakeholders are challenged to maximize this investment by producing a knowledgeable workforce large enough to support the growing demand for scientists and engineers.¹² Opportunities to improve efficiency exist in ensuring that students receive equitable primary, secondary, and post-secondary preparation in science and mathematics, thereby promoting the growth and diversification of students in the sciences and engineering. One such opportunity is to prevent the loss of students in post-secondary introductory chemistry, which is described as a gateway course required for many degree programs within science and engineering.³ In chemistry courses, students enter with various degrees of preparation in areas which have been related to success (e.g., mathematics).⁴–¹⁸ College-level chemistry students with lower scores on pre-college math test scores have been labeled as academically at-risk.⁷,⁸,¹⁹,²⁰ The early identification and provision of aid to at-risk students offers an opportunity to minimize student attrition and increase participation in STEM.

The number of articles observed in the literature relating to the identification of at-risk students in chemistry is significant. However, the literature lacks detail as to the nature of these challenges with the course. For example, differences in at-risk student performance on topics (e.g., reactions in solution, stoichiometry) and subject-related skillsets (e.g., predicting the dissociation of ionic compounds, enacting mol-mol conversions) could inform how best to
intervene. These details are necessary to better understand the challenges at-risk students face and improve efforts directed toward their success. As stated by Suter and Camilli in 2019 (p. 59),

"STEM workforce capacity is a very disparate collection of skills, education, and disciplinary knowledge, policy recommendations would benefit from a clear specification of the exact skills that are needed to improve specific capacities."

Therefore, this dissertation applies a multi-methodological approach toward understanding the content-related challenges with which chemistry students identified as at-risk, most incommensurately struggle. The results of these studies are presented with the intent that doing so will provide the information necessary to move beyond sourcing new ways to predict student failure and, instead, apply these understandings toward the effective design, delivery, and assessment of prerequisite coursework, chemistry courses, and interventions. By providing lessons tailored toward meaningful student comprehension on the aspects of chemistry which impact at-risk students the most, the potential to deliver equitable outcomes to introductory chemistry courses may be realized.

Precollege Mathematics Preparation Predicts Academic Chemistry Performance

While a gateway course to many postsecondary STEM degree programs, general chemistry has also been identified as a considerable obstacle for students’ advancement and retention in STEM education. As such, investigations of the disparities observed in students’ general chemistry performance comprise a robust literature base predominately concerned with identifying relevant predictors of academic achievement.

A variety of predictors have been investigated to identify chemistry students at-risk of unfavorable outcomes in general or introductory, chemistry courses including demographic predictors such as students’ gender, race or ethnicity, and age. Affect-related qualities have also been identified as potential predictors of students’ performance in chemistry. Other predictors included study habits, prior knowledge of chemistry, including procedural
understandings of chemistry, and measures of cognitive ability such as self-evaluation, proportional reasoning, formal thought, or scientific reasoning ability.

Of the various predictors identified, standardized measures of precollege preparations in mathematics, such as the math composite of the SAT (SATM) and, to a lesser extent, the American College Test (ACTM), are robustly referenced as predictors of chemistry students’ academic performance. As such, differences in the mathematics preparation of chemistry students are the focus of this dissertation.

The SATM features three primary foci described as the mastery of linear equations and systems, quantitative literacy via problem-solving and data analysis, and the manipulation of complex questions. Students’ scores on SATM were correlated with introductory chemistry course grades (transformed into a numerical scale) with values ranging from correlations of 0.42 to 0.63. Repeated using ACTM scores, these correlations ranged from 0.36-0.76. One study applying historical data to latter cohorts to predict which students will pass or fail chemistry determined the predictive accuracy of the SATM as 72.8%, another reports a predictive accuracy spanning from 65-75%.

Overall, students with lower scores on pre-college math tests (e.g., SAT and ACT) have been regularly observed to disproportionately withdraw or achieve D’s and F’s in undergraduate chemistry courses when compared to their peers with higher math test scores. As such, general chemistry students scoring in the bottom 25-30% of math test scores, or with SATM scores below 600, are most at-risk for adverse assessment performance.

Taken in sum, these studies concern predictors of student performance and are critical as they inform the structures in which disparities exist and a means by which to identify the students who are most at-risk for marginalization by these structures.
Implications for Supporting At-Risk Chemistry Students

The implications posed by researchers investigating disparities that arise amongst chemistry students with differences in mathematics performance present hypothesized, sometimes oppositional, approaches to support at-risk student performance. For example, some researchers call for more accurate means of prediction and propose the combination of demographics, cognitive, and noncognitive measures to produced predictive models of greater sensitivity. Some propose the use of predictive measures to implement institutional cutoffs for enrollment in chemistry courses offering remedial instruction to support students’ preparation. This approach was critiqued as students identified as at-risk but capable of success in the course could be unfairly penalized. Instead, some researchers advance the use of predictive measures solely as a tool for advising student enrollment.

Alternatively, proposals have been made for more curricula-level approaches wherein predictors are used as a warning device to identify and intervene with students at-risk. However, educational institutions have been called to reconsider these practices as communicating low expectations for student performance has the potential for perpetuating unequal outcomes through a self-fulfilling prophecy effect. Beyond these opposing perspectives, investigations of these intervention approaches (remedial instruction and general intervention) have generated concerns related to the cost and limited efficacy of these approaches.

Remedial Instruction in Mathematics

Assigning prerequisite coursework in mathematics to prospective chemistry students scoring below a threshold on math test scores is a practice nearly ubiquitously observed in the United States. Yet, systematic evaluations of remedial instruction in mathematics suggests the benefit of this practice is arguable.
For example, Boatman and Long evaluated the short- and long-term effects of remedial mathematics coursework on at-risk students. Math remediation negatively impacted the retention of students both at-risk and high-achieving with no significant improvement in the outcomes of either high- or low-performing at-risk students in follow-on coursework. At the University of Akron, the percentage of at-risk students who passed on-sequence, first semester General Chemistry before institutional adoption of mathematics remediation was 76.3% (or 694 of 909) across five semesters of historical data. Two semesters following remediation, the percentage of passing at-risk students increased to 82.1% (or 161 of 196), an increase of 11 students. Whether the connection between mathematics and chemistry is dependent on “actual mathematics knowledge” or the development of “other, higher-order cognitive, skill sets required in science” is a question posited by the authors of the study who conclude:

“…if success in chemistry stems from [the] mastery of specific mathematics skills and knowledge, then those specific skills must be identified and built upon for students who will take college-level chemistry classes” (p. 381).

This quote reflects the state of the literature concerning remediation in mathematics wherein descriptions of the persistence and nature by which at-risk students so differentially perform have not yet been elucidated.

Currently, remedial coursework relies on a reasonable, but an unfounded, assumption that students of low math test scores will struggle with mathematically laden topics in chemistry. The fallacy of this assumption may be implicit to works evaluating the efficacy of remediation in mathematics for at-risk chemistry students but is made explicit in work conducted by Scott in 2012. Scott analyzed students’ response processes following practice with mathematical tasks analogous to chemistry tasks and found no measurable change in students’ chemistry performance after practicing comparable mathematics problems. The implication of this result is the limited transfer of mathematical skill sets toward preparation for the study of chemistry.
Remedial Instruction in Chemistry

An alternative approach to remedial instruction using introductory chemistry coursework to establish foundational skillsets for at-risk students enrolling in General Chemistry has also demonstrated little evidence of efficacy. For example, a six-year study concerning the benefit of remedial chemistry at Texas Tech University found no positive effects to placement remediation on academic performance while resulting in higher attrition rates than what was observed before its integration. Similar results were attained at the University of Toledo following the integration of an elementary chemistry course to support the preparation of at-risk chemistry students wherein no significant differences in the performance of students were evident.

While no evidence of improvement following remediation was detected, negative impacts related to time and financial costs to institutions, instructors, and students have been observed along with an increase in attrition rates for at-risk students. These works suggest current approaches to remediation via prerequisites assigned by institutional cutoffs are not a viable means to promote at-risk student success despite their ubiquity in educational institutions. The lack of viability emphasizes the importance of describing the nature of students' differential performance before adapting the learning environment.

Interventions of Supplemental Instruction

While remedial coursework has provided limited evidence toward altering the trajectories of at-risk chemistry students, other interventions related to supplementation instruction have been evaluated. An intervention at Columbia University in 1975 involved supplementary instruction in the form of problem-solving practice. Students scoring below a 610 on the SAT math test were asked to enroll in supplemental instruction that ran parallel to the course and involved additional practice using algorithms to solve problems in chemistry. There was an increase in the number of at-risk chemistry students attaining A’s or B’s from 10% to 19% when comparing historical and experimental data. An independent-samples t-test was
conducted to compare differences in the mean final grades of at-risk students following the supplemental instruction, where a statistically significant difference was detected ($t = 2.54, p < 0.01$). While a measure of substantive significance (e.g., effect sizes) could not be calculated from the information provided, the authors weigh the practical significance of these results and conclude that there was a cost-effectiveness problem wherein drastic changes in the curriculum did not follow with considerable differences in student performance.

At Penn State in 1997, differentials in assessment performance were used to evaluate the impact of cooperative learning on a cohort comprised of a majority of at-risk students' ($M_{SAT Math} = 530$) when compared to that of a more mathematically prepared cohort ($M_{SAT Math} = 572$) who received instruction via lecture. The at-risk students who received 100 minutes of additional instructional time and engaged in learning groups performed 10% higher than their peers on an interim exam and 3% higher on a cumulative final exam (significance not reported). Similar results were observed at the University of Texas at San Antonio in 2001, wherein 19 at-risk students were placed in a small class restricted only to these students, and 17 at-risk students were assigned to a large-group lecture ($n = 210$). The tests conducted to compare student outcomes were not named; however, a non-significant result was reported ($p = 0.1158$).

**Multi-Faceted Interventions for At-Risk Chemistry Students**

In 2014, Duke University integrated a multifaceted approach to the retention and promotion of at-risk student success. The approach included training teaching assistants to moderate study groups toward facilitating collaborative learning, hiring a chemistry learning specialist to evaluate curriculum design, and leveraging strong departmental support for absorbing program costs associated with innovating pedagogical approaches. This program was piloted with a small group of at-risk students ($n = 30$). At-risk students in the treatment group were observed to perform similarly to those of higher math test scores in first- and second-semester General and Organic Chemistry. In addition to this successful outcome,
notable retention rates for students of the intervention pilot were observed (68% compared to historical rates of 29% and 27% amongst students who declined participation in the intervention). Interpreting the findings of this study is problematic as the design does not allow for an investigation as to which change in the learning environment served equity.

While quasi-experimental designs are common in Chemistry Education Research, the cited works evaluating interventions on at-risk students enact designs with considerable threats to the internal validity, rendering interpretations of the results problematic. Except for the study conducted at Duke University, the cited interventions present no discernable evidence as to a reduction in differential performance between at-risk students and their peers. Further, there were threats to the internal validity of the data collected were observed. These include the use of historical data to compare experimental and control groups, differences in time-on-task or other competing variables, and failing to control for differences in incoming preparation in chemistry students who were non-randomly assigned to treatment. The limitations of these designs could confound the progression of evidence-based practices for the advancement of retention and performance for students at-risk in chemistry.

**Incorporating Collaborative Learning Environments**

Literature concerning attrition in first-year STEM courses calls for instructional reforms that progress from the passive transmission of knowledge to more active and collaborative classroom environments. One such reform involves the promotion of cooperative learning — any instructional technique wherein small groups of students work to achieve a common goal — where evidence supports enhancements in academic achievement, and equity, as evidenced by the retention of students belonging to underrepresented minority (URM) groups as described by the National Science Foundation. While meta-analyses of cooperative learning in chemistry, and active learning in STEM, indicate a consistent
positive effect of these pedagogical techniques on student performance overall, evidence
toward improving at-risk student performance is not as well supported.

Lewis and Lewis,\textsuperscript{54} explicitly investigated the impact of a peer-led guided inquiry (PLGI; a collaborative learning pedagogy) to mediate the relationship between SATM and student performance in chemistry. It was found that the pedagogy improved students’ average academic performance regardless of their SATM score but had no significant impact on the relationship between SATM and student performance in chemistry. Similar results concerning the effectiveness of PLGI to the overall cohort but no evidence for reductions to the dependence of student performance on incoming preparations in mathematics were found in a later study.\textsuperscript{55}

Previous evaluations of another form of collaborative learning, peer-led team learning (PLTL), in science courses posit improvements in student performance,\textsuperscript{56–59,19,60} and equity from the perspective of student retention for underrepresented minorities.\textsuperscript{61,56–60,62} PLTL can be described as a learning structure involving undergraduate students who 1) were previously successful with the course and 2) are trained to facilitate small-group learning for a new cohort of students.\textsuperscript{63,64} Evidence for the efficacy of PLTL was been attributed to fostering student-to-student interaction and prompting mastering in the course material via problem-solving.\textsuperscript{65}

Given a prior study describing the quality and quantity of not-at-risk and at-risk students’ study habits outside of the classroom,\textsuperscript{7} it is plausible problem-solving in a collaborative environment (as is structured within PLTL) could improve at-risk student performance. In this study, at-risk chemistry students engaged in PLTL were observed to study more than their not-at-risk peers, where a high rate of study was observed to mitigate lower incoming SATM scores. A cross-case analysis of the interviews conducted with at-risk students Mary and Bella revealed an expressed relationship with students’ quality and quantity of studying and their participation in study groups. Bella states:
“If you have a problem, you can go to your peers… They can explain it to you, and [sic] they have a problem, you can help them, also when you are helping people, you are kind of learning yourself.” (p. 887)

However, it is unclear whether PLTL stands as a plausible intervention for cultivating equity amongst chemistry students of differential preparations in mathematics.

**Operationalizing Equity in the Context of Science Education Reform**

National imperatives to maximize investments in science, technology, engineering, and mathematics (STEM) education, recognize the importance of equity toward promoting the growth and diversification of a scientifically, mathematically literate society.\textsuperscript{1,2,66–68} Cultivating equity has been described along numerous perspectives, including 1) advancing outcomes of a marginalized group, or *equity of individuality*,\textsuperscript{69} 2) addressing differences in outcomes, or *equal outputs*,\textsuperscript{69,70} and 3) achieving *equality of outcomes* wherein all students have an equal opportunity to attain academic success.\textsuperscript{70} In evaluations of science education reforms, each perspective offers unique insight as to how equity can be cultivated.

Sharon Lynch advances an “equity schema” to frame science education reforms targeted toward cultivating equity and achieving equality of outcomes (characterized as distribution curves of academic performance across groups wherein the number of high- and low-achieving students within each group is similar).\textsuperscript{70}

“There is an implicit, long-term, but clearly definable goal that involves both closing achievement gaps while raising standards for all. This ultimately will result in equal representation of various groups in science-related careers at all levels.” (p. 30)\textsuperscript{70}

The schema can be envisioned as a feedback cycle wherein evaluators: 1) *investigate* where equal outputs are observed by students of various groups, 2) if not, *describe* the mechanisms through which differences persist to achieve an understanding of how to advance equity of individuality within the marginalized group, 3) *adapt* the learning environment to address
disparities in performance, 4) *evaluate* the extent to which equality of outcomes has been achieved, and 5) if addressed, *advance* standards for all (see Figure 1).

![Diagram](image)

**Figure 1.** A visual summary of the cycles by which equality of outcomes can be evaluated.

These perspectives on equity and evaluations of science education reform are pertinent in explorations of the literature base, informing the disparities observed in academic science performance amongst students of variable preparations in mathematics and the evidence-informed practices intended to cultivate equity in STEM education.
Past Works and Lynch’s Framework for Evaluating Science Education Reforms

For example, studies seeking to predict which students are at-risk are widely representative of the first stage in the cycle toward attaining equality of outcomes (see Figure 2) and share a common objective in *investigating* disparities amongst chemistry students.

![Image of equity schema]

Figure 2. Citations related to each stage of the equity schema demonstrating a progression from investigate to adapt and evaluate with limited research describing the nature of the disparities observed.

A single study involving at-risk student study habits,⁷ was identified to *describe* the persistence or nature of differences amongst chemistry students of variable preparations in mathematics.

This suggests that while much of the literature base calls for *adaptation* involving either 1) remedial or supplemental coursework in mathematics and introductory chemistry or 2) changes in pedagogical approaches to emphasize collaborative learning environments centered around group work, few works *describe* the nature of the disparities observed. This lack of description and understanding surrounding how these disparities in performance manifest in the course serve as a potential explanation for the limited success followed by past interventions.
As such, understanding and describing may be a necessary predecessor to adaptation and evaluation toward realizing commitments to equity and diversity amongst a well-prepared STEM workforce.

Directions of this Work

Prior works present a compelling and consistent evidence-base for the use of pre-college math test scores in investigating the extent of the disparities observed and predicting which students are most at-risk for unfavorable outcomes in general chemistry courses. \(^7,8,19,20\)

However, there is a dearth of the literature describing which aspects of the content in chemistry courses present the greatest difficulty to at-risk students and why these differentials persist.

This dissertation concerning inequities observed amongst chemistry students of differential preparations in mathematics was informed by the following research objective: Describe the pedagogies, content, and reasonings with which at-risk students most incommensurately struggle in general chemistry courses. This research seeks to inform research and practice related to instructional, assessment, and intervention practices targeting the advancement of at-risk student success and toward cultivating equity amongst students with differential preparations in pre-college mathematics.

References


(54) Berry, L. Collaborative Learning: A Program for Improving the Retention of Minority Students; 1991.


CHAPTER 2:

METHODOLOGIES USED TO EVALUATE ACADEMIC PERFORMANCE

This chapter describes relevant aspects of the setting, methods, and data analyses for the presented works. Methodologies specific to individual studies will be described in subsequent chapters.

Note to Reader

Portions of this chapters have been previously published in Chemistry Education Research and Practice, 19: 867-884, 20: 570-593, and 21: 287-306. These works have been reproduced for this dissertation with permission from the Royal Society of Chemistry (see Appendix B for Copyright Permissions).

Research Setting

The studies presented throughout this dissertation were conducted at a large, public and doctorate-granting institution in the Southeastern United States. The institution offers multiple classes of first- and second-semester general chemistry (GCI and GCII, respectively) during the fall and spring semester. The institution also offers a single class of second-semester general chemistry during a 6-week summer term. These studies analyze data collected fall and spring semesters only with data collection spanning seven semesters.
Class Sizes

On average, eight classes of first-semester general chemistry courses were attended by 241 students in fall (on-sequence) semesters with four classes of 176 students attending in the spring (off-sequence). Typically, three GCII classes were offered in the fall (off-sequence) semester and attended by an average of 250 students with five classes in the spring (on-sequence) semester, attended by an average of 248 students.

Content and Prerequisites

GCI courses involved instruction as to the properties of substances and reactions, thermochemistry, atomic-molecular structures, bonding, and the periodic properties of elements and compounds. The institution list prerequisites for enrollment in GCI list as a 570 on the math composite of the SAT or complete a college-level algebra course with a passing grade (defined as a “C” or better), in addition to either one year of high school chemistry or the completion of the university’s introductory chemistry course with a passing grade. For enrollment in GCII courses, prerequisites include either GCI or General Chemistry for Engineers. GCII courses included instruction on the principles and applications of solutions chemistry, chemical thermodynamics, kinetics, equilibria, electrochemistry, and nuclear chemistry.

Coordination and Grading Schemes

All general chemistry classes were coordinated across instructors of the same course with a shared textbook, learning objectives, syllabus, grading scheme, and online learning management platform. Students’ grades were comprised of three interim exams (45% of total grade, 15% each exam), a final exam (25% of final grade), online homework (10%) and participation driven grading systems (e.g., clickers, quizzes, or participating in peer-led sessions where applicable) for the remaining 20%.
Pedagogy

All sections of GCI are conducted using peer-led team learning (PLTL), a pedagogy described in the introduction. GCI students attended regular lecture sessions twice per week and a peer-led problem-solving session once per week.\(^1,2\) In GCII, students self-select into sections that do or do not enact PLTL (relying, instead, on lecture-based or didactic instruction). GCII classes met twice weekly for 75-minutes each with students in PLTL attending one lecture-based class meeting and another meeting per week for a PLTL problem-solving session where students were assigned a set of ten to fifteen problems to work through and encouraged to work together on the assignment.

Whether or not PLTL is conducted in GCII was dependent on a variety of institutional factors including funding, instructor availability, and enrollment. To promote the sustainability of the reform via developing a population of instructors with experience enacting PLTL, all faculty were encouraged to participate in implementing this pedagogy. When teaching PLTL, instructors assigned 10% of student grades to student responses on the classroom response system that were used during both the lectures and the problem-solving sessions. They also assigned 5% to student attendance to the problem-solving sessions and 5% to weekly online quizzes due before each problem-solving session. In courses not using PLTL, both class sessions per week employed the lecture format similar to above with the instructor providing didactic instruction and working problems in front of the class. Instructors using didactic instruction also assigned 10% of student grades to responses in the classroom response system, which was used throughout the lecture instruction.

Retention and Pass Rates

Retention (operationalized as the percentage of students enrolled who go on to complete the GCII final exam) for students attending GCII classes for which PLTL was enacted had a comparable rate (94.6%) to those receiving didactic instruction (90.8%). Consistent with
prior evaluations of PLTL, pass rates (percentage of students receiving an A, B, or C in the course) were considerably higher for GCII students receiving PLTL, 93.5%, than those who did not, 84.8%.3–9

**Instruments Assessing Chemistry Knowledge**

Interim exams were common across all classes within a semester and course (GCI or GCII) and were written by a committee of the students’ instructors. These exams consisted of 20 multiple-choice assessment items with four distractors (five answer choices in total) and a series of six true-or-false items following the Measure of Linked Concepts format to emphasize the links across topics in the course.10 Each multiple-choice item on the tests was worth seven points, and each true-or-false item was worth three points for a correct response or one point for selecting unsure (in an attempt to reduce chance guessing) for a total of 158 points.

Final, cumulative exams in GCI were either the 2015 First Term General Chemistry Exam available from the ACS Examinations Institute,11 or followed a similar format with 45 multiple-choice and ten true-or-false assessment items. Final exams in GCII were invariably the latter. Students’ final ACS exam scores were calculated from the percent correct achieved on the ACS exam multiplied by a total of 250 points. For simplicity in comparisons, all scores reported herein are formatted as a percentage of possible points earned. While no two interim tests for the responses collected were identical, each set of tests were written from the same list of learning objectives (presented in the Appendix C) and administered to students throughout the semester in roughly equal intervals of time.

The data presented herein are of student responses to multiple-choice assessments administered as exams normally collected in the instructional setting. Students’ assessment responses were collected in a controlled setting wherein exams were proctored by teaching assistants and instructors. Students received a cover sheet with a periodic table, were permitted to bring a nonprogrammable, scientific calculator, and were offered scratch paper.

21
Ethics

All data collection adhered to guidelines set by the institutional review board at the research setting in the evaluation of the effectiveness of curricular approaches on student success in chemistry. Beyond this, a number of considerations in the design and approach of these research studies were undertaken from the perspectives of impacts on the instructors and students. Where evaluative measures were taken from the context of an instructor’s classroom, scores were grand mean centered within instructor to prevent cross comparison across instructors. This was of importance as I believed instructors should not be compared to their peers or otherwise penalized for experimenting with pedagogical implementations. When a specific instructor or student is referenced in the data, a pseudonym (for which gender, race, or ethnicity should not be assumed) was always used to both provide the reader a name with which to associate the data and protect the identity of the participant.

Student performance is always presented in the collective (or by group) as, similarly to measures related to a specific instructor, to prevent self- or cross-comparison between students. Student groups, however, were directly compared in an effort to examine equity from both the perspective of comparative equity (are student groups performing similarly) and individuality (are interventions useful in supporting the group of interest). These two perspectives were often used in tandem. While I do not intend to situate not-at-risk chemistry students (or those with higher math test scores) as the “ideal” or “normative” group, examining the extent to which inequities are perpetuated is a valuable discussion toward cultivating equity. Further, these works are adopted with the perspective that at-risk chemistry students (those scoring in the bottom quartile of math test score within each cohort) are not in some way deficient but are performing differentially as a potential indicator for inequities perpetuated in the systems derived by their learning environment. These dissertation works are presented with an acknowledgement of bias in the aversion toward deficit perspectives used in equity research.
Identifying Disparities Amongst General Chemistry Students

At the research setting, correlations between students’ math composite scores and final, cumulative exam performance in GCI ranged from 0.37 to 0.56 in the seven semesters for which data were collected. The range is similar to those previously published in the literature (0.42 to 0.63),\textsuperscript{12–18} supporting the trend of precollege mathematics scores predicting students’ college-level chemistry performance. The average correlation between these measures for students enrolled in off-sequence semesters (0.39) was less than that of on-sequence semesters (0.54). This suggests that math composite scores were more predictive of variations in student performance in on-sequence semesters than off-sequence semesters.

At-risk students were identified as students who scored in the bottom quartile of math composite scores for the student cohort in GCI, a method consistent with the prior literature.\textsuperscript{19–22} As such, a student identified as at-risk in GCI would also be at-risk in GCII. Students enrolled with math composite scores comprised the majority at this research setting with 80-93% of cohorts represented by these measures (see Table 1).

<table>
<thead>
<tr>
<th>Measures</th>
<th>N</th>
<th>n_valid</th>
<th>n_missing</th>
<th>Median (SD)</th>
<th>Top</th>
<th>Bottom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 2016</td>
<td>902</td>
<td>772</td>
<td>130</td>
<td>560 (64)</td>
<td>590</td>
<td>520</td>
</tr>
<tr>
<td>Fall 2016</td>
<td>1991</td>
<td>1841</td>
<td>150</td>
<td>600 (74)</td>
<td>650</td>
<td>560</td>
</tr>
<tr>
<td>Spring 2017</td>
<td>714</td>
<td>568</td>
<td>146</td>
<td>560 (65)</td>
<td>590</td>
<td>530</td>
</tr>
<tr>
<td>Fall 2017</td>
<td>1994</td>
<td>1852</td>
<td>142</td>
<td>610 (69)</td>
<td>650</td>
<td>560</td>
</tr>
<tr>
<td>Spring 2018</td>
<td>755</td>
<td>634</td>
<td>121</td>
<td>560 (64)</td>
<td>590</td>
<td>520</td>
</tr>
<tr>
<td>Fall 2018</td>
<td>2035</td>
<td>1881</td>
<td>154</td>
<td>610 (70)</td>
<td>660</td>
<td>580</td>
</tr>
<tr>
<td>Spring 2019</td>
<td>795</td>
<td>662</td>
<td>133</td>
<td>570 (71)</td>
<td>620</td>
<td>530</td>
</tr>
</tbody>
</table>

Quartiles for students’ math composite scores on the SAT (SATM) or concorded scores on the ACT (ACTM) where SATM was not available,\textsuperscript{23,24} within each cohort of GCI students were
calculated by semester and listed in Table 1. At-risk students were identified as those scoring at or below each of these cutoffs for the given semesters of GCI with on-sequence cutoffs ranging from 560 to 580 and off-sequence from 520 to 530. These scores are near the range of scores (below 550) identified in prior works for which students were observed to differentially achieve negative outcomes in general chemistry, and were below the institutional cutoff permitting students to enroll for the course only after completing a college-level algebra course.

Concerns Regarding Equity and Diversity

Of the 8,210 students entering the university with math composite score, 6,083 and 2,124 were identified as not-at-risk and at-risk, respectively. Students at the research setting are asked to identify from a predetermined set of races or ethnicities. To investigate the impact of risk determinations made by math composite score on diversity and representation at the research setting, the percent of students who self-identify as each of the races and ethnicities were calculated (see Table 2).

Table 2. Race or ethnicity disaggregated by risk for students enrolled in GCI at the research setting.

<table>
<thead>
<tr>
<th>Race or Ethnicity</th>
<th>Not-At-Risk (NAR)</th>
<th>At-Risk (AR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>6083</td>
<td>2124</td>
</tr>
<tr>
<td>White</td>
<td>47%</td>
<td>38%</td>
</tr>
<tr>
<td>Asian</td>
<td>14%</td>
<td>8%</td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>20%</td>
<td>25%</td>
</tr>
<tr>
<td>Black or African American</td>
<td>10%</td>
<td>18%</td>
</tr>
<tr>
<td>Non-Resident Alien</td>
<td>5%</td>
<td>7%</td>
</tr>
<tr>
<td>Ethnicities less than 5%(a) or Unknown</td>
<td>4%</td>
<td>4%</td>
</tr>
</tbody>
</table>

\(a\) Native Hawaiian/other Pacific Islander or American Indians or Alaska Native

Amongst all ethnicities, a higher percentage of students having self-identified as Hispanic or Latino, Black or African American, or Non-Resident Alien were categorized as part of the at-risk
group than the not-at risk group. These results suggest preparation in secondary-school mathematics, using SATM or ACTM as a proxy, is not equitably achieved by students of underrepresented minorities in the sciences as defined by the national science foundation.²⁸

This result also corresponds with national trends in the U.S. where the average SATM score for Black or African American students is 489 and Hispanic or Latino students is 463, each below the overall national average of 531.²⁴ As such, addressing disparities in at-risk student performance could serve to address the lack of representation observed amongst those engaged in the sciences.²⁸⁻³¹

The higher than expected percentage of Black/African American or Hispanic/Latino students in the at-risk group suggests aiding at-risk students has the potential for increasing retention and improving the diversity of students who complete undergraduate degrees in chemistry and potentially other STEM fields. An analysis of the performance for students who identify as an underrepresented minority is not within the scope of this dissertation. However, the commonality between at-risk students and those of underrepresented minorities is worth noting as efforts to understand the persistence of differential performance between at-risk students and their peers are also likely to support students of underrepresented minorities.

Examining the Extent of Inequity Observed at the Setting

Given the strength of correlations between students’ math composite scores and final, exam performance, disparities between students identified as not-at-risk (NAR) and at-risk (AR) were expected. To examine the extent to which these disparities persisted, distributions of course outcomes (final course grades) achieved by students in GCI and GCII were disaggregated by risk (see Figure 3).
Figure 3. Sankey diagram where the number of not-at-risk and at-risk students (left) is represented by the width of the bands connecting students to the final grades achieved (sorted in descending likelihood) in first- and second-semester general chemistry courses (GCI and GCII, respectively; middle and right).

Figure 3 is a representation that depicts the frequency of final course outcomes in GCI and GCII for NAR and AR students in descending order of likelihood. NAR students were most often observed to attain A’s in GCI and GCII whereas AR students were most likely to achieve B’s and not enroll in GCII. More precarious differences are observable in what students of each group were least likely to achieve. While NAR students were least likely to attain C’s and a DFW in GCI and GCII (respectively), AR students were least likely to attain A’s in GCI and GCII.

To investigate and quantify the progress needed to achieve equality of outcomes as described by Lynch, the percent of NAR and AR students who achieved each course outcome were calculated (see Table 3).
Table 3. The percentage of students from each group (not-at-risk or NAR and at-risk or AR) who achieved each final course outcome in first- and second-semester general chemistry courses (GCI and GCII, respectively).

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>DFW</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAR (75%)</td>
<td>90%</td>
<td>69%</td>
<td>52%</td>
<td>56%</td>
</tr>
<tr>
<td>AR (25%)</td>
<td>10%</td>
<td>31%</td>
<td>48%</td>
<td>44%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Group</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>DFW</th>
<th>Did not enroll</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAR (75%)</td>
<td>92%</td>
<td>76%</td>
<td>58%</td>
<td>58%</td>
<td>67%</td>
</tr>
<tr>
<td>AR (25%)</td>
<td>8%</td>
<td>24%</td>
<td>42%</td>
<td>42%</td>
<td>33%</td>
</tr>
</tbody>
</table>

Comparing distributions of these final course outcomes to the percent of the cohorts these students generally comprised, AR students attained 44% of the DFW’s in GCI and only 10% of the A’s despite representing 25% of the student cohort (those scoring in the bottom quartile of math test scores within their GCI cohort). In GCII, these disparities worsened with 42% of AR students attaining DFW, 33% choosing not to enroll in the follow-on course, and only 8% achieving an A.

Having achieved equality of outcomes, wherein all students have an equal opportunity to succeed in general chemistry courses despite differences in incoming mathematics preparation, these percentages would be closer to 75% and 25% for the NAR and AR student cohorts. To achieve a curriculum wherein students are not penalized for differences in preparation, the equity of individuality, or success of the marginalized group, must be improved by roughly 15%. This could be achieved via remedial coursework, pedagogical innovation, or intervention if the persistence of these differences despite these measures were identified and explained.

To be clear, we are not proposing that chemistry performance and mathematics proficiency are unrelated, as multiple chemistry topics require quantification and mathematic manipulation. Instead, we hypothesize that students enrolled in general chemistry at the research setting have the requisite algebraic manipulation skills needed for general chemistry
(given the requirements to enroll in the course described above). Part of the goal of this dissertation is to explore the feasibility of this hypothesis.

If the hypothesis holds, students have the requisite mathematics skillsets needed to succeed in chemistry and math composite scores measure additional differences between the students that should be explored to cultivate equity in general chemistry course. Alternatively, if the hypothesis fails and algebraic manipulation skills manifest as the reason for student struggles this would call for additional preparation aimed at developing this skill set.

References


CHAPTER 3:
AN EVALUATION OF COLLABORATIVE LEARNING AND ITS IMPACT ON STUDENT PERFORMANCE AND EQUITY

Note to Reader

Portions of this chapters have been previously published in Chemistry Education Research and Practice, 21: 287-306. These works have been reproduced for this dissertation with permission from the Royal Society of Chemistry (see Appendix B for Copyright Permissions).

The Evidence in Evidence-Based Practices

Literature concerning attrition in first-year science, technology, engineering, and mathematics (STEM) courses, call for instructional reforms that progress from the passive transmission of knowledge to more active and collaborative classroom environments.\textsuperscript{1-3} One such reform involves the promotion of cooperative learning —any instructional technique wherein small groups of students work to achieve a common goal — where the evidence supports enhancements in academic achievement,\textsuperscript{4,5} and greater equity evidenced by the retention of students belonging to underrepresented minority (URM) groups as described by the National Science Foundation.\textsuperscript{4,6} As with any reform, challenges have arisen in evaluating the impact of cooperative learning, raising concerns on the robustness of the evidence-base.\textsuperscript{7} As Michael posits, while active learning often concerns what students do in these courses, it is the teacher that fosters the environment, and exploring what teachers do in implementing reforms is woefully underexplored. As federal legislation mandates the use of evidence-based instructional practices,\textsuperscript{8,9} these concerns could hinder the convergence of viewpoints among practitioners, researchers, and policymakers.
A Shift in the “Gold Standard”

Randomly assigning students to one of two instructional conditions for comparisons, enacting a randomized control trial, was the original gold standard in evaluating large-scale efficacy or replication studies of a well-developed intervention.\textsuperscript{10} Random assignment is thought to reduce alternative explanations as to the effect of a reform that may be explained by student-level differences in incoming preparation but requires large sample sizes spanning multiple sites with similar implementations in the interest of establishing generalizability. Random assignment within authentic educational settings is often challenging owing to the common practice of self-directed student enrollment in courses.\textsuperscript{10,11}

This “gold standard” has shifted in educational research, replacing randomized control trials with research designs enacting a variety of complementary methodologies to identify, measure, and explain treatment effects.\textsuperscript{12–15} In line with this progression toward methodological pluralism, quasi-experimental studies have sought to account for student-level differences in incoming preparation by matching groups of students on a measure thought to be explanatory of these differences (e.g., pre-college math test scores, final exams scores for prerequisite courses) or using these measures as covariates in statistical models to control for their impact.\textsuperscript{11} While more easily implemented in an educational setting, quasi-experimental designs present considerable alternative explanations for differences in outcome that could be attributable to instructor-level differences, such as differences in implementation, times of day, classroom resources, attrition rates,\textsuperscript{11,16} Hawthorne effects related to effectiveness over time,\textsuperscript{17,18} and contamination effects related to students sharing course materials.\textsuperscript{19,20}

Challenges to Enacting Randomized Control Trials in Educational Settings

Beyond instructor-level differences and trends in effectiveness over time, the statistical tests used to analyze data collected of these designs adhere to assumptions that may be
tenuous in an educational setting: 1) data are obtained via random sampling from a well-defined population, 2) observations are independent, and 3) measurements of interest have a normal distribution. The second assumption becomes even less tenable in STEM courses where advancements in instructional reforms often promote cooperative learning, heightening the impact that one student’s learning experience has on another. Further, these tests require sufficient sample sizes to detect meaningful differences between student groups that could pose additional challenges when applied to educational settings representative of smaller class sizes.

To demonstrate the utility of randomization tests in complementing works seeking to establish evidence-based practices for the instruction of science, the impact of peer-led team learning (PLTL) on students’ academic performance will be examined as one form of cooperative learning shown to increase student learning and improve retention in STEM majors. PLTL can be described as a learning structure involving undergraduate students who 1) were previously successful with the course, and 2) are trained to facilitate small-group learning for a new cohort of students. PLTL is thought to be effective owing to the student-to-student interactions and collaborative environment towards mastering the course material via problem-solving. Previous evaluations of PLTL in science courses posit improvements in student performance, and equity from the perspective of student retention. This study seeks to apply a novel methodology to advance the methodological pluralism observed in the research literature concerning evaluations of PLTL as an instructional reform in comparison to the reference (didactic or lecture-based) instructional pedagogy.

Objectives of this Study

In the interest of advancing this shift toward methodological pluralism as a gold standard for evaluating STEM educational reforms, this study seeks to present randomization tests as an underutilized, quantitative methodology that 1) complements non- and quasi-experimental research designs, 2) is not beholden to parametric statistical assumptions and sample sizes,
and 3) addresses the instructor-level differences that could threaten the internal validity of
evidence collected in an evaluation of an instructional reform. The work was guided by two
research objectives:

1. Introduce the STEM education community to an alternative, experimental and
   randomized methodology that could prove a helpful complement to other research
designs establishing evidence-based instructional reforms.

2. Enact randomization tests in an evaluation of the impact of peer-led team learning on the
effectiveness of chemistry instructors toward promoting improved and more equitable
academic achievements amongst students in second-semester general chemistry
courses.

Introducing Randomization Tests

A Shift in Randomization

Randomized control trials applied in an educational setting involve the random
assignment of students to a reference (control) or reformed instructional practice and require
uniform, multisite implementation to promote generalizability.\(^{20}\) Alternatively, randomization tests
investigate the impact of a reform enacted on a single case (e.g., a small group of, or individual,
instructors/students/schools) serving as its own control and randomly assigned to reforms
during distinct phases of the study.\(^{21,34}\) In the context of educational research, single-case
research designs have been used to evaluate and establish evidence-based practices in the
behavioral sciences,\(^{34}\) and special education.\(^{35}\)

To illustrate the differences in these approaches to randomization, consider a research
study that evaluates an instructional reform by comparing academic outcomes of students
taught by two science teachers using either a reference (control) or reformed instructional
pedagogy (see left tile of Figure 4).
Figure 4. Illustration of the differences between two methodologies that enact randomization in the assignment of students (randomized control trial) versus the assignment of pedagogy over a period of time (randomization tests).

In a **randomized control trial**, students would be randomly assigned to one of the two science teachers, each enacting a different pedagogy. The data compared could consist of students’ academic performance (test scores), attrition rates, differentials observed between groups of interest, changes in affect, and observational data concerning students’ participation or engagement. Differences in student performance attributable to differences in incoming preparation are assumed non-existent owing to the random assignment to groups. As is the case with any research design, the approach has its limitations. Instructor-level differences, trends in effectiveness over time within an instructor (relating to gains in experience implementing the reform), and the assumption of statistical independence as applied to an educational setting could pose threats to the internal validity of the evidence acquired resulting in limited or distorted detections of differences between student groups.

Randomized control trials are also limited by way of sample-size. To enact an independent-samples *t*-test comparing mean academic performances for students participating in either the reference or reformed instructional pedagogy, each group would (at minimum) require 50 students for sufficient statistical power ($\pi = 0.8$) to detect statistically significant
differences ($\alpha = 0.05$) of medium effect size or $d = 0.5$. This sample-size requirement limits the educational settings where this methodology is applicable, potentially removing settings with smaller class sizes — such as smaller institutions, upper-level STEM courses are commonly offered as a solitary section each semester attended by fewer than 50 students, or specialized courses such as teacher training — from consideration in evaluations of instructional reform.

Restructuring the study to a randomization tests design, these two science teachers could enact either the reference or reformed instructional pedagogy at the beginning of the school year and randomly select starting points to toggle to the alternative pedagogy for their respective students (see right tile of Figure 4). Using repeated student-level measures related to academic performance, differential performances amongst groups of interest, attrition rates, affect, or observational data, comparisons of the data collected before and after a change in pedagogy are conducted using randomization tests (see Figure 5).

Figure 5. Graphical representation of hypothetical data used in randomization tests design for students of one or more instructors.
Differences in student performance are then compared within each group of students and between phases of the instructional pedagogy applied (reference or reformed). For example, consider instructor 1 in the randomization tests approach. This instructor implemented the reference pedagogy at the start of the semester and toggled to the reformed pedagogy at the fourth measure (fourth test, observation, survey collected, etc.). The differences in student performance for phase A (measures 1-3) and B (measures 4-7), compared to signals that arise with all possible arrangements of the data, represents the signal denoting the efficacy of the reform pedagogy. For example, the reform could have taken place after the second measure, and the data for measures 1 through 2 would be compared to 3 through 7. The number of hypothetical arrangements that meet or exceed the observed signal informs the probability that the change in student performance is significant (details to be discussed). This allows for the performance of each instructor’s students to serve as its own control in estimating the impacts of the reference and reformed instructional pedagogies and can be enacted in any educational setting regardless of sample size.

In this approach, differences in academic performance are not likely the result of student-, instructor-, or institutional-level differences (as these variables are held constant) and does not rely on a theoretical distribution of data adherent to assumptions tenuous in an educational setting (e.g., randomly sampled, independent, normally distributed). Thus, randomization tests provide a means to quantify the effect of an instructional reform and address some of the challenges in enacting a randomized test control in an educational setting. The design does involve a trade-off in generalizability in favor of reliability, as these data are specific to a group of students taught by a single instructor at a particular educational setting. Even so, randomization tests represent an underutilized methodology in education research that can serve as a strong complement to prevalent research designs.
How Randomization Tests Work

A hypothetical example is detailed to describe how to enact a randomization test. Consider a researcher designs a single-subject research study concerning whether the number of questions posed to students by a science instructor changes following the enactment of a different instructional pedagogy during an academic year. The reform is hypothesized to increase the number of questions an instructor poses to students. An observer counts the number of content-related questions asked by the instructor in each of 20 randomly selected lessons before and after the implementation of the instructional reform (see Table 4).

Table 4. The number of questions (Q's) posited by an instructor to students, observed over 20 lessons with at least four observations per phase (possible starting points highlighted).

<table>
<thead>
<tr>
<th>Lesson</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q's</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>14</td>
<td>12</td>
<td>12</td>
<td>15</td>
<td>14</td>
<td>14</td>
<td>15</td>
<td>17</td>
<td>20</td>
<td>16</td>
<td>17</td>
<td>15</td>
<td>22</td>
<td>18</td>
<td>11</td>
</tr>
<tr>
<td>Actual</td>
<td>Reference</td>
<td>Reform</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Possible</td>
<td>13 Potential Starting Points</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As upward trends in the baseline can introduce threats to the internal reliability of the data collected in single-case research designs, the researcher evaluates trends in the data as it is received. Having observed no upward trends in the data after the 4th lesson, the researcher determines the baseline stable. Using this information, the researcher determines each phase of the study (Reference and Reform) should have a minimum of 4 lessons to enact a similar evaluation of trend and within-phase stability. Given 20 lessons observed and a minimum of 4 lessons for each phase, 13 potential starting points for the intervention are determined with the reform, possibly starting anytime from the 5th lesson through the 17th lesson (see the final row in Table 4). Randomly selected, the enactment of the reformed instructional pedagogy began with the 6th lesson (see the second row in Table 4).

This design is known as an “AB” design wherein phase A constitutes the reference, and phase B the reformed instructional pedagogy. Reversal designs (e.g., ABA, ABAB, and so on)
can also be implemented to investigate whether the effect of the intervention can be demonstrated multiple times in cases where the reform can be meaningfully removed from the setting.

For the 13 lessons in which the intervention could have begun, the absolute value of mean differences for each phase (or $|\bar{X}_B - \bar{X}_A|$) is calculated (see Table 5).

Table 5. Absolute values of differences in the mean number of questions asked in each lesson (actual starting point in blue).

| Starting Point | $\bar{X}_B$ | $\bar{X}_A$ | $|\bar{X}_B - \bar{X}_A|$ |
|----------------|-----------|-----------|-----------------|
| Lesson 5       | 15.06     | 5.50      | 9.56            |
| Lesson 6       | 15.80     | 5.20      | 10.60           |
| Lesson 7       | 15.93     | 6.67      | 9.26            |
| Lesson 8       | 15.85     | 8.14      | 7.70            |
| Lesson 9       | 16.17     | 8.63      | 7.54            |
| Lesson 10      | 16.27     | 9.33      | 6.94            |
| Lesson 11      | 16.50     | 9.80      | 6.70            |
| Lesson 12      | 16.78     | 10.18     | 6.60            |
| Lesson 13      | 17.00     | 10.58     | 6.42            |
| Lesson 14      | 17.00     | 11.08     | 5.92            |
| Lesson 15      | 16.50     | 11.71     | 4.79            |
| Lesson 16      | 16.60     | 12.00     | 4.60            |
| Lesson 17      | 16.50     | 12.31     | 4.19            |

1 Mean number of questions asked in phase B.
2 Mean number of questions asked in phase A.

These values can serve to describe the likelihood that a random starting point generates a difference between phases $|\bar{X}_B - \bar{X}_A|$ that is larger than or equal to the difference in the sixth lesson when the intervention was enacted. The probability that the effect of the intervention is attributable to chance (analogous to $p$, the conventional measure of statistical significance) is equal to the number of values for $|\bar{X}_B - \bar{X}_A|$ greater than or equal to the test statistic (10.60, blue row in Table 5), divided by the number of possible arrangements for the randomization distribution of the data.
To calculate the number of possible arrangements \((O)\), the researcher would consider the total number of observations \((N)\), the minimum number of observations within each phase that would support evaluations of trends in the data \((n)\), the number of changes between the instructional pedagogies the instructor will enact \((k)\) in the design \((e.g., AB, ABA, ABAB)\), and the number of phase arrangements \((c)\) considered for the design \((e.g., AB vs. BA)\); see Equation 1 below.\(^{37}\)

Equation 1. Permutations formula for randomization tests.

\[
O = \frac{(N - n(k + 1) + k)!}{(N - n(k + 1))! k!} \cdot c = \frac{(20 - 4(1 + 1) + 1)!}{(20 - 4(1 + 1))! 1!} \cdot 2 = \frac{13!}{12! 1!} \cdot 2 = 13 \cdot 2 = 26
\]

In this example, 20 lessons are randomly observed \((N = 20)\), with a minimum of 4 lessons in each phase \((n = 4)\), enacted with one change between phases \((k = 1)\), and with either AB or BA arrangement possible \((c = 2)\), there are 26 possible arrangements of the data. From Table 5, it is found that only one value was equal to or greater than the test statistic of 10.60. Thus, the probability of obtaining a test statistic as large or greater than the test statistic for all possible random assignments is \(p = 1/26 = 0.0385\). Using a statistical significance threshold of 0.05, the researcher can conclude the two instructional pedagogies statistically differ in effectiveness for promoting the instructor to pose questions to their students.

**Methods**

**Instructors and Survey Items**

Four instructors each comprise a case used within the study to investigate the impact of changes in instructional pedagogy on the academic performances of each instructor’s students. While instructors were aware that student performance data were being collected for general research purposes, the application of these data to this research study was not made known
until after data collection was complete. The results of this study were shared with instructors with informed consent received by each before publishing. Pseudonyms are used to communicate findings relevant to each case (Parker, Davis, Morales, and Stacy) and are not intended to communicate instructors' gender, race, or ethnicity.

In an effort to better understand emerging trends in the data concerning the impacts (or lack thereof) following enactments of reformed instructional pedagogies, an exploratory, qualitative study was facilitated by the administration of open-ended survey prompts to the four instructors whose cases were evaluated. A case study approach was a natural, qualitative complement to the quantitative single-subject research design enacted and has been used to successfully navigate instructor perspectives under the lens of the pedagogical content knowledge (PCK) framework. These insights were intended to support descriptions of the instructor’s experience, comfort with, and perceived benefits or challenges in enacting the reference (didactic, or lecture-based) and reformed (PLTL) instructional pedagogies.

Prompts were modeled after Loughran’s CoRes approach, to identify factors that may influence instructor’s practices in implementing these instructional reforms (see Box 1).

**Box 1. Survey items administered to second-semester general chemistry instructors.**

1. For how many semesters were you the primary instructor (at any institution) of second-semester general chemistry?
2. What past experiences inform your teaching?
3. On a scale from 0 (ineffective) to 10 (exceptionally effective), how effective are you at delivering lecture-based instruction? Please explain.
4. On a scale from 0 (ineffective) to 10 (exceptionally effective), how effective are you at enacting peer-led team learning? Please explain.
5. How has participating in peer-led team learning affected your delivery of lecture-based teaching?
6. Describe any challenges you have had in directing instructional time toward students participating in peer-led problem-solving?
7. Why did you want to teach using peer-led team learning?
These data were collected with the informed consent of the participating instructors and the approval of an Institutional Review Board to better understand the impact of pedagogical reform in gateway courses in chemistry on students’ academic performance and persistence in STEM disciplines.

Effectiveness and Equity as Outcome Measures

Problematizing Effectiveness: Mean Standardized Residual (MSR)

This study defines cases on the instructor-level so that each instructor serves as their own control to model the use of randomization tests in evaluations of instructional reforms. As mentioned, multiple classes of second-semester General Chemistry (GCII) were given the same exams concurrently at the research setting. During the semester, four different exams are given. To compare data across various tests and different semesters, with tests varying in the content covered and difficulty, each exam was standardized. This way, a positive standardized score indicates that the class outperformed the cohort of classes that took the same exam. There was also a concern that across semesters each class may have student-level differences in incoming preparation. At the setting, first-semester General Chemistry (GCI) also enacts common exams. To control for differences in incoming preparation, GCI final exams were standardized each semester, and a regression was conducted where students’ most recent standardized GCI final exam scores predicted each standardized GCII exam, and the residuals were saved. For each exam and each class, the mean standardized residual (MSR) was calculated. Finally, to protect instructor confidentiality and prevent comparisons of instructors, the set of MSRs for each instructor was centered, making the average MSR for each instructor equal to zero.

Authentic to the research setting, changes from didactic instruction to PLTL within a semester are drastic enough to limit the practicality of enacting this shift mid-semester. Rather
than interpreting the effects of the reform on the same group of students within a semester, these data reflect whether students’ performance for multiple groups exceeded or fell short of predictions made using their performances in first-semester General Chemistry. This, of course, is dependent on the assumption that students’ performance on their final exams in GCI predict their performance in GCII, an assumption that seems reasonable given the results of simple linear regressions calculated to predict student performance on each of their assessments in GCII (see Table 6).

Table 6. Students’ GCI Final exam scores regressed onto each of their GCII exams.

<table>
<thead>
<tr>
<th>Regressions</th>
<th>Unstandardized (b)</th>
<th>p-value</th>
<th>Adjusted R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>GCII Test 1</td>
<td>(Constant) -.117</td>
<td>&lt; .001</td>
<td>.433</td>
</tr>
<tr>
<td>GCI Final</td>
<td>.694</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant) -.126</td>
<td>&lt; .001</td>
<td>.432</td>
<td></td>
</tr>
<tr>
<td>GCI Final</td>
<td>.702</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant) -.124</td>
<td>&lt; .001</td>
<td>.390</td>
<td></td>
</tr>
<tr>
<td>GCI Final</td>
<td>.660</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(Constant) -.150</td>
<td>&lt; .001</td>
<td>.513</td>
<td></td>
</tr>
<tr>
<td>GCI Final</td>
<td>.757</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Students’ GCI final exam performance was a significant predictor of their performances on each of the four GCII assessments accounting for 39.0% to 51.3% of the variances observed between students. To demonstrate how mean standardized residuals were used as a measure of students’ academic outcomes, consider a hypothetical student with average performance on the GCI final exam would have a standardized score of zero. Applying the GCII Final regression equation, the student would be predicted to score 0.150 of a standard deviation below the mean. Should the student’s actual standardized score on the final exam have been 0.712, their standardized residual for the final would be 0.862, indicating their performance was 0.862 of a standard deviation better than predicted. As a measure of whether a class performed above or below what was predicted via their incoming preparation (GCI score) on each test, the mean of standardized residuals for each semester and on each test constitute the outcome data used for this single-subject research design. Thus, the mean standardized residual (MSR) for the
performance of each instructor’s students was operationalized as an outcome measure of effectiveness.

**Problematizing Equity: Differential Standardized Residual (DSR)**

In addition to *effectiveness* measuring improvements in student performance, *equity* was also considered as an outcome measure in this evaluation of PLTL as an instructional reform. The National Research Council,\(^43\) reviewed changes in instructional reform and took note that while impacts to student learning were identified, the differential impact to students across groups of interest was rarely addressed stating that many of the studies reviewed by the committee were not designed to examine differences in students’ characteristics.

In chemistry, a strong relationship has been identified between students’ pre-college math test scores (e.g., SAT and ACT) and their academic performance in the course.\(^44-46\) So strong is this relationship that the majority of studies evaluating PLTL use student performance on the SAT as a means to control for differences in incoming preparation,\(^27,28,30,47,48,23,31,33\) but do not examine differences in the effectiveness of PLTL on students with variable performances on the SAT. This predictive relationship between pre-college math test scores and chemistry performance has been used to identify students scoring in the bottom quartile of a cohort’s math test scores as inequitably at-risk for achieving unfavorable outcomes in chemistry courses,\(^49-52\) for which students belonging to underrepresented minority groups (as described by the National Science Foundation) were overly represented.\(^50\) As was suggested by the NRC, this study was designed to investigate the impact of PLTL as a pedagogical reform on equity amongst college chemistry students differing in pre-college math test scores.

In past research concerning equity in scientific reform, equality (relating to fairness in that all students receive the same resources with the expectation that all students will then perform similarly) and equity (a system of trade-offs that provides students with a fair chance to succeed) have been differentiated.\(^53\) Much as methodological pluralism has been advanced as
the gold standard of educational research given the variety of perspectives and insight that can be achieved, so too have advancements in equity research called for the problematization of the measurement of equity to occur along multiple measures for greater insight as to the impact of scientific reform on differences between students of interest.\textsuperscript{53–56} The need for explicit operationalization when discussing equity was demonstrated by Rodriguez \textit{et al.} when data purported to reflect gender equity were reevaluated, revealing the persistence of inequity that went undetected as a result of its operationalization.\textsuperscript{55} This study seeks to advance the use of multiple, competing operationalizations of equity described as follows.

1. \textit{Equity of Individuality} – relating to advancements in performance amongst students of the group of interest following the reform.\textsuperscript{54–56}

2. \textit{Equity of Parity} – relating to a social justice perspective wherein the differences between majority and marginalized groups are reduced thereby benefiting the “less prepared” students more than the “well prepared” students.\textsuperscript{53,55}

The equity of individuality model is measured by the success for marginalized groups and does not consider differentials that may exist among groups, potentially perpetuating differential performance.\textsuperscript{55} While the equity of individuality model neglects differential performance, equity of parity (also referred to as equal outputs as described by Lynch) measures differential performance in an effort to seek meaningful ways to address the inequalities observed in student outcomes.\textsuperscript{53} The exclusive focus on differential gaps, however, potentially overlooks the extent each group has achieved success. For example, groups may perform equally; however, the overall success rate remains below expectations. Ideally, all students would advance following the enactment of an instructional reform (equity of individuality) to a degree in which there is no need for explorations of difference in student performance (equity of parity) as these differences would be non-existent. Adopting both perspectives of equity allows for investigations as to how close a reform is to achieving this ideal.
In the current study, student performance in GCI (as described for the MSR outcome variable) was controlled for differences in incoming preparation by their performance in GCI. In the context of equity of individuality, the MSR of at-risk students (those scoring in the bottom quartile of the cohort’s pre-college math test scores) belonging to instructors implementing didactic-instruction and PLTL will be compared. Should PLTL improve equity of individuality via improving performance amongst those at-risk for unfavorable academic outcomes, the MSR of at-risk students would be positive when PLTL is enacted.

To quantify equity of parity, differences in MSR between students at-risk and their peers in the not-at-risk group (those scoring in the top-three quartiles of pre-college math test scores) is operationalized as DSR or differential standardized residuals. DSR is thus a measure of differences in student performance attributed to precollege mathematics preparation. It is worth noting that the differences are not thought to reflect differences in incoming chemistry preparation given these differences were controlled using students’ performance on their first-semester General Chemistry (GCI) exams. Should students of the not-at-risk and at-risk groups perform more equitably following the implementation of PLTL, DSR would be negative, indicating a reduction in the gap, when PLTL is enacted.

**Interpreting Effect Sizes**

Cohen’s $d$ reflects the magnitude of change expressed in units of standard deviation and can be interpreted as small, medium, and large effects derived from parametric independent sample t-tests (see Equation 2).

$$d = \frac{\bar{X}_B - \bar{X}_A}{SD}$$

Equation 2. Cohen’s $d$ formula where $MD$ is the difference in mean outcome measures in the B (reformed) and A (reference) phases and $SD$ is the pooled standard deviation across phases.
The same measure can be used to quantify the effect of these differences indirectly within a single-subject research design. However, single-subject research designs generally produce considerably larger effect sizes, owing to the relatively small within-case standard deviations observed. While Cohen’s $d$ can provide an indirect estimate of effect; the measure should be interpreted tentatively for single-subject research designs and always in the context of the data under investigation. While these values are presented throughout the results, non-overlap indices were used to more directly measure the substantive impact observed between phases.

Non-Overlap Indices

For a more direct estimate of effect, non-overlap indices can be used to index effect size. A stable index for use where trends in the data could influence effect is the nonoverlap of all pairs (NAP). To calculate the NAP index, the following are first determined: the total number of pairs ($N$) of data points and the number of overlapping pairs ($n_O$) where the outcome metric of a data point is within the range of the other phase. The NAP index (see Equation 3) is calculated as the proportion of nonoverlapping pairs ($N - n_O$) from the total number of pairs ($N$), scaled from 0 to 1.

Equation 3. Nonoverlap of all pairs formula where the number of pairs ($N$) and the number of overlapping pairs ($n_O$) are used to calculate the index.

$$NAP = \frac{N - n_O}{N}$$

A NAP closer to 1 indicates fewer datapoints between phases that overlap and would support an inference of a substantive impact of the intervention. For example, consider the data shown in Figure 6.
Given 16 data points in phase B (blue circles) and 4 in phase (red squares), the total number of pairs is 64 \( (N = n_A \times n_B) \). Of these pairs, five overlap between phases (wherein the data values in phase B are less than or equal to data values in phase A). This results in a NAP of 0.922, which can be interpreted as 92.2% of phase B has no overlap with phase A indicating a considerable difference in the outcome metric for students receiving the experimental pedagogy.

When compared to Cohen’s \( d \), NAP subsumes comparisons throughout the full extent of the data provided (each pair of observations) rather than averages of observations within each phase.\(^{58}\) Field tests of 200 published AB comparisons commonly found in medical studies and tentative interpretations of NAP ranges (see Table 7) were provided by Parker and Vannest in 2009.\(^{58}\)

<table>
<thead>
<tr>
<th>NAP</th>
<th>EFFECT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 - 0.65</td>
<td>Small</td>
</tr>
<tr>
<td>0.66 - 0.92</td>
<td>Medium</td>
</tr>
<tr>
<td>0.93 - 1.0</td>
<td>Large</td>
</tr>
</tbody>
</table>

Figure 6. Example case for calculating nonoverlap of all pairs (NAP) – numbers over data points in phase A indicate overlap with phase B.
Thus, the data shown in Figure 6 present a NAP that can be interpreted as a medium-large effect size.

This index is helpful, particularly where the number of randomization distribution outcomes is less than 20 (where statistical significance cannot be detected), and comparisons between pedagogies can be evaluated only using visual analysis. Visual analysis (or descriptions concerning the overall pattern of the data) is common practice in single-subject research designs and involves an evaluation of baseline stability, variations within and between phases, trend or slope, level and overlap between adjacent phases (as described above), and comparisons of data across similar phases to determine if there are reasonable demonstrations of the effect. Resources describing the best-practices for enacting visual analyses,\textsuperscript{59–61} randomization tests,\textsuperscript{21,34,62,63} and evaluating the results of single-subject research designs,\textsuperscript{64–66} are referenced for the reader.

Results

Instructor-Level Differences in the Impact of PLTL on Overall Student Performance

The results of each instructor’s responses to the survey items in Box 1 and their performance feedback will be presented for each case (instructor) in order of increasing design complexity (e.g., AB/BA, ABA/BAB, and ABAB/BABA).

Professor Parker (BA)

In response to the faculty survey, Professor Parker indicates teaching second-semester General Chemistry (GCII) for over ten semesters. Parker’s teaching is informed by workshops in alternative pedagogies (e.g., process-oriented guided inquiry learning or POGIL and PLTL), reading literature on learning cycles and enacting PLTL, and reflecting on past teaching experiences. Parker indicates general effectiveness when using either didactic or PLTL
instruction, but also describes a belief of improved effectiveness when implementing PLTL.
Parker supports this improvement in effectiveness by stating reservations as to whether lectures “promote the skills necessary for students to succeed in the class” and values the trainings peer leaders receive. Parker communicates the influence of PLTL experience on instruction within didactic courses.

“My lectures, in particular working problems but also bringing up conceptual explanations, have become a lot more segmented. By that, I mean I stop and explain each decision rationale at more frequent intervals. I also have looked for more opportunities to engage students during lecture via clickers.”
Parker wanted to teach using PLTL given the effectiveness demonstrated in other studies and finds challenges in enacting PLTL around selecting the most pertinent content for students to engage with during problem-solving.

During the study, Prof. Parker taught five semesters of GCII and enacted a BA design wherein PLTL instruction was conducted first, establishing a baseline, followed by the enactment of didactic instruction in semester 7. Parker’s baseline was comprised of students’ performance under PLTL instruction wherein students, on average, were observed to perform 0.043 standard deviations above what was predicted based on final exam scores from GCII (see Figure 7 on the next page). In semester 7, Parker enacted didactic instruction for which student performance was markedly lower ($M = -0.174$). For Parker, students with PLTL did 0.217 MSR (difference between 0.043 and −0.174) better than students with didactic instruction. The difference in student performance was considerable, with little overlap in the data between phases ($NAP = 0.922$, a medium to large effect).
To evaluate the significance of these differences using randomization tests, the number of randomization distribution outcomes had to be determined. Given the 20 observations collected (5 semesters, each with 4 observations) and one phase change (B to A), 13 potential starting points for the permutation were identified with 2 possible arrangements (AB or BA). Thus, the probability of obtaining a test statistic as large or larger than $|\bar{X}_{exp} - \bar{X}_{ref}|$ for Parker’s observational data was $p = \frac{x}{26}$. One case met this criterion, and that case was the test statistic itself, resulting in significant differences detected between instructional pedagogies favoring PLTL ($p = 0.038$, $d = 2.09$). Thus, for Parker, the enactment of PLTL resulted in a favorable and significant impact on student outcomes contributing to performances that exceeded what was predicted based on their GCI final exam scores.

Professor Davis (AB)

Having taught 6 semesters of GCII, Professor Davis’s teaching was informed by past experiences as a student, teaching assistant, mentor and professor. Davis reports experiences with metacognitive learning strategies as pivotal in graduate coursework and self-identifies as an effective instructor regardless of pedagogy but more effective so when implementing PLTL.
In lecture-based instruction, Davis works hard to engage students “including real-life examples, working problems out step-by-step, engaging the audience by asking questions and answering questions, and incorporating technology”.

Davis describes challenges in implementing PLTL related to engaging students who prefer to work individually and balancing time among the different groups of students during peer-leading sessions. PLTL has influenced Davis’s teaching in lecture-based pedagogical implementations as follows.

“Participating in peer-led team learning has given me insight into the ineffective approaches I used in the past. I can see now that increasing the amount of examples that are worked out step-by-step during lecture doesn't necessarily increase students' understanding of the material. It is better for them to work examples themselves instead of watching me work them multiple times.”

Davis wanted to teach using PLTL to get involved with an approach that others described as beneficial. Having participated in the implementation of PLTL, Davis valued interacting with students and guiding the learning process.

Prof. Davis enacted an AB design across 4 semesters of assessment performance data. On average, Davis’s students with PLTL did 0.157 MSR better than students with didactic instruction (see Figure 8 on the next page). Performance outcomes of students improved from semester 5 ($M = -0.182$) to semester 6 ($M = 0.164$), both semesters under didactic instruction, suggesting a potential trend in improvement unrelated to changes in instructional pedagogy. A NAP of 0.771 was calculated (medium effect) suggesting the change in pedagogy substantively improved student performance.
Given one phase change from A to B and 16 observations, the number of randomization distribution outcomes was less than 20 and could not reach the critical value of 0.05. Thus, statistical significance could not be tested with the data available.

In summary, both cases of AB or BA design suggest students (on average) performed more favorably when either instructor implemented a PLTL pedagogy. Two medium effects were demonstrated favoring the enactment of PLTL for these two instructors. However, the interpretations of these data could be limited by potential competing explanations owed to the AB design enacted such as the upward trend over time identified in Davis’s results.

**Professor Morales (AB and ABA)**

Professor Morales’s response to the faculty survey indicates experience in teaching GCII that exceeds 10 semesters. Further, Morales’s teaching is informed by past academic work and training in science education and communication. Morales identifies as an effective instructor and perceives an increase in effectiveness when enacting PLTL describing these improvements as driven by an improvement in student understanding of concepts. A challenge Morales communicates regardless of instructional pedagogy is the need to adapt teaching to changes in

---

Figure 8. Graphic display of Professor Davis' performance feedback from semester 3 to 7.
Morales describes the challenges of adopting PLTL as adjusting to the reduction in the time available for exposition, students’ carelessness concerning attendance and overcoming a few students’ preference for receiving instruction more passively. Morales motivation to teach using peer learning was described as:

“I like talking with students about chemistry and showing them that it is not so hard as many believe to comprehend. I like showing them how it applies to their lives and making the connections to what they already know.”

The influence of PLTL on didactic instruction was described by Morales as allowing for the removal of superfluous material and an increased understanding of “what students actually understand vs. what I think they understand because they smiled and nodded at me during lecture”.

Professor Morales had a unique case wherein the instructor enacted both pedagogies within the same semester in differing classes of the same course. This facilitated comparisons of both an AB design and a reversal design (ABA) within the same case (see Figure 9 on the next page). The baseline was conducted using didactic instruction where students were observed to perform 0.132 standard deviations below predicted. For the AB design (didactic in semester 3 followed by semesters of PLTL shown at the top of Figure 9), improvements in student performance were detected ($M_{PLTL} = +0.033$) with a NAP of 0.906 (medium effect).

Given 13 potential starting points for the intervention, the probability of obtaining a test statistic as large or larger than $|\bar{X}_{exp} - \bar{X}_{ref}|$ for Morales’ observational data was $p = x/26$. One case met this criterion and statistically significant differences between phases were detected ($p = 0.038, d = 1.77$). However, there may be a trend in student improvement as semester 3 (baseline) progresses.
Figure 9. Graphic display of Professor Morales’s performance feedback from semester 2 to 7 for an AB (upper section) and ABA (lower section) design.

Whether these improvements are the result of experience of the pedagogy enacted can be better argued via the consideration of the data comprising the reversal design (ABA shown at the bottom of Figure 9). Here the trend over time may be inconsequential as student performance returned back to baseline following the removal of PLTL. Within 20 observations, 4 observations at minimum per phase (three tests and a final exam), and 2 changes in phase (from A₁ to B and B to A₂), the number of randomization distribution outcomes was 45. Given two possible arrangements (ABA or BAB), the probability is equal to x/90. One case (the test
statistic) met these criteria and a statistically significant difference ($p = 0.011, d = 1.17$) was detected. This suggests that, for Prof. Morales, significant differences in student outcomes were detected upon the enactment of PLTL.

**Professor Stacy (ABAB or BABA)**

Professor Stacy, whose teaching is informed by past experiences as a teaching assistant, has 7 semesters of experience in teaching GCII. Stacy identifies as an effective instructor regardless of instructional pedagogy as supported by students’ performance on assessments but is more effective when implementing PLTL. When comparing effectiveness between pedagogies, Stacy describes an increase in the use of clicker questions to allow students receiving didactic instruction to practice and a concern that “some of them might skip processing the questions and just choose A, B, C, D, or E”. Stacy describes PLTL’s benefits as:

> “It gives students an opportunity to practice step-by-step during class time with the guidance of well-trained peer leaders. Gen Chem 2 has a lot of math and concepts. Students really need to understand the concept and then process the problem. Practicing in groups with students while the leaders are present really helps them digest the materials. Also, the best way to learn is to teach! While explaining to each other, they also digest the materials.”

Stacy shares the perspectives of Profs. Parker and Morales in that PLTL has influenced enactment of didactic instruction by elucidating the topics on which students need more time and the mistakes students may commonly make when engaging with these topics. When asked of the challenges experienced enacting PLTL, Stacy describes concern over the students who may prefer to work by themselves rather than within groups.

In a final case, Prof. Stacy demonstrates that, for some instructors, which instructional pedagogy is selected may not play an instrumental role in students’ performance. Stacy enacts
the experimental instructional pedagogy first using an alternating treatment (BABA) design (see Figure 10).

![Figure 10. Graphic display of Professor Stacy’s performance feedback from semester 1 to 7.](image)

Phase averages indicate that Stacy’s students performed similarly under didactic instruction ($M = -0.030$) and with PLTL ($M = +0.015$); NAP was calculated across all three changes in phase and indicate no detectable effects for the change in instructional pedagogy. Given 24 observations, a minimum of 4 observations per phase, 3 changes in phase ($A_1$ to $B_1$, $B_1$ to $A_2$, and $A_2$ to $B_2$), and 2 possible arrangements (ABAB or BABA), the number of outcomes ($O$) was equal to 330. Many cases (77) met these criteria resulting in no detectible significant differences in student performance given either pedagogy ($p = 0.233$, $d = 0.397$).

**Summary of PLTL Impact on Student Outcomes**

In summary, student performance (on average) improved for all cases, significantly so for the students of Prof. Parker, Davis, and Morales (see Table 8).
Table 8. A summary of mean standardized residuals (MSR) as outcome measures used to evaluate the reformed (PLTL) instructional pedagogy and its impact relative to the reference (didactic) approach.

<table>
<thead>
<tr>
<th>Case</th>
<th>MSR&lt;sub&gt;PLTL&lt;/sub&gt;</th>
<th>MSR&lt;sub&gt;didactic&lt;/sub&gt;</th>
<th>p-value</th>
<th>Cohen’s d</th>
<th>NAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parker (BA)</td>
<td>0.043</td>
<td>-0.174</td>
<td>0.038</td>
<td>2.09</td>
<td>0.922</td>
</tr>
<tr>
<td>Davis (AB)</td>
<td>0.118</td>
<td>-0.039</td>
<td></td>
<td>1.30</td>
<td>0.771</td>
</tr>
<tr>
<td>Morales (AB)</td>
<td>0.033</td>
<td>-0.132</td>
<td>0.038</td>
<td>1.77</td>
<td>0.906</td>
</tr>
<tr>
<td>Morales (ABA)</td>
<td>0.040</td>
<td>-0.060</td>
<td>0.011</td>
<td>1.17</td>
<td>0.875</td>
</tr>
<tr>
<td>Stacy (BABA)</td>
<td>0.015</td>
<td>-0.030</td>
<td>0.233</td>
<td>0.233</td>
<td>0.292</td>
</tr>
</tbody>
</table>

* Insufficient data were available to calculate statistical significance.

The effect of PLTL on overall student performance ranged from no detectable difference (NAP = 0.125) to substantively significant differences of medium effect (NAP = 0.922). For Prof. Stacy, either instructional pedagogy resulted in comparable student outcomes. However, for Prof. Parker, the difference in student performance was clearly meaningful. This spectrum in the effectiveness of PLTL indicates instructor-level differences had an impact on the effectiveness of either instructional pedagogy, with PLTL’s effectiveness ranging from a meaningful positive effect to no measurable effect.

Instructor-Level Differences also Moderate the Impact of PLTL on Equity

*Equity of Individuality*

The impact on equity of individuality following the enactment of an instructional reform was operationalized as improvements to the performance of students scoring in the bottom-quartile of math test scores and at-risk (AR) for achieving unfavorable outcomes in the course. In all cases except one (Morales with the ABA design), at-risk students in second-semester General Chemistry performed above what was predicted by their scores on their first-semester General Chemistry exams following the enactment of PLTL (see Figure 11).
Figure 11. Bar graphs depicting the average MSRs (mean-centered by each case or instructor) for students of the not-at-risk (NAR) and at-risk (AR) group receiving didactic-instruction and peer-led team learning (PLTL).

For those students at-risk, regardless of instructor, performance was less favorable following the receipt of didactic instruction. While improvement was observed across all cases, the degree to which improvements in at-risk student performance were observed (calculated as the difference between at-risk students receiving PLTL and didactic) ranged from $+0.079$ to $+0.269$ indicating a considerable range between instructors. The positive values of these differences are indicative that at-risk students with PLTL enacted performed better versus at-risk students with didactic instruction enacted. Changes in not-at-risk student (NAR) performance ranged from $-0.020$ to $0.225$ reflecting the impact of instructor-level differences on both student performance and equity as defined by individuality.

**Equity of Parity**

To quantify changes in the performance gaps observed between at-risk students and their peers, DSR (of differences in MSR between student groups) were calculated, mean-centered by instructor, and evaluated using randomization tests. As DSRs were centered for
each instructor, a positive DSR represents a larger differential (inequity) in student performance with the pedagogy in place whereas a negative DSR indicates a reduction in differential (greater equity) in student outcomes. For all but one case (Davis), DSRs for PLTL were lower (more equity) than that of didactic (see Table 9).

Table 9. A summary of differential student performance (DSR) as outcome measures used to evaluate the reformed (PLTL) instructional pedagogy and its impact relative to the reference (didactic) approach.

<table>
<thead>
<tr>
<th>Case</th>
<th>DSR&lt;sub&gt;PLTL&lt;/sub&gt;</th>
<th>DSR&lt;sub&gt;Didactic&lt;/sub&gt;</th>
<th>p-value</th>
<th>Cohen's d</th>
<th>NAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parker (BA)</td>
<td>-0.004</td>
<td>0.017</td>
<td>0.462</td>
<td>-0.116</td>
<td>0.469</td>
</tr>
<tr>
<td>Davis (AB)</td>
<td>0.007</td>
<td>-0.002</td>
<td>a</td>
<td>0.054</td>
<td>0.458</td>
</tr>
<tr>
<td>Morales (AB)</td>
<td>-0.040</td>
<td>0.161</td>
<td>0.038</td>
<td>-0.969</td>
<td>0.656</td>
</tr>
<tr>
<td>Morales (ABA)</td>
<td>-0.031</td>
<td>0.046</td>
<td>0.322</td>
<td>-0.278</td>
<td>0.313</td>
</tr>
<tr>
<td>Stacy (BABA)</td>
<td>-0.069</td>
<td>0.138</td>
<td>0.009</td>
<td>-1.173</td>
<td>0.813</td>
</tr>
</tbody>
</table>

*Insufficient data were available to calculate statistical significance.

This suggests that, at best, the pedagogy could support reductions in the differential performances observed of students as a result of differences in precollege math test scores and, at worst, no deleterious effects are observed following the enactment of PLTL.

As calculated using randomization tests, significant reductions of differential performance between students of the not-at-risk and at-risk student groups following the enactment of PLTL were detected in 2 of the 5 cases evaluated (Stacy and Morales’s AB design), each at medium effect sizes. For Morales’s AB case, student performance increased, and differential performance decreased significantly. For Parker, Davis, and Morales’s ABA case, PLTL significantly improved student performance and showed a reduction in the differential performance, though not significantly. For Stacy, the lone case where no difference in student performance was detected, a significant decrease in differential performance was observed.
The Impact of PLTL on Equity

Informing equity from both perspectives (individuality and parity), changes in student performance for both the NAR and AR student groups following the enactment of PLTL were positive (indicating improvement) but were also similar enough across the cases of Parker, Davis, and Morales’s ABA design) to have no significant impact on the differential between groups. However, differentials were reduced significantly amongst students of Morales AB design (wherein AR student performance nearly doubled in MSR) and Stacy (wherein AR student performance increased considerably with a slight decrease in the performance of NAR students). Overall, PLTL could be described as a more equitable instructional practice than didactic instruction wherein further understanding as to why this change in pedagogy positively impacts the performance of at-risk students could drive the advancement of equity in chemistry courses.

Discussion

The intent of the work was to introduce randomization tests and to demonstrate how this experimental, randomized design could advance efforts to evaluate PLTL from a perspective of methodological pluralism considered the gold standard of STEM education evaluation. In 2006, Lawrenz and Huffman unpacked this shift in the gold standard of STEM education evaluation as follows.

“Appreciation and use of a variety of techniques is a reflection of the STEM disciplines themselves and permits comprehensive insight into STEM education evaluation.” (p. 30)

The introduction of randomization tests also serves to offer an alternative mechanism for conducting evaluation work that is particularly well suited for research designs that focus on
smaller sample sizes or when a comparison group is not readily available, which are particularly common among upper-level STEM courses.

The results of this study indicate a spectrum of differences among instructors in the effectiveness of PLTL toward improving student performance and ameliorating performance gaps between chemistry students of variable preparations in mathematics (see Tables 5 and 6). As such, differences in effect of an instructional reform on the instructor-level are meaningful for evaluations of PLTL and demonstrate a contribution to which single-case subject designs add to the knowledge base from past works evaluating the instructional reform.

An Evaluation of Peer-Led Team Learning in the Context of Prior Literature

In the context of prior literature, measures of student performance and instructor-level differences have been considered using a variety of methods (see Table 10 on the following page). The majority of the studies in Table 10 used success rates (the percent of students receiving passing grades in the course) to measures differences in student outcomes and evaluate the success of PLTL as an instructional reform.\textsuperscript{23,27–30,32,33} While each define a variety of grading schemes, all present an increase in the percentage of students who passed the course following the implementation of PLTL ranging from 9-25%. Exam performance was also used to measure student success following the enactment of PLTL.\textsuperscript{23,29,33,67} Of these four works, no statistically significant differences in student performance on exams were detected. This suggests that improvements in student outcomes following PLTL are either attributable to marks unrelated to exams or were not detectable on students’ final exams.
Table 10. A summary of the methods used in studies evaluating the impact of PLTL in science courses.

<table>
<thead>
<tr>
<th>Authors</th>
<th>Sub- or discipline</th>
<th>Instructor-Level Considerations</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bramaje and Espinosa, 2013</td>
<td>High School Chemistry</td>
<td>Common Instructor</td>
<td>Relative to didactic instruction, student performance on a concept inventory improved by $d = 0.45$ (medium effect)</td>
</tr>
<tr>
<td>Lewis, 2011</td>
<td>General Chemistry</td>
<td>Each Class an Observation, Common Instructors Similar Time-on-Task</td>
<td>10% increase in success rates (% ABC), comparable outcomes on ACS final exams</td>
</tr>
<tr>
<td>Shields et al., 2012</td>
<td>General Chemistry</td>
<td>Common Instructors</td>
<td>Significant improvement with PLTL amongst students diagnosed as underprepared ($d = 1.22$, a large effect)</td>
</tr>
<tr>
<td>Hockings et al., 2008</td>
<td>General Chemistry</td>
<td>Not Addressed</td>
<td>9% increase in success rate, improvement in student performance by ⅔ of a grade point</td>
</tr>
<tr>
<td>Mitchell et al., 2012</td>
<td>General Chemistry</td>
<td>Each Class an Observation</td>
<td>13% increase in success rate, comparable outcomes on ACS, treatment effects lost after return to didactic instruction</td>
</tr>
<tr>
<td>Chan and Bauer, 2015</td>
<td>General Chemistry</td>
<td>Common Instructor, Similar Time-on-Task</td>
<td>Comparable outcomes on exams</td>
</tr>
<tr>
<td>Tien et al. 2002; Lyle and Robinson 2003</td>
<td>Organic Chemistry</td>
<td>Common Instructor</td>
<td>11% increase in success rate, significant improvement in student performance ($d = 0.64$), significant gains in success rates for females (15%) and students who identify as an URM (11%)</td>
</tr>
<tr>
<td>Wamser, 2006</td>
<td>Organic Chemistry</td>
<td>Common Instructors</td>
<td>16% increase in success rate, 6% increase in student performance on ACS exams</td>
</tr>
<tr>
<td>Snyder et al., 2016</td>
<td>Biology</td>
<td>Not Addressed</td>
<td>25% increase in success rate, reduction in differential success rates between students who identify as an URM and their peers</td>
</tr>
<tr>
<td>Aldridge, 2011</td>
<td>Biology</td>
<td>Common Instructors</td>
<td>Statistically significant improvement in final exam performance ($d = 0.73$, large effect) amongst students in PLTL</td>
</tr>
<tr>
<td>Akinyele, 2010</td>
<td>General, Organic, and Biological Chemistry</td>
<td>Common Instructor, Similar Time-on-Task</td>
<td>15% increase in success rate, significant improvement in students’ overall course performance ($d_{AVG} = 0.71$, over 6 semesters)</td>
</tr>
</tbody>
</table>

The current study is distinct from these prior works as the impact of PLTL is considered for each instructor only within the context of past student performance with didactic instruction from the
same instructor. The collected data indicate increases in student performance (on average) across all four instructors where three cases show practical impact with medium effect sizes (as measured by NAP) and two cases that were statistically significant.

Few studies in the past considered the impact of PLTL on equity.\textsuperscript{27,32} These works operationalized equity as either improvements in success rates for underrepresented minorities or URMs,\textsuperscript{27} or a reduction in the differences in success rates,\textsuperscript{32} between students who identify as URMs and their peers. Given that at-risk chemistry students (or those scoring in the bottom-quartile of precollege math test scores) are disproportionately comprised of students who identify as an URM,\textsuperscript{50} and the reduction in differential performance (on average) observed across four of the five cases evaluated (see Table 9), the findings of the current study suggest that PLTL may effectively reduce the differential performances observed in these past works. The current study distinguishes itself from these past works by measuring equity on exams administered throughout the semester rather than students’ success rates as communicated in prior works. This approach to measuring equity supports that these improvements in equity are not along a single measure but can be observed throughout the semester and across cohorts of students. Future research should consider including a measure of reductions in differential performance as a means for evaluating the success of instructional reforms.

Methods Addressing Instructor-Level Differences

While measures of student outcomes and equity are valuable to the body of work seeking evaluations of instructional reform, so too is the contribution of the current work suggesting instructor-level differences have a meaningful impact on the effect of changes in pedagogy. The selected studies address instructor-level differences using a variety of designs including those that are naturalistic, collectivistic, and individualistic (see Figure 12).
Naturalistic designs posit the data is representative of a naturalistic educational setting wherein differences between instructors are inherent and thereby are not directly considered. For example, the study by Mitchell et al. (2012) investigates differences in student retention from GCI through GCII amongst students who received either PLTL or didactic instruction. Among instructors the methods of instruction, in-class exams, and assignments were not controlled. As such, the data of this and other works that were not designed to account for instructor-level differences are reflective of a naturalistic implementation of PLTL across instructors who have a variety of teaching styles and experience with the pedagogical reform.\(^\text{30,32,33}\)

Collectivist designs feature instructors who toggle back and forth between PLTL and didactic instruction. For example, Wamser in 2006 used a collectivist design to evaluate PLTL wherein two different instructors, differing in experience with enacting the reference and reformed instructional pedagogy at the research setting, alternated between teaching sections.

---

**Figure 12.** An illustration representative of the various methodological approaches toward addressing instructor-level differences.
with and without PLTL over five years. As instructors toggled between teaching the course using PLTL and not, there were no common years wherein both instructors taught the same group of students using the same pedagogy. Thus, the results emphasize collective differences between students who self-selected to participate in PLTL and those who did not with either instructor rather than comparing for differences between instructor. In Lewis’s 2011 study, instructor-level differences were treated as both naturalistic and collectivistic.

Pass rates were compared for student groups receiving either the referenced or reformed instructional pedagogy from a common set of instructors (collectivistic) and across instructors overall (naturalistic). When implementing PLTL, this common set of instructors observed a 10% increase in retention. Overall, improvements in success rate were 15%. Although both changes in pass rate (when compared to didactic instruction) were significantly different with large effect sizes, these findings suggest instructor-level differences may have had an impact on retention. This impact on retention is explained by Lewis as potentially relating to instruction decisions (e.g., awarding extra credit, accepting late work, and curving of test scores) but was not reflective of differences in time-on-task which, distinctive from the previous studies, were controlled.

Individualistic designs were also observed wherein the data compares student groups receiving either the reference or reformed instructional pedagogy from a single instructor. This design reflects a trade-off by eliminating between-instructor differences (via the consideration of a single educator) at the cost of broader generalizability. Chan and Bauer’s study, reflects data collected amongst students of an experienced instructor enacting both pedagogies at a setting wherein PLTL has been incorporated continuously and students’ time-on-task were measurably comparable. Sharing a common dataset, Lyle and Robinson, and Tien et al. also review data collected from a single instructor who was described as consistent, well-trained, and involved in the training and facilitation of PLTL. Similarly, the current study involves
examinations of student performance on the instructor-level (individualistic) and across multiple instructors (collectivistic).

The current study complements other designs in that instructor-level data were explicitly evaluated and different effects were observed among instructors. In doing so, the results suggest how effective a pedagogy can be for a group of students is strongly swayed by these instructor-level differences. Amongst the cases presented effect sizes favoring the implementation of PLTL for Profs. Parker, Morales, Davis, and Stacy varied widely (2.09, 1.77, 1.30 and 0.233, respectively). This reflects the variability observed in Lewis’s evaluation of PLTL\textsuperscript{23} when comparing the data naturalistically and collectively and could offer a potential explanation as to why no differences in student performance or retention were observed for the study by Chan and Bauer.\textsuperscript{67}

Further, these data present a novel contribution in the evaluations of within-instructor trends across semesters. Both Davis and Morales present cases wherein a trend was suspected given increases in student performance over time. For Davis, student achievement improved along two semesters of didactic instruction prior to the semester PLTL was implemented. In Morales’s case, there was a positive trajectory for student achievement observed during the semester wherein didactic instruction was administered. In both cases, familiarity with students, the topics, the teaching sequence, and the pedagogy could have contributed to increases in student achievement over time. Further, cases wherein reversals were enacted (Morales, ABA; Stacy, BABA) present trends that suggest positive impacts of PLTL can linger into follow-on semesters taught using didactic instruction. These trends in effectiveness over time within instructors add nuance to the interpretations of data and speak towards evaluations of consistency that do not assume instructors remain constant in their effectiveness across a semester or as their experiences with an instructional pedagogy increase.
Limitations

As an introduction to the methodology of randomization tests within single-subject research designs, the data presented herein presents some considerations worth further exploration. For example, students’ incoming preparation was assumed mitigated by their first-semester General Chemistry performance, a variable that may not encapsulate other factors that contribute to differences amongst students such as motivation, problem-solving ability, and language comprehension. While controlling for first-semester General Chemistry performance is meant to address differences in incoming preparation, it may not account for all differences among the historical groups of students represented in the data.

While naturalistic, the assignment of PLTL and didactic instruction was not truly random as the intent for the pedagogical reform was to foster multiple faculty gaining experience with the reform to promote the sustainability of the reform. Additionally, assessments were not common across semesters and may have other differences in measure that cannot be accounted for via standardization. Finally, baselines for the instructors were markedly unstable (or presented with considerable reference phase instability) and may suggest that instructor experience at the setting or the differential benefits of PLTL by topics (e.g., PLTL may aid more with a sub-set of topics) could inform these differences in student performance. Given the demonstration that positive changes in student achievement in three of the four cases presented, and a marginal positive impact for the fourth case, coincided with the enactment of PLTL, it is arguably less plausible that differences in incoming student characteristics coincided with the enactment of PLTL, and more plausible that the enactment of PLTL was responsible for the differences observed. Owing to this judgment, the authors’ interpretation of the data generated in this evaluation remains endorsing the use of PLTL as an effective instructional pedagogy.
Implications for Research and Practice

Ultimately, instructor-level differences are a meaningful consideration in the evaluation and generalizability of effects following instructional reforms. As institutions and communities of education researchers continue in the evaluation of instructional reform, these instructor-level differences warrant attention as do the nuances amongst instructors’ implementations of reforms that could positively or negatively contribute to student success, retention, and equity. Randomization tests could be a methodology by which instructors evaluate the pedagogical value of changes in their classrooms without limitations concerning sample size and the ethical considerations when selecting a control group. Further, these findings suggest the value in evaluating instructional reforms over periods of time capable of measuring change with their familiarity in enacting the instructional reform and cautions against assumptions made that one reform could serve as a cure-all when applied to educational settings.

As posited by Michael, faculty development is a critical component of evaluating instructional reforms:

“One of the critical issues is faculty development, helping teachers to become familiar with new approaches to teaching and helping them gain experience actually implementing them.” (p. 164)

Faculty development and differences between instructors could provide other explanations for observed student differences when evaluating instructional reforms. Observed differences in student outcomes could also be related to differences in the instructors’ implementation of or experience with the reformed pedagogy, differences in retention rates, or more practical differences such as the times of day courses are administered, classroom resources, and differences in time-on-task. As such, a variety of research designs can be used to assess the weight of these differences and paint a more complete picture of a reform’s effectiveness.
Future Works

Randomization tests poses a method that can contribute to research designs relying on a statistical assumption of independence that may be inappropriate for education settings but can be further contextualized using this alternative approach. As such, future works evaluating changes in pedagogical reform could be advanced via studies applying this methodology and offer novel contributions including a consideration of instructor-level differences and how these differences interact with student performance over extended periods of time. Additionally, the approach is not dependent on large sample sizes and can be conducted with a single student making this methodology applicable to upper-level courses, students of rare-incidence populations, and smaller institutions.

Given the limitations of this introduction, future works seeking to enact an evaluation of pedagogical changes enacted at random periods of time with a common group of students serving as their own control would provide an application of randomization tests more authentic to the intent of the design. As equity, instructor-level differences in implementation, and methodologically pluralistic evaluations of evidence-based practices are rare, a reconceptualization of what constitutes an evidence-base could considerably advance conversations concerning the evaluation of education reforms.

Conclusion

Randomization tests, a quantitative method for evaluating changes in an outcome measure following the random assignment of an intervention, were used to measure student achievement across four instructors at the research setting when enacting PLTL. The trends across the cases demonstrated a positive impact on overall student performance and reductions in the differential performances of not-at-risk and at-risk chemistry students. For three of the four cases for which statistical significance could be calculated, the enactment of PLTL had a significant impact on student outcomes. PLTL also positively improved at-risk student
performance in all cases with varying degrees of success wherein two of the four cases presented a statistically significant reduction in the differential performance observed of chemistry students with variable preparation in mathematics. In addition to trends and lingering effects of reversals on follow-on data, these findings suggest instructor-level differences could meaningfully impact student performance and the equity (both via individuality and parity) of a course. Further advances in methodological pluralism, such as the inclusion of single-subject research designs and other novel designs for evaluations of education reforms, could prove effective complements towards works that seek to establish evidence-based practices.

References


(6) Berry, L. Collaborative Learning: A Program for Improving the Retention of Minority Students; 1991.


Ginsburg, A.; Smith, M. S. Do Randomized Controlled Trials Meet the “Gold Standard.” American Enterprise Institute March 18, 2016.


(48) Aldridge, J. N. FROM ACCESS TO SUCCESS IN SCIENCE: AN ACADEMIC-STUDENT AFFAIRS INTERVENTION FOR UNDERGRADUATE FRESHMEN BIOLOGY STUDENTS. (Doctor of Education), University of Delaware, Newark, Delaware, 2011.


CHAPTER 4:
CHEMISTRY TOPICS POSING INCOMMENSURATE DIFFICULTY TO STUDENTS WITH LOW MATH APTITUDE SCORES

Note to Reader

Portions of this chapters have been previously published in Chemistry Education Research and Practice, 19: 867-884. These works have been reproduced for this dissertation with permission from the Royal Society of Chemistry (see Appendix B for Copyright Permissions).

Introduction

Identifying At-Risk Students in Chemistry

The early identification of students at risk of poor academic performance (e.g., performance on examinations, course completion, or final grades) in chemistry has encompassed the vast majority of published work on this population of students since the 1970’s.1–3 The list of predictors has been used to identify at-risk students in chemistry includes a variety of demographic, affective and cognitive traits. Considered demographic predictors include students’ gender,4,5 age,6 years in college,5 and race or ethnicity.5 Adaptive, affect-related qualities were also identified as predictors including self-efficacy,7,8 chemistry competency beliefs,8 study habits,9 chosen undergraduate major,5 test anxiety, emotional satisfaction,7 and self-reflective or metacognitive accuracy.10

Measures of cognitive traits via standardized pre-admissions testing such as the SAT and, to a lesser extent, the American College Test (ACT) have been historically well-referenced and are considered a reasonably strong predictor for at-risk chemistry students.4–6,9,11–17 Other
considerations included pretest measures of prior knowledge in math and/or chemistry,\textsuperscript{6,18} and measures of cognitive ability such as formal thought,\textsuperscript{15} and spatial ability.\textsuperscript{19}

The math sub-section of the SAT, herein referred to as SATM, features a consistent relationship with chemistry performance across a wide literature base. The current structure for the questions of the verbal and mathematical components of the SAT was established in 1952 sparking interest in the test’s ability to predict a variety of college student outcomes. Three major foci of SATM were identified by the College Board as the mastery of linear equations and systems, quantitative literacy via problem-solving and data analysis, and the manipulation of complex questions.\textsuperscript{20} In the 20-30 years following its establishment, students’ scores on SATM were correlated with introductory chemistry course grades (transformed into a numerical scale) with values ranging from correlations of 0.42 to 0.51.\textsuperscript{13,14,21}

To address the numbers of students achieving various grade levels differing from that predicted by SATM, studies began to converge on the idea that while various ranges of SATM scores (scored between two- to eight-hundred points possible on the assessment) are represented by students who ultimately received a wide variety of high and low grades in chemistry courses, students scoring within the lower percentiles of the SATM were much more unlikely to achieve the high grades amongst their peers.\textsuperscript{4} In 1979, Andrew & Andrews concluded that students scoring high on the SATM are not guaranteed to pass, yet students with low scores are strongly predicted to perform poorly in chemistry. Spencer,\textsuperscript{5} explored demographic and other background variables, namely years in college, ethnicity, gender, and major, for the extent these variables interact with the relationship of SATM and academic performance in chemistry. Insufficient evidence of interactions was found, except students who had declared chemistry majors over-performed the expected outcome based on SATM, leading to the finding that the relationship of SATM with chemistry performance was consistent across these measures.
As SATM is well-established to serve as an identifier for students at academic risk in first-semester chemistry courses, the next logical distinction to be made is where along the range of SATM scores should the cutoff for at-risk students exist? Lewis and Lewis,\textsuperscript{15} used a regression of SATM and SATV (verbal sub-section) related to a cumulative final exam to identify combinations of student scores on SATM and SAT-V that were predicted to score in the bottom 30\% on the cumulative exam and found that scoring below 500 on both sub-scores, or other combinations where a higher score on one measure could be offset by a lower score on the other, identified a cohort of students who scored below the threshold 70.5\% of the time. More recent studies report accuracy in identifying students at risk of low academic performance in college chemistry as the bottom quartile in SATM of each semester’s cohort.\textsuperscript{9,16,17} In this study, students scoring in the bottom quartile of their semester’s cohort will be classified as at-risk. Combined, these studies present a compelling and consistent argument for the use of SATM scores to predict performance in first-semester chemistry. None of the studies address which aspects of a student’s cognitive ability, specifically those essential for success in college chemistry, are predicted by the SATM. With regard to how the SATM predicts success in chemistry courses, the standing assumption has been that students entering college with lower SATM perform poorly in introductory chemistry owing primarily to a lack of quantitative skills.

**Assisting At-Risk Students in Chemistry**

The research literature on identifying at-risk students or relating cognitive, affective or behavioral measures to student performance in chemistry often includes prescriptive suggestions for assisting at-risk students. A common suggestion is the assignment of a prerequisite course or remedial course in mathematics or chemistry.\textsuperscript{11,12,18,22} Additionally, the evaluation of cohorts following prerequisite courses in mathematics may be problematic as they create alternative cohorts of students eligible to enroll in the chemistry coursework that follows. In-class suggestions include adopting more frequent and higher quality study habits,\textsuperscript{9} or self-
regulation to promote intellectual engagement with coursework.\textsuperscript{10} The in-class suggestions are resultant from the findings within the literature but were not explicitly evaluated within the same literature.

The research literature also includes examples of evaluating interventions designed to aid at-risk students. Mason and Verdel,\textsuperscript{23} identified at-risk students as those voluntarily participating in university-sponsored retention improvement programs that sought to aid minority students, first-generation students and student-athletes. A group of 36 at-risk students was divided with 17 students attending a traditional large lecture and 19 students attending a small lecture class with attendance restricted to their enrollment. Both groups received lecture-based instruction and dedicated time for individual or group work. The results showed that the students in the large lecture class outperformed the students in the small lecture class though there was insufficient evidence to make a case of statistical significance.

Another example is an evaluation of the Science Advancement through Group Engagement program.\textsuperscript{17} The program was available to students scoring in the bottom quartile of the SAT\textsubscript{M} and included extracurricular work with learning specialists, teaching assistants, and class-related group-work over four semesters including two semesters of introductory chemistry and two of organic chemistry. The evaluation found that among at-risk students enrolled in the program, 68% completed their coursework through organic chemistry as compared to those not involved with the program (27%) and historical student records of retention (29%) with female students and students of underrepresented minority groups most benefiting from the program.

There is also a well-established body of literature evaluating interventions designed to aid all students in a setting, where the observed benefits likely extend to at-risk students within the setting. Meta-analyses of cooperative learning in chemistry,\textsuperscript{24,26} and active learning in STEM,\textsuperscript{26} indicate a consistent positive effect of these pedagogical techniques on student performance overall. Lewis and Lewis,\textsuperscript{15} explicitly investigated the impact of a peer-led guided inquiry pedagogy to mediate the relationship between SAT\textsubscript{M} and student performance in
chemistry. It was found that the pedagogy improved the average academic performance for students regardless of their SATM score but had no significant impact on the relationship between SATM and student performance in chemistry, thus it did not ameliorate the differential performance between at-risk students in comparison to the rest of the cohort.

**Purpose of the Study**

Despite a wide literature base showing a consistent predictive ability of SATM to identify at-risk students, none of the studies report the chemistry topics where at-risk students disproportionately struggle or succeed. As a result, both the design of prescriptive suggestions and interventions are without an evidence base on the chemistry topics to target. Further, evaluations of interventions in the form of pedagogical change or extracurricular support have shown benefits for all students at the setting including at-risk students, but there is a lack of evidence that the differential performance of at-risk students has been remedied. This study seeks to characterize these topics to inform efforts to aid at-risk students. The investigation will be guided by two principal research questions:

1) Which topics most consistently pose differential difficulty for at-risk students, where differential performance is measured by the difference between at-risk students and the remainder of the cohort?

2) How critical are the identified topics with differential difficulty to student success within introductory chemistry, particularly among at-risk students?

**Methods**

**Treatment of Missing Data**

The primary analyses to follow occur with the 2,979 (52.9%) of students who both have an SATM score and completed all assessments (three interim and one ACS final). Of these 2,979 students with complete data, 753 (25.3%) were identified as having scored within the
bottom quartile of the SATM and are herein referred to as the at-risk cohort. Students without SATM scores, with ACT math composite scores (ACTM), and/or did not take either the ACS exam or the three interim assessments were considered separately within the missing data analysis found in the Appendix E. Students who were missing an individual item (e.g., left a question blank) on a particular assessment but otherwise completed the assessment were assumed to have been unable to answer that particular question and thus their response was treated as an incorrect response.

**Assessment Items Scored and Mapped onto Major Topics**

Assessment items from interim tests ($N = 78$ for each semester, $N_{total} = 312$) for the four semesters were written from a common list of learning objectives. A list of major topics was created to map onto learning objectives in such a way that learning objectives of similar topic matter were combined. All assessment items from two semesters, interim ($N = 156$) and ACS ($N = 70$), were then mapped onto the topic list via an iterative process. The percent agreement was calculated, and the discrepancies were reviewed creating revisions to the topics list. This process was iteratively continued until a topic list was generated that resulted in 89.8% agreement between the authors with the remaining discrepancies discussed to reach a consensus. This resulted in a list of 16 major topics that were mapped onto learning objectives (see Appendix C). All remaining interim assessment items were then mapped onto the 16 major topics by a single author.

As a check on the consistency in which items could be assigned topics, two instructors at the research setting were provided the refined topic list and independently assigned items from different semesters. The percent agreements between these instructors and the researchers were 83% and 80% (Cohen’s Kappa 0.86 and 0.83; Gwet’s AC1 0.86 and 0.84 respectively). These values are interpreted as strong levels of agreement and suggest topics are distinct enough to provide consistent assignment of assessment items. The 16 major topics
are arranged in order of teaching sequence at the setting and are followed by both distinctive, two-letter codes and frequencies on interim semester assessments in Table 11.

Table 11. Topics in order of teaching sequence followed by two-letter codes and frequencies.

<table>
<thead>
<tr>
<th>#</th>
<th>Topics (Codes)</th>
<th>Spring 2016</th>
<th>Fall 2016</th>
<th>Spring 2017</th>
<th>Fall 2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Structure of the Atom (SA)</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>The Mole Concept (MC)</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>Nomenclature (NM)</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>Stoichiometry (ST)</td>
<td>3</td>
<td>7</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>5</td>
<td>Molarity (MR)</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>Reactions in Solution (RS)</td>
<td>1</td>
<td>3</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>7</td>
<td>Redox Reactions (RR)</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>8</td>
<td>The Gas Laws (GL)</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>Changes in Enthalpy (CH)</td>
<td>7</td>
<td>4</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>Changes in Energy (CQ)</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>11</td>
<td>Properties of Light (PL)</td>
<td>5</td>
<td>7</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>12</td>
<td>Electron Configurations (EC)</td>
<td>11</td>
<td>10</td>
<td>5</td>
<td>14</td>
</tr>
<tr>
<td>13</td>
<td>Periodic Trends (PT)</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>Lewis Structures (LS)</td>
<td>9</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>15</td>
<td>Molecular Geometry (MG)</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>Valence Bond Theory (VB)</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

Note that later topics such as Lewis structure, molecular shapes and valence bond theory were occasionally covered after the last interim assessment and therefore were only tested on the ACS exam. For the ACS Exam, the following topics had fewer than three items per topic: reactions in solution (RS), redox reactions (RR), changes in energy (CQ) and properties of light (PL). Mean topic difficulties (P) for the interim (Spring 2016, ‘17, Fall ‘16, and ‘17) and ACS exam assessments were calculated for each semester as the percent correct of assessment items belonging to each topic (see Figure 13).
Figure 13. A wide range of mean topics scores were observed on interim and ACS exams overall.

Here, the interim average represents the average difficulty for all interim assessment items related to a particular topic. The ACS average represents the same for those assessment items on the ACS exam. The overall average (in blue) represents the interim average plus the ACS average divided by 2 thereby giving equal weight to each source of assessment items. Overall, students appear to have experienced a wide range of topic difficulties over the four semesters. ACS assessment topics (dashed, gray line) appear to have greater topic difficulties than topics measured within interim assessments (dotted, gray line). The topics on which students performed with the least difficulty were the structure of the atom (SA; $P = 78.65\%$), properties of light (PL; $P = 76.50\%$), and reactions in solution (RS; $P = 69.40\%$). Molecular geometry (MG; $P = 38.50\%$), only tested on the ACS exam, posed a particularly notable difficulty to students overall. Other topics on which students encountered moderate difficulty include redox reactions (RR; $P = 55.43\%$), changes in enthalpy (CH; $P = 56.32\%$), and molarity (MR; $P = 59.70\%$).

The data presented in Figure 13, and ensuing figures, represents categorical data on the x-axis (major topics). The choice was made to represent this data in a line graph to visually emphasize the relationship among the interim, ACS and overall averages across topic and it is
not meant to portray continuity in performance between topics. To explore the topics on which at-risk students most disproportionately struggled, mean topic scores were calculated for at-risk and not at-risk students. Differences between the groups’ mean topic scores were measured via effect size (Cohen’s $d$) to control for the variability within the scores.\textsuperscript{29} Effect size represents a standardized measure of the differential performance between the two groups where a positive number indicates the not at-risk group outperformed the at-risk group. A $d = 0.2$, $0.5$, and $0.8$ are described as having small, medium, and large effect sizes, respectively.

**Results**

**At-Risk Students are Disproportionately Challenged on Mole Concept and Stoichiometry**

**Assessment Items**

Mean effect sizes averaged across four semesters for both interim and ACS assessments are provided in Figure 14.

![Figure 14](image)

**Topics (in order of teaching sequence)**

- **Overall**
- **Interim**
- **ACS**

Figure 14. Mole concept (MC) and stoichiometry (ST) are topics posing incommensurate difficulty to students with low math aptitude scores.
Interim, ACS and overall averages used the same approach as above except each item was scored with Cohen’s $d$ as the metric of differential performance. Molecular geometry (MG) and valence bond theory (VB) were only assessed in the ACS exam. The effect sizes for the topics of the gas laws (GL) and electron configurations (EC) diverge when comparing interim (dotted line) to ACS (dashed line) assessments. All topics featured at least an effect size between small and medium which indicates that at-risk students struggle on all first-semester chemistry topics when compared to the not at-risk cohort; however, larger effect sizes on particular topics identify areas of particular concern. Two topics emerge as posing the most disproportionate and consistent challenge to at-risk students: mole concept (MC) and, to a greater extent, stoichiometry (ST) with medium to large effect sizes of 0.60 and 0.70, respectively.

To determine the consistency of differential performance on mole concept and stoichiometry within the data, each semester was analyzed separately (results are described in the Appendix D). It was found that stoichiometry was a relative peak in each of the four semesters and mole concept was a relative peak in three of the four semesters. The consistent differential performance found for these topics across each semester discounts the possibility that the differential performance observed is a result of random noise within the data.

Mole Concept and Stoichiometry are Essential to Student Success

To determine the degree to which the SATM may measure a student’s potential performance on the topics of mole concept and stoichiometry as compared to other topics, correlations for each topic to the SATM, interim and summative assessments were computed in Table 12 (on the next page).
Table 12. Pearson’s correlation values for topic scores, the SATM, and assessments.

<table>
<thead>
<tr>
<th>Pearson’s Correlations (R)</th>
<th>SATM</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>ACS Exam</th>
<th>Average Topic R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test 1</td>
<td>0.44</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test 2</td>
<td>0.45</td>
<td>0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Test 3</td>
<td>0.37</td>
<td>0.57</td>
<td>0.61</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACS Exam</td>
<td>0.49</td>
<td>0.70</td>
<td>0.71</td>
<td>0.67</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA</td>
<td>0.35</td>
<td>0.57</td>
<td>0.43</td>
<td>0.42</td>
<td>0.56</td>
<td>0.37</td>
</tr>
<tr>
<td>MC</td>
<td>0.44</td>
<td>0.68</td>
<td>0.56</td>
<td>0.50</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td>NM</td>
<td>0.34</td>
<td>0.70</td>
<td>0.53</td>
<td>0.53</td>
<td>0.69</td>
<td>0.46</td>
</tr>
<tr>
<td>ST</td>
<td>0.49</td>
<td>0.73</td>
<td>0.60</td>
<td>0.51</td>
<td>0.75</td>
<td>0.49</td>
</tr>
<tr>
<td>MR</td>
<td>0.39</td>
<td>0.64</td>
<td>0.59</td>
<td>0.48</td>
<td>0.69</td>
<td>0.46</td>
</tr>
<tr>
<td>RS</td>
<td>0.27</td>
<td>0.54</td>
<td>0.50</td>
<td>0.42</td>
<td>0.57</td>
<td>0.39</td>
</tr>
<tr>
<td>RR</td>
<td>0.37</td>
<td>0.48</td>
<td>0.57</td>
<td>0.47</td>
<td>0.62</td>
<td>0.41</td>
</tr>
<tr>
<td>GL</td>
<td>0.39</td>
<td>0.58</td>
<td>0.65</td>
<td>0.53</td>
<td>0.71</td>
<td>0.46</td>
</tr>
<tr>
<td>CH</td>
<td>0.39</td>
<td>0.52</td>
<td>0.69</td>
<td>0.55</td>
<td>0.71</td>
<td>0.45</td>
</tr>
<tr>
<td>CQ</td>
<td>0.33</td>
<td>0.41</td>
<td>0.55</td>
<td>0.42</td>
<td>0.56</td>
<td>0.36</td>
</tr>
<tr>
<td>PL</td>
<td>0.25</td>
<td>0.36</td>
<td>0.40</td>
<td>0.44</td>
<td>0.48</td>
<td>0.32</td>
</tr>
<tr>
<td>EC</td>
<td>0.32</td>
<td>0.49</td>
<td>0.55</td>
<td>0.64</td>
<td>0.63</td>
<td>0.41</td>
</tr>
<tr>
<td>PT</td>
<td>0.33</td>
<td>0.48</td>
<td>0.50</td>
<td>0.66</td>
<td>0.64</td>
<td>0.40</td>
</tr>
<tr>
<td>LS</td>
<td>0.30</td>
<td>0.52</td>
<td>0.53</td>
<td>0.70</td>
<td>0.70</td>
<td>0.43</td>
</tr>
<tr>
<td>MG</td>
<td>0.29</td>
<td>0.62</td>
<td>0.51</td>
<td>0.46</td>
<td>0.66</td>
<td>0.44</td>
</tr>
<tr>
<td>VB</td>
<td>0.12</td>
<td>0.30</td>
<td>0.36</td>
<td>0.53</td>
<td>0.54</td>
<td>0.31</td>
</tr>
</tbody>
</table>

Conditional formatting presenting as various shades of blue in Table 12 emphasize larger effect sizes (darker shades blue) from smaller effect sizes (lighter shades of blue). Also included in the last column of Table 12 is the average topic correlation which was found by correlating the topic in the row with each other topic and then computing the average of those correlations.

The topics of mole concept (MC; \( r = 0.44 \)) and stoichiometry (ST; \( r = 0.49 \)) are moderately correlated with a student’s SATM score; and are slightly higher than the correlation of SATM with other topics. ACS exam scores are highly correlated with the topics of stoichiometry (ST; \( r = 0.75 \)), mole concept (MC), gas laws (GL), and changes of enthalpy (CH)
each with values of $r = 0.71$. The topics of mole concept and stoichiometry are also among the topics with the strongest relation with student success in other topics; stoichiometry had the highest correlation with all other topics ($ST; r = 0.49$) and the topics of mole concept ($MC; r = 0.45$), molarity ($MR; r = 0.46$), gas laws ($GL; r = 0.46$), and changes in enthalpy ($CH; r = 0.45$) were also strongly related. The remaining topics had correlations between 0.31 and 0.41 suggesting an overall interconnected nature of the topics or a common skill set needed to succeed.

Upon reviewing assessment items and learning objectives, examples of applied forms of stoichiometry and mole concept appears in subsequent topics. Assessment items from interim exams were identified that explicitly relied on applied forms of stoichiometry and mole concept within subsequent topics and exploring the differential performance indicates that the recurrence of stoichiometry and mole topics can partially explain some differential performance observed in later topics. Four exemplary items of this point are described in Figures 15, 16, 17, and 18; for copyright reasons items from the ACS Exam were explicitly not included as exemplars. Each figure includes the assessment item with the correct answer in red, the assigned learning objective (LO), percent correct ($P$), the differential between the percent of not at-risk versus at-risk students answering correctly ($MD$) and the effect size ($d$).

**F17.T2.Q8.** A 0.465 g sample of an unknown compound occupies 245 mL at 298 K and 1.22 atm. What is the molar mass of the unknown compound?

<table>
<thead>
<tr>
<th>Option</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A)</td>
<td>12.2 g/mol</td>
</tr>
<tr>
<td>B)</td>
<td>26.3 g/mol</td>
</tr>
<tr>
<td>C)</td>
<td>33.9 g/mol</td>
</tr>
<tr>
<td>D)</td>
<td>38.0 g/mol</td>
</tr>
<tr>
<td>E)</td>
<td>81.8 g/mol</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Topic:</th>
<th>The Gas Laws (GL)</th>
<th>Frequencies (f):</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO: 8.c. Use the ideal gas law to determine molar volume, density, and molar mass of a gas.</td>
<td></td>
<td>A: 96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B: 76</td>
</tr>
<tr>
<td>P:</td>
<td>78.01%</td>
<td>C: 43</td>
</tr>
<tr>
<td>MD:</td>
<td>25.07%</td>
<td>D: 862</td>
</tr>
<tr>
<td>d:</td>
<td>0.627</td>
<td>E: 27</td>
</tr>
</tbody>
</table>

Figure 15. Interim assessment item on mole concept in gas laws.
**F17.T1.Q12.** A flask containing 450 mL of 0.500 M HBr was accidentally knocked to the floor. How many moles of $K_2CO_3$ would be required to completely react with HBr according to the following equation? 

$$2 \text{HBr(aq)} + \text{K}_2\text{CO}_3\text{(aq)} \rightarrow 2 \text{KBr(aq)} + \text{CO}_2\text{(g)} + \text{H}_2\text{O(l)}$$

A) 0.113 mol  
B) 0.225 mol  
C) 0.250 mol  
D) 0.450 mol  
E) 1.00 mol

<table>
<thead>
<tr>
<th>Topic:</th>
<th>Molarity (MR)</th>
<th>Frequencies ($f$):</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO: 5.c. Perform stoichiometric calculations for reactions in aqueous solutions.</td>
<td>A:</td>
<td>1043</td>
</tr>
<tr>
<td></td>
<td>B:</td>
<td>429</td>
</tr>
<tr>
<td>P:</td>
<td>60.72%</td>
<td>C:</td>
</tr>
<tr>
<td>MD:</td>
<td>30.62%</td>
<td>D:</td>
</tr>
<tr>
<td>$d$:</td>
<td>0.851</td>
<td>E:</td>
</tr>
</tbody>
</table>

Figure 16. Interim assessment item on stoichiometry in molarity.

**S16.T1.Q3.** What volume of $CO_2$ gas (MM = 44.01 g/mol) can be produced from the reaction of 2.00 L of 1.00 M HCl reacting with excess CaCO$_3$ at standard temperature and pressure? 

$$2 \text{HCl (aq)} + \text{CaCO}_3\text{(aq)} \rightarrow \text{CO}_2\text{(g)} + \text{H}_2\text{O(l)} + \text{CaCl}_2\text{(aq)}$$

A. 0.509 L  
B. 2.00 L  
C. 4.00 L  
D. 22.4 L  
E. 44.8 L

<table>
<thead>
<tr>
<th>Topic:</th>
<th>The Gas Laws (GL)</th>
<th>Frequencies ($f$):</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO: 8.e. Relate stoichiometry calculations to the ideal gas law</td>
<td>A:</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>B:</td>
<td>61</td>
</tr>
<tr>
<td>P:</td>
<td>57.45%</td>
<td>C:</td>
</tr>
<tr>
<td>MD:</td>
<td>11.13%</td>
<td>D:</td>
</tr>
<tr>
<td>$d$:</td>
<td>0.226</td>
<td>E:</td>
</tr>
</tbody>
</table>

Figure 17. Interim assessment item stoichiometry in gas laws.

**S16.T2.Q2.** Given the chemical reaction:

$$C_3H_8(g) + 5 O_2(g) \rightarrow 3 CO_2(g) + 4 H_2O(g) \quad \Delta H = -2044 \text{ kJ}$$

How much energy would be released if 2.0 moles of $O_2$ were reacted?

A. $8.2 \times 10^2 \text{ kJ}$  
B. $2.1 \times 10^3 \text{ kJ}$  
C. $4.1 \times 10^3 \text{ kJ}$  
D. $2.0 \times 10^4 \text{ kJ}$  
E. $1.3 \times 10^5 \text{ kJ}$

<table>
<thead>
<tr>
<th>Topic:</th>
<th>Changes in Enthalpy (CH)</th>
<th>Frequencies ($f$):</th>
</tr>
</thead>
<tbody>
<tr>
<td>LO: 9.b. Describe the changes in $\Delta H$ when manipulating a chemical reaction (reversing, multiplying by a constant) as per Hess’s Law.</td>
<td>A:</td>
<td>471</td>
</tr>
<tr>
<td></td>
<td>B:</td>
<td>65</td>
</tr>
<tr>
<td>P:</td>
<td>73.01%</td>
<td>C:</td>
</tr>
<tr>
<td>MD:</td>
<td>19.78%</td>
<td>D:</td>
</tr>
<tr>
<td>$d$:</td>
<td>0.453</td>
<td>E:</td>
</tr>
</tbody>
</table>

Figure 18. Interim assessment item on stoichiometry in changes in enthalpy.
Two of these assessment items (Figures 15 and 16) have effect sizes greater than 0.6; commensurate with the effect sizes observed for mole concept and stoichiometry overall and above the average effect sizes for gas laws and molarity, respectively (see Figure 14). Two other assessment items (Figures 17 and 18) still feature a differential performance of 0.226 and 0.453, respectively, but are below the overall average observed for their respective topics. By item analysis reveals considerable variation among the differential performance metric that is evened out by the topic score. Individually, each item could be impacted by various features. The item in Figure 17, for example, may have less differential owing to students selecting 22.4 L as the standard molar volume of a gas rather than performing the intended operation and with at-risk students disproportionately using this heuristic. Systemic investigation of by-item features that relate to differential performance can provide additional insight but was considered beyond the scope of the current investigation.

**At-Risk Students Achieving Proficiency in Mole Concept and Stoichiometry Performed Equitably or Surpassed their Not At-Risk Peers**

To further explore whether the topics of mole concept and stoichiometry are essential to success for at-risk students, a subgroup of at-risk students who performed well in these topics was identified. This subgroup is termed “Risers” and was identified as at-risk students who scored better than 65% on the interim and ACS assessment items for either the mole concept or stoichiometry topics. If stoichiometry and mole concept were topics imperative to the success of at-risk students, students achieving proficiency on these topics would be expected to have higher academic outcomes than their peers within the at-risk student group.

The mean SATM score of risers was 499, close to the mean SATM score for the at-risk group overall at 488 and considerably below the not at-risk students 598. The academic performance of risers surpassed not only that of their at-risk peers but was quite similar to that of the not at-risk students (see Figure 19 on the next page).
Figure 19. At-risk students proficient in mole concept and stoichiometry (Risers) perform similarly to the not at-risk cohort.

Risers were compared to not at-risk students by topic to determine if the result is only related to their performance on mole concept and stoichiometry. The results are shown in Figure 20 where a positive value indicates not at-risk performed higher than the risers.

Figure 20. Topic performance of at-risk students proficient in mole concept and stoichiometry (Risers).

Overall, the differential performance across topics tended to be proximate to zero, suggesting equal performance between the groups, or negative, indicating the risers scored higher on those
topics. These results suggest that at-risk students proficient on stoichiometry and mole concept succeed in the course at a comparable rate as the rest of the cohort and further supports the importance of these topics for at-risk student performance.

Explorations of the Missing Data

The above analyses focus on the 2,979 students, 52.9% of the overall cohort, with complete data (incoming SATM score and completion of all interim and ACS assessments). Students with missing data were analyzed, where possible, to investigate the extent the key finding, that at-risk students disproportionately struggle with mole concept and stoichiometry, was applicable.

Among the 890 students with no SATM score, 15.8% of the cohort had ACT scores and complete data on all interim and ACS assessments. In the U.S. system, the ACT serves as a nationally available alternative to the SAT, and also features a mathematics composite score (ACTM). To determine whether at-risk students could be identified using ACTM scores, concordance tables published by the College Board in 2016 provided a conversion between scores of the two standardize math aptitude tests. These conversions were developed using the equipercentile concordance method in which scores from each test were found to share the same percentile rank.30 For example, scores found at the 75th percentile nationally for the SATM correspond to the same percentile for converted scores on the ACTM. The bottom quartile of ACTM scores for this group of 890 students ranges between 20 to 23 depending on the semester. Referencing the College Board’s concordance tables, SATM scores corresponding to these scores are between 510 and 570. As the bottom quartile for SATM scores at this setting is between 490 and 570, students defined as at-risk by either relative SATM or ACTM performance appear reasonably comparable. At-risk students, among those without SAT scores, were then defined as those in the bottom quartile of the ACTM and mean effect sizes measuring differential performance is presented in Figure 21 (on the next page).
Figure 21. Trends related to disproportionate topic difficulty for at-risk students are consistent with students of lower ACTM scores.

The topics of mole concept and stoichiometry share the highest effect sizes observed indicating these topics also comprise the greatest differential between bottom quartile ACTM scores compared to the top three quartiles. With this measure, however, mole concept is observed as having a greater contribution to the difficulty of at-risk students than that of stoichiometry. Thus, the finding of differential performance on these topics appears to extend to the cohort of students without SAT scores but with ACT scores.

Another source of missing data are students who missed one or more of the interim assessments but did take the ACS assessment. Among those with SATM or ACTM scores, there were 575 students, 10.2% of the cohort, who missed one or more of the interim assessments but completed the ACS assessment. Reasons for missing an interim assessment are varied and include scheduling conflicts, ailments or emergencies and students were provided an alternative exam at a later date in these situations. Owing to the alternative interim assessments, the analysis focused on differential performance by topic on the ACS assessment and used either SATM when available or ACTM to classify at-risk students. The results (see Appendix E) found that mole concept and stoichiometry topics provided the largest differential performance between at-risk and not at-risk students mirroring the results for the larger cohort.
The remaining source of missing data includes students who did not take the ACS exam, 530 students or 9.4% of the cohort, and those who did not have SAT or ACT scores, 662 students or 11.7%. The ACS exam was required for successful completion of the course and students who did not take this exam either withdrew from the course, stopped attending the course or in rare cases arranged to complete the exam in a following semester. Among this group, the amount of interim assessments completed is inconsistent which prevents an exploration of differential topic performance. More descriptively, the students in each group (at-risk or not) that took the first test are not the same as those who took the second test and therefore a comparison across topics using the interim assessments would not be appropriate. Finally, there was no way to identify an at-risk cohort among the 11.7% of students who did not have SAT or ACT scores so the key finding could not be explored for this group either. That said, the differential performance by topic was consistent among the three groups: those with complete test scores and SAT scores, those with complete test scores and ACT scores but no SAT scores, and those who completed the ACS exam while missing one or more interim assessments. Combined these three groups represent 78.9% of the overall cohort.

Discussion and Implications

Of the overall cohort comprised of 5,636 students, the primary analyses presented in this article reflect a consistent group of 2,979 students entering first-semester chemistry with SATM scores and having completed all interim and ACS assessments items within their respective semesters. 753 at-risk students demonstrated statistically significant differences in mean performance on all topics in introductory chemistry in comparison to not at-risk students. Of these at-risk students, students self-identified as Black/African American and Hispanic/Latino comprised a greater percentage than that of the not at-risk cohort.

The mean difficulties (average percent correct) for students overall suggest reasonable performance on assessment items related to the topics of mole concept (MC) and stoichiometry.
(ST); each with respective means of 64.84% and 62.49% (see Figure 13), however, these topics also featured the largest differential in performance between at-risk as compared to not at-risk students. The effect sizes between these two groups are medium to large with $d = .60$ for mole concept and $d = .70$ for stoichiometry (see Figure 14) across four semesters of data, and this finding was also consistent among those with missing data where applicable trends could be explored. This differential performance indicates these topics present consistent and significant challenges to students with low SATM scores.

Stoichiometry is not only the topic most correlated to performance on the ACS exam but is also the topic most correlated with SATM and other topic scores (see Table 12). Explorations of other topics with medium effect sizes such as molarity (MR; $d = .53$), gas laws (GL; $d = .51$), and changes of enthalpy (CH; $d = .51$) revealed mole concept and stoichiometry may partially explain the differential performance observed within these topics (see Figures 15, 16, 17, and 18). At-risk students demonstrating a proficiency of 65% or higher on assessment items of either mole concept or stoichiometry assessment items far outperformed their peers with comparable SATM scores in the at-risk cohort. Additionally, these proficient at-risk students performed either equitably or exceeded that of the not at-risk cohort (top three quartiles of the SATM) on both interim and ACS assessments (see Figures 19 and 20). The results reflect the findings of Tai et al., who found students’ self-report of time spent on stoichiometry in high school was varied and had the strongest relationship with first-semester chemistry grades among a set of other chemistry topics.

The research literature offers promising instructional techniques for promoting students’ understanding of stoichiometry. Kimberlin and Yezierski, developed, implemented, and evaluated two inquiry lessons to promote understanding of stoichiometry concepts. One lesson involved using particulate models and questions to elicit and address students’ common incorrect ideas regarding mole ratios. The other lesson targets the interpretation between symbolic, algorithmic and macroscopic descriptions of stoichiometry. The lessons were
implemented in five high schools (secondary school) introductory chemistry courses and were evaluated using a pre/post implementation of conceptual stoichiometry tests. Comparing the post-test to the pre-test, statistically significant gains with a large effect size were observed.

These findings provide potential paths forward toward improving students understanding of stoichiometry concepts. Coupled with the findings presented here on the differential performance of at-risk students on stoichiometry topics, it is hypothesized that the effective implementation of such activities could promote the success of at-risk students. That said, Cacciatore and Sevian,\textsuperscript{33} found that an inquiry laboratory experiment on stoichiometry led to improved performance on stoichiometry questions but not on indirect applications of stoichiometry making the transfer of knowledge still an important hurdle to realizing this hypothesis.

Returning to the aforementioned hypothesis that low SATM performance would foreshadow a lack of quantitative skills, which is responsible for differential performance, the results are arguably more nuanced. First, mole concept and stoichiometry as they are articulated here include more than just algebraic manipulation. In particular, these topics rest on both proportional reasoning and the translation between mathematical and symbolic representations, with the latter serving as the language for communicating chemistry concepts.\textsuperscript{34} The difficulty in translation from symbolic to mathematical representations matches a common trend observed in the literature base exploring students’ problem-solving strategies in chemistry where students can apply algorithmic strategies to assessment items but are unable to meaningfully transfer their understanding to more conceptual or unique scenarios of the topic.\textsuperscript{23,35–37}

The results here suggest that this issue may be particularly pronounced among students with low SATM. Second, other topics in the course reliant on more direct algebraic manipulation such as gas laws, changes of energy, and properties of light had differentials closer to the average of observed topics. These results suggest that while at-risk students appear to struggle
on topics such as mole concept and stoichiometry, their difficulties appear less pronounced in response to topics that focus on quantitative reasoning with less emphasis on chemical formulas or reactions (e.g., gas laws, properties of light).

The focus on topic averages rather than analyses of individual assessment items or the process by which students answer these items is a limitation of this work. While the intent of the article was to elucidate the topics on which at-risk and first-semester chemistry students overall most struggle, efforts to aid at-risk students would benefit by further characterizing the tasks on which students most struggle in chemistry. Carrying out this work may be achieved via item analysis and qualitative investigations to offer additional insight into efforts towards supporting at-risk students. Alternatively, explorations as to the learning progressions of students formulating concepts fundamental to stoichiometry are limited and could be further explored via repeated measures using surveys, interviews or assessments to measure student solution strategies or self-reflective explanations of said strategies. Characterizing such a learning progression may provide a framework on which educators can better assess and design learning experiences to facilitate high conceptual understanding of these topics,\textsuperscript{38,39} and promote the success of at-risk students.

An additional limitation of this work is that the data collection was limited to one institution and is reliant on closed-ended assessments and potentially the sequencing of topics in this curriculum. With this acknowledgement, it remains uncertain the extent that at-risk students would struggle on these topics at varying institution types, assessment methods and curricula. Further studies into the performance of at-risk students are needed to substantiate a more generalizable claim and can also clarify whether proficiency in mole concept and stoichiometry is predictive of student success in general chemistry as this would establish greater need still for effectively designed lessons and assessments on these fundamental topics with evaluations that explicitly consider at-risk students. Should these studies be tested and implemented into consideration for the instruction of at-risk students, potential for the reduction
of attrition and inclusion of both students of inequitable pre-college math preparation and students of underrepresented minorities could be realized.

Conclusion

This study presents a trend towards the characterization of the topics on which at-risk students (as comprised by students having scored relatively in the bottom quartile of each semester’s cohort on the quantitative composite of the “old” SAT, the “new” SAT and the math composite of the ACT) most disproportionately struggle: the mole concept and stoichiometry. This trend for at-risk student struggle with these two topics was observed along two academic years (2016-2017) across four semesters of data. The results call attention to an emphasis on these topics to achieve the goal of an equitable class setting that is responsive to a range of incoming student preparations. This attention is further warranted given the correlations of these topics to success in the course overall and the presence of these topics in subsequent topics such as gas laws and thermodynamics. Additionally, at-risk students who were successful on these topics had similar assessment performance as the not at-risk cohort. The existing research literature offers potential avenues toward promoting student success on these topics which can serve as a fruitful area of future research. The methodology presented herein provides a path for future research to better understand other sources of inequitable success within chemistry courses. Finally, students of the at-risk cohort were also comprised of a greater number of students having self-identified as Black/African American and Hispanic/Latino and thus studies aimed at improving at-risk student outcomes have potential for ameliorating both differences observed amongst students of differential math preparation and towards diversification of STEM careers.
References


CHAPTER 5:
AN EXPLANATIVE BASIS FOR THE DIFFERENTIAL PERFORMANCE OF STUDENTS WITH LOW MATH APTITUDE

Note to Reader
Portions of this chapters have been previously published in Chemistry Education Research and Practice, 20: 570-593. These works have been reproduced for this dissertation with permission from the Royal Society of Chemistry (see Appendix B for Copyright Permissions).

Introduction
Prediction for the Sake of Intervention
Cognitive measures of math aptitude are heavily used in the research literature to identify students at-risk for unfavorable outcomes in college-level chemistry.\textsuperscript{1-10} Each study provided a means by which instructors could identify at-risk chemistry students via incoming preparation in mathematics. More recent works identify at-risk students as those scoring at or below the bottom quartile of math aptitude scores for each semester’s cohort, which will be used in the current study.\textsuperscript{8,10,11} While each of these studies presents evidence for the utility of math aptitude in predicting students’ assessment performance in chemistry, the work falls short in developing an understanding as to why students of low math aptitude perform poorly in general chemistry.

Many of the cited works provide practical implications for their findings. Implications include students’ assignment to remedial instruction,\textsuperscript{1,2,7} procedures at the institutional level to
apply cut-offs for math aptitude scores necessary for enrollment,¹ the assignment of remedial coursework or instruction with an emphasis in problem-solving,⁴ and the promotion of productive study habits by assigning practice problems for the students to complete before attending lecture.¹⁰ Each set of implications relies on the reasonable assumption that students of low math aptitude scores struggle with the mathematical components of chemistry. However, in 2012, Scott analyzed student response processes in the form of handwritten solutions to analogous mathematics and chemistry problems and identified no measurable change in students’ chemistry performance after practicing analogous mathematics problems.¹² In 2009, Donovan and Wheland hypothesized that the connection between mathematics and chemistry is unclear and may not be dependent on “actual mathematics knowledge” but rather the development of a higher-order cognitive, skill sets required in science.¹³ Additionally, limited or no success has been observed as to the influence of prerequisite coursework on performance gaps.²,¹³,¹⁴ From the research literature, it is clear that efforts to improve the success of students with low math aptitude are widespread and would benefit by furthering the evidence base as to why differentials persist in academic chemistry performance.

Prior work sought to identify the topics on which at-risk chemistry students most differentially perform.¹¹ Students scoring in the bottom quartile of math aptitude were observed to perform most differentially on items belonging to the mole concept and stoichiometry. Assessment of these topics rarely progressed beyond multiplication or division. Additionally, this previous study observed at-risk chemistry students who attained a proficiency of 65% or higher on each semester’s collective mole concept or stoichiometry assessment items to outperform both their peers in the at-risk group and those students not-at-risk. Despite the identification of the topics on which at-risk chemistry students most differentially perform and the success of students following proficiency on these topics, the study was unable to address why at-risk chemistry students differentially performed on these topics. More specifically, amongst the assessment items within these topics, what features of the assessment items posed differential
challenges to at-risk chemistry students? The current study, thereby, seeks to understand why the mole concept and stoichiometry present a disproportionate challenge to at-risk chemistry students and evaluate features of assessment design that result in the observed differential performance.

**Challenges Chemistry Students Face with the Mole Concept and Stoichiometry**

No prior work was identified to describe challenges with the mole concept or stoichiometry amongst chemistry students with low math aptitude scores. However, other works (described below) have outlined the challenges of low-performing chemistry students on these topics. The use of symbolism to communicate chemical equations has often been observed in the instruction and assessment of stoichiometry and has been regularly described as difficult for students to interpret.\(^\text{15}\) **Representational competence** comprises the skill sets by which elemental symbols and chemical formulae or reactions are interpreted.\(^\text{16}\) To provide an example of representational competence in the context of problem-solving, suppose a stoichiometry prompt requests the mass of a product given the mass of a reactant. One would then use the provided nomenclature, chemical formulae, or a chemical reaction to interpret coefficients and subscripts and discern the number of atoms comprising each substance and the stoichiometric ratio between reacting chemicals. In the context of stoichiometry, low-performing students were observed to conflate the roles of coefficients and subscripts in balanced chemical.\(^\text{17}\) These challenges with the interpretation of coefficients and subscripts appeared lessened via student involvement in the illustration of submicroscopic diagrams (pictorial depictions of chemical reactions) to visualize proportional relationships between chemical quantities.\(^\text{18}\) These studies describe student struggles with representational competence likely impeding their ability to quantify chemical proportions (e.g., deriving solutions to stoichiometry problems). A study published by Schank and Kozma in 2002,\(^\text{19}\) examined the impact of learning environments emphasizing representational competence via an interactive computer program that provides
visualizations of chemical. In this study, chemistry teachers communicated the importance of representational competence and provided ideas for the use of these skill sets toward student’s understanding of the law of mass conservation and how chemicals rearrange in a chemical reaction retaining stoichiometric proportions.

Similar to representational competence approaches for enhancing instruction to support student difficulty with systematic approaches, or the step-by-step solution processes used to arrange conversion factors, seemed to promote increasingly visual instructional. Phillips’ dissertation describes the efficacy of graphic organizers toward supporting young students’ transitions from Piagetian concrete-operational reasoning, or logic reasoning applied to physical entities, to that of formal-operational or abstract reasoning where deductive reasoning concerns hypothetical situations without the need to connect to a physical entity. An example of these forms of reasoning could involve how students solve the following prompt: “If Rob is taller than Jon and Jon is taller than Brandon, who is tallest?” If the student can deduce who is taller without creating a concrete representation, the student is likely relying on formal-operational reasoning. If, however, the student illustrates a picture to support their reasoning, the student may be in the concrete-operational stage but can use the image to reason, nonetheless.

Connections between concrete- and formal-operational reasoning to students of low math aptitude scores have been made in the literature. Phillips posits graphic organizers including a solution flowchart where bidirectional arrows between mass and moles are labeled with the operations involving molar mass required to interconvert these values can support students of concrete-operational reasoning to progress toward more formal-operational reasoning. Just as more visual instructional practices appeared to promote students’ progression toward formal-operational reasoning by proving a framework from which students can connect abstract concepts, so too have increasingly visual instructional practices positively impacted the academic success of high school chemistry students presenting with relatively low proficiency in proportional reasoning when solving stoichiometry problems. "Supplemental, less
mathematical, and more visual approaches” were recommended for students who struggled with proportional reasoning (p. 175).²¹

Interestingly, where visuals have been successful, practice with proportional reasoning in mathematical contexts was not observed to transfer to the context of chemistry.¹²,²³ In these studies, mathematical skill sets were not inherently transferred to analogous chemistry skill sets particularly where proportional reasoning was involved (as is observed with stoichiometry).

Thus, a final theme concerning the challenges observed amongst chemistry students in the literature was conceptual understanding regarding measures particular to the mole concept and its application to stoichiometry.²⁴ In the works cited previously, students were unable to distinguish between relevant units of measure concerning stoichiometric conversions (e.g., mass, moles, and molar mass). Analyses of textbooks and instructional materials suggest that some of these challenges could result from students’ introduction to these quantities.²⁵,²⁶ A literature review conducted in 2002 also describes challenges with the conceptual understanding of stoichiometry as related to how these entities are introduced during instruction.²⁷ The review identifies a consensus in the literature concerning students’ conceptual misunderstanding of units used in stoichiometry as attributable to students’ misconstruction of mole as a mass or number particular to the property of a substance such as atomic mass. From the perspective of students’ solution processes, a tendency toward avoidance of the mole was observed amongst students and was thought to stem from this lack of conceptual understanding of terms in the context of students’ solution processes.²⁸ Instead, students were often observed to rely on proportional reasoning techniques (e.g., cross-multiplication) to replace methods modeled step-by-step by their instructors suggesting students lean on their mathematical skill sets to compensate for a lack of understanding of calculations in the context of chemistry.

Ultimately, the themes described in the research literature concerning student challenges with the mole concept and stoichiometry cited above can be summarized as related
to 1) representational competence, 2) systematic approaches, and 3) conceptual understanding. Each theme is summarized succinctly below.

1. **Representational competence**: interpreting elemental symbols and chemical formulae or reactions to identify proportionality between chemical substances and inherent physical properties relevant to the problem-solving process.

2. **Systematic approaches**: an algorithmic or step-by-step approach resulting in a solution that presents coherent units of measurement.

3. **Conceptual understanding**: the underlying terminology and theories of science used to reason a viable solution deductively.

This review of the literature describing how low-performing chemistry students and chemistry students as a whole struggle with the mole concept and stoichiometry was used to develop a conceptual framework as to how these challenges may manifest and evaluate whether or not these challenges are a description of those observed amongst at-risk chemistry students.

**Purpose**

Chemistry courses serve as a primary conduit in the STEM pipeline, and remediation of the differential outcomes observed for these students could serve to diversify the population of students who emerge successfully.\(^ {29}\) There is currently a lack of evidence on the topic matter, assessment strategies, and experiences contributory to the differential performance observed amongst chemistry students with low math aptitude scores, despite a sizable literature base relating math aptitude to chemistry performance suggesting this differential performance remains widespread.\(^ {1-10}\) Further, given the nearly ubiquitous adoption of compulsory prerequisite courses in mathematics before enrollment in college-level chemistry,\(^ {30}\) it is clear that instructors and administrators are rightfully concerned about differential performance and would benefit from an expansion of the relevant evidence base.
This study sought to analyze the solution processes of 348 chemistry students (including 101 students with relatively low math aptitude scores) to mole concept and stoichiometry assessment items where differential performance has been previously identified at the research setting. The results can inform instructional and assessment practices while providing a greater insight as to the skill most strongly predicted by chemistry students’ precollege math aptitude scores. Understanding how these students experience inequitable difficulty in chemistry has the potential to inform practices that serve to ameliorate gaps in performance and support more equitable, evidence-based practices for intervention, instruction, and assessment. Without this knowledge, current practices may unintentionally contribute to the propagation of these inequitable outcomes in first-semester general chemistry.

Research Objectives

Two primary research objectives guided the development of this study:

1. Describe the challenges observed amongst at-risk chemistry students in solution processes concerning the mole concept and stoichiometry assessment items.
2. Compare and contrast the challenges that present amongst not-at-risk and at-risk chemistry students to describe potential causes for the differential performance observed.

Methods

Development and Design of Post-Assessment Survey Prompts

Following the finding that at-risk chemistry students at the research setting struggled most differentially on mole concept and stoichiometry assessment items, test items related to these topics were selected from those drafted by a committee of the course instructors. Items were selected a priori (before the information was available as to students’ performance on the
item). Each assessment item was expected to require a mole-to-mole conversion in different contexts: mole concept, theoretical yield, gas laws, and changes in energy.

Survey prompts were administered via students’ online learning management system. Following each of the three interim exams, students were asked to respond to open-ended prompts designed to elicit their solution processes within a week of their in-course interim chemistry tests. Surveys were administered following interim exams to receive students’ solution processes after students prepared for, completed, and had access to the answer key for the assessment items presented in the surveys.

The first two prompts followed the first interim test, the next followed the second interim test, and the last followed the third interim test. No topics were identified as involving the mole concept or stoichiometry on the third interim test. Students were able to freely edit their answers on the survey until the established deadline. The survey prompts followed a consistent format as presented in Box 2 (see next page). Students were asked to review these assessment items and, given the correct answer, describe their solution processes for attaining this answer. Students received 0.5% added to their final grades for the completion of each of the three surveys. Students’ were also informed “There are no correct answers to the survey, but repetitive or non-responsive answers may not result in extra points” regarding the expectations for their receipt of these extra credit points.

Coding Students’ Responses

Responses were first collected into a database using the export function of the learning management system, merged with roster and assessment data files using SPSS. Then, responses were imported to the qualitative analysis data software MAXQDA, as a means to catalog and organize codes, memos, and logbook (comprised of a dated and timed series of entries wherein the researchers reflect on the data and the process by which the data has been analyzed) while interacting with the data.
Box 2. Survey prompts for each stoichiometric task.

The Mole Concept Item
For the test question (below), please explain how to arrive at the correct answer (“E”).
How many moles of oxygen are in 15 g of K₂SO₄?

A) 0.022 mol B) 0.043 mol C) 0.086 mol D) 0.17 mol E) 0.34 mol

The Stoichiometry Item
For the same test question (below), please explain how to arrive at the correct answer (“C”).
Calculate the amount of NH₃ gas produced by reacting 45.6 g of H₂ gas with an excess of N₂ gas according to the following chemical equation.

\[ 3 \text{H}_2 (g) + \text{N}_2 (g) \rightarrow 2 \text{NH}_3 (g) \]

A) 85.6 g B) 171 g C) 257 g D) 385 g E) 579 g

The Gas Laws Item
For the same test question (below), please explain how to arrive at the correct answer (“A”).
If 126 L of H₂(g) at 15°C and a constant pressure of 0.920 atm reacts with excess N₂(g) as shown below, how many moles of NH₃(g) are produced?

\[ 3 \text{H}_2 (g) + \text{N}_2 (g) \rightarrow 2 \text{NH}_3 (g) \]

A) 3.27 mol NH₃ B) 4.90 mol NH₃ C) 7.36 mol NH₃ D) 55.7 mol NH₃ E) 84.0 mol NH₃

The Changes in Energy Item
For the same test question (below), please explain how to arrive at the correct answer (“B”).
Consider the following balanced chemical equation:

\[ 4 \text{NH}_3(g) + 5 \text{O}_2(g) \rightarrow 4 \text{NO}(g) + 6 \text{H}_2\text{O}(l) \quad \Delta H_{\text{rxn}} = +1168 \text{kJ} \]

How much heat is absorbed/released when 2.05 mol of NH₃(g) reacts with excess O₂(g) to produce NO(g) and H₂O(l)?

A) 5.99 x 10² kJ of heat is released. B) 5.99 x 10² kJ of heat is absorbed.
C) 2.40 x 10³ kJ of heat is released. D) 2.40 x 10³ kJ of heat is absorbed.
E) 1.02 x 10⁴ kJ of heat is absorbed.
The first author devised a deductive coding scheme using themes observed from the literature base cited in the introduction. These works served as a conceptual framework concerning why students (in general) tend to struggle with the mole concept and stoichiometry assessment items. From this framework, a coding scheme was developed to describe the errors students make in their solution processes for the assessment items selected for the survey (Table 13).

Table 13. A priori coding scheme comprised of the challenges by which students have been observed to struggle with the presented mole concept and stoichiometry assessment items in past works.

<table>
<thead>
<tr>
<th>Code</th>
<th>Code Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Representational competence</strong></td>
<td></td>
</tr>
<tr>
<td>Atomic vs. Molecular</td>
<td>Atoms comprising species are misrepresented as covalent compounds using subscripts (e.g., 4 O in potassium sulfate vs. O₄).</td>
</tr>
<tr>
<td>Misrepresented Proportions</td>
<td>Proportions identified by the subscripts and/or coefficients of a chemical formula are misrepresented.</td>
</tr>
<tr>
<td>Elemental Symbols</td>
<td>Atomic masses are incorrectly assigned to elements.</td>
</tr>
<tr>
<td>Sign Conventions</td>
<td>The meaning of + and - signs in the context of scientific notation or interpreting energy transfer is reversed.</td>
</tr>
<tr>
<td>Nomenclature</td>
<td>Chemical symbols are misinterpreted when determining the predominate state of matter or name of a chemical species.</td>
</tr>
<tr>
<td><strong>Systematic approaches</strong></td>
<td></td>
</tr>
<tr>
<td>Inaccurate Conversion Factors</td>
<td>Selected conversion factors between units of measure (e.g., mass to moles; moles to mass) are incapable of resulting in a viable measure for the converted value (e.g., converting mass to moles using Avogadro’s number).</td>
</tr>
<tr>
<td>Conversion Arrangement</td>
<td>Numbers and units are arranged improperly result in units incapable of achieving a coherent number and unit of measure (e.g., 12.1 moles O multiplied by 1 mol O/16.0 g O to achieve mass of O).</td>
</tr>
<tr>
<td>Sequence</td>
<td>Steps in the predicted solution path(s) are present but are rearranged in a manner that does not reflect a solution path that is chemically plausible.</td>
</tr>
<tr>
<td><strong>Conceptual Understanding</strong></td>
<td></td>
</tr>
<tr>
<td>Interchemical Unit Identity</td>
<td>Moles of a compound are indistinguishable from moles of a constituent element.</td>
</tr>
<tr>
<td>Unit Conflation</td>
<td>Differing units of measure are used interchangeably (e.g., molar mass as mass).</td>
</tr>
<tr>
<td>Intrachemical Unit Identity</td>
<td>Terminologies inherent to particular units are misapplied (e.g., using the term “the number of atoms” to describe the moles of the same chemical).</td>
</tr>
</tbody>
</table>
When challenges emerged similar to or beyond those described by the conceptual framework, the erred processes were used to generate one-to-four-word descriptions summarizing the challenges demonstrated by the students who authored them. Similar codes were then combined via pattern coding \(^3\) to group these summaries, assess their commonality, and consider each in the context of the conceptual framework in concerning the difficulties observed in students struggling with the mole concept and stoichiometry. Student solutions were coded so that one student's response could have multiple codes applied. This iterative process started with both authors coding common subsets of the data, discussing differences in the manner by which codes were applied, and altering the codebook to improve parsimonious and consistent coding of students' responses.

Subsets of the data used for interrater were selected from inaccurate student responses of which there were 340 responses from the 247 not-at-risk (NAR) students and 199 responses from the 101 at-risk (AR) students. Coding rounds occurred by item (mole concept, stoichiometry, gas laws, and changes in enthalpy) where rounds were randomized and selected so that the number of inaccurate responses from the NAR and AR cohorts was equal. For example, the coding round wherein codes (see Table 13) were applied to student's response processes for the mole concept item included 62 (31 from each group) of the 138 (or 45\%) inaccurate responses. The researchers coded these items independently using a spreadsheet to record their applications of the codes, and later met to discuss the discrepancies and make amendments to the codes to promote greater clarity in their meaning and applicability. This cycle continued for each assessment item until rounds of coding emerged with few disagreements, each discussed by the authors until an agreement was achieved. In total, 539 inaccurate responses were collected and 196 (36.4\%) were coded in rounds by both researchers until the discrepancies were rare enough to suggest the coding scheme was reliable for application to the remaining 343 inaccurate responses.
This methodology was regularly evaluated by the researchers for trustworthiness, using frameworks presented in the literature base for qualitative education research. Primary considerations included the alignment of the methodology with the research questions and previous literature, the execution of data collection and analysis, and the contributions these data and the presentation of relevant findings may provide to the literature. As a last effort to ensure the trustworthiness of the findings presented, the findings were shared with two undergraduate researchers (who were not participants in the study). These students demonstrated strong academic performance in their experiences with general chemistry and served as peer-leaders and tutors at the research institution and were able to provide a students’ perspective for some of the challenges observed. Undergraduate researchers were given an overview of the methodology and deidentified samples of the dataset to summarize (these samples consisted of pastiches presenting three or more student responses grouped by code for each assessment item). Then, the students were given the results and discussion sections presented below and were asked to interject, argue, and debate their perspectives of the data where alignment was not achieved. The students were influential in this process and identified nuanced additions, contrapositives, and students’ perspectives of the data and implications for assessment design.

Results

The results presented below consist of students’ assessment response data and solution response processes related to the topics of mole concept, stoichiometry, and applied stoichiometry (e.g., stoichiometry used in the context of the gas laws). By-item statistics include the number of students who responded to the item (N), the percent by which students overall (P) selected the correct answer, and the standard deviations (SD) observed within each group demarcated by at-risk (AR) and not-at-risk (NAR). Below item statistics, the rationales for each of the distractors and the frequency of student selection (n and %) are provided for each group.
The mean differential between NAR and AR was calculated as the mean percentage of students in the not-at-risk student cohort subtracted by that of the at-risk cohort \((MD = NAR - AR)\). These statistics are then contrasted with qualitative descriptions of students’ challenges supported with participant-voiced pastiches comprised of exemplary student responses for each code color-coded via figure legend. Following the description of prevalent themes in the challenges observed amongst at-risk chemistry students, a qualitative comparison of those observed amongst the not-at-risk cohort is described.

**Challenges of At-Risk Chemistry Students**

**Interchangeable Chemical Identity**

At-risk chemistry students were commonly observed to interchange the chemical identity of a numerical value between different chemical species sometimes via common units such as mass or moles. An error of this quality was coded as “interchangeable chemical identity.” Take, for example, responses to the mole concept assessment item in which students were prompted to calculate the moles of oxygen in 15 grams of \(K_2SO_4\) (see Figure 22 on the next page). Two, accurate solution processes to the mole concept item were described by students as depicted below.

The Mole Ratio Process

\[
(15 \text{ g } K_2SO_4) \left( \frac{1 \text{ mol } K_2SO_4}{174.27 \text{ g}} \right) \left( \frac{4 \text{ mol } O}{1 \text{ mol } K_2SO_4} \right) = 0.34 \text{ mol } O
\]

The Mass Percent Process

\[
(15 \text{ g } K_2SO_4) \left( \frac{64.0 \text{ g } O}{174.27 \text{ g } K_2SO_4} \right) \left( \frac{1 \text{ mol } O}{16.00 \text{ g } O} \right) = 0.34 \text{ mol } O
\]
Figure 22. The mole concept interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

A presentation of interchangeable chemical identities was observed by students who divided the provided mass of K₂SO₄ —15 grams — by the molar mass of oxygen.

"Well, you'd set 15 grams of K₂SO₄ to proper conversion [sic], with the molar mass of oxygen on the bottom, set over the continuing equation…” (Erin)

Erin describes a conversion between 15 grams of potassium sulfate and “the molar mass of oxygen”. One plausible interpretation of this solution process is displayed below.

\[
(15 \, g \, K_2SO_4) \left( \frac{1 \, mol \, O}{16 \, g} \right)
\]

Whether Erin chose the atomic mass of oxygen or the molecular mass of oxygen (an error of consequence to be discussed), there was interchangeability in the chemical identity of the given mass and the atomic or molar mass included in the solution process. A student could perceive
this process as resulting in the cancellation of grams to arrive at moles of oxygen via dimensional analysis. A variety of responses were observed to share this interchangeability (see Figure 23) involving the mass or molar mass of oxygen applied to the mass or moles of potassium sulfate in solution processes.

| “mol of K₂SO₄ then multiply the molar mass of oxygen” (♀ Pamela) |
| “take 15g multiplied by 1 then divide by the mass of oxygen” (♂ Michael) |
| “You take the 15g and multiple that by the mass of the oxygen and divide by the mass of K₂SO₄” (♂ Darryl) |

Figure 23. Varieties by which students communicated interchangeability between numbers sharing the unit of mass but differing in chemical identity.

Much of the differential performance observed for the item shown in Figure 22, was attributed to the “skipped mole ratio” option (choice C) which matches the answer choice predicted by students who exemplified interchangeable chemical identity by using the molar mass of oxygen with the mass of potassium sulfate.

At-risk chemistry students were also observed to describe calculating the mass percent of oxygen in potassium sulfate and misattribute this value with the moles of oxygen in potassium sulfate as the two were near in numerical value for the assessment item (0.37 or 37% oxygen by mass and 0.34 mol of oxygen in potassium sulfate). Students’ responses shown in Figure 24 describe this inaccurate application of the mass percent algorithm and resulted in students either selecting the correct answer on the exam or presenting this process in their survey responses as what should have been enacted.
“First, we write out the formula for converting grams to moles, which is moles = grams/molar mass. We have the grams, we now need to find the molar mass. To do this, we need a periodic table to find the mass of K, S, and O. K is 39.1 grams, and there are two of them. S is 32 grams, with one in the compound. O is 16 grams, with 4 moles in the compound. This amounts to 173 grams. You then take the total oxygen molecules, which is 64 (16x4) and divide it by the total molar mass. That totals to 0.34 moles.” (Pete)

“You would have to divide the molar mass of oxygen (64) by the molar mass of K₂SO₄ (174.259) then multiply by 100%.” (Jan)

“I divided the amount of molar mass in O₂ by the total molar mass of K₂SO₄ and came up with the answer of 0.34.” (Angela)

“First you would find the mass of 4 oxygen and divide it by atomic mass of the equation. Then you would take that answer and convert it into moles by multiplying by 1/molar mass.” (Charles)

This interchangeability was characteristic of solution processes for all four assessment items and indicates a characteristic challenge in a student’s conceptual understanding of the meaning behind the numbers and units of a solution process. In the stoichiometry assessment item, students were tasked with calculating the mass of NH₃ produced from 45.6 g of H₂ and excess N₂ (see Figure 25 on the next page).

A predicted solution path for the item could be described as follows.

\[
\left(45.6 \text{ g } \text{H}_2\right) \left(\frac{1 \text{ mol } \text{H}_2}{2.016 \text{ g}}\right) \left(\frac{2 \text{ mol } \text{NH}_3}{3 \text{ mol } \text{H}_2}\right) \left(\frac{17.034 \text{ g}}{1 \text{ mol } \text{NH}_3}\right) = 257 \text{ g } \text{NH}_3
\]
The Stoichiometry Item

Calculate the amount of NH₃ gas produced by reacting 45.6 g of H₂ gas with excess N₂ gas according to the following chemical equation.

3 H₂ (g) + N₂ (g) → 2 NH₃ (g)

A) 85.6 g
B) 171 g
C) 257 g
D) 385 g
E) 579 g

Figure 25. The stoichiometry interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

Students’ processes involved the interchangeability of chemical identities sharing a common unit of measure (often mass, molar mass, or moles) as observed in their responses to other assessment items.

“Convert 45.6 g [of H₂] gives [sic] to moles by dividing it by the total mass of NH₂ [sic].

Then multiply the moles found by the 2 that are in hydrogen, and lastly convert your new moles back to ‘g’” (Kevin)

\[(45.6 \text{ g } H₂) \times \left(\frac{1 \text{ mol } NH₂}{g \text{ NH₂}}\right) \times (2 \text{ mol } H₂)\]

Here the interchangeability between H₂ and “NH₂” seems reinforced by both the cancellation of grams and the quality of NH₂ as containing molecular hydrogen. Kevin’s response was coded as both interchangeable chemical identity and an error in the interpretation of chemical symbolism wherein the identities of elements capable of bonding as diatomic molecules are inaccurately attributed molecular or elemental forms.
Atomic vs. Molecular

The code labeled “atomic vs. molecular” was most pronounced where the comprising elements of a compound were capable of forming a diatomic molecule and were involved in a component of students’ solution processes. In Figure 26, responses to the mole concept assessment item reflect this challenge with representational competence.

“15g K₂SO₄ * (1mol K₂SO₄/174g K₂SO₄) * (4 mol O₄/1mol of K₂SO₄)=” (mland James).

“You convert 15g to moles of K₂SO₄ and get .086 then use that and multiply it by 4 for the O₄” (mland Ryan)

“Take 15 grams and multiply by 1mol K₂SO₄/174.27, then use molar ratio to find the moles of O₄.” (mland Toby)

Figure 26. Returning to the mole concept assessment item, the code “atomic vs. molecular” was prevalent.

Here, students tend to represent 4 mol of oxygen atoms in potassium sulfate as either O₄ or O₂. One plausible interpretation of these errors in the solution process could concern the stem of this item. Students could have misinterpreted “moles of oxygen” with “moles of molecular oxygen” and thereby selected “D” as a plausible response to the item. Only twenty-nine students (4.8%) selected answer choice “D” discounting this as a common interpretation of the question but the question stem would be improved by specifying “oxygen atoms”. As shown in Figure 26, students commonly misrepresented oxygen as O₄. James and Ryan still utilized a coefficient of 4 with O₄ and thus this misconception did not prevent their selecting the correct answer to the assessment item. Inaccurate solution processes that arrive at correct answer choices could reinforce students’ misconceptions and students’ future interpretations of chemical symbolism could be negatively impacted as a result.
The Role of Molar Mass

Save for the mole concept assessment item, all other stoichiometry-related assessment items involve the application of stoichiometric coefficients interpreted from a balanced chemical reaction to solve an assessment item. For the stoichiometry assessment item in Figure 25, students struggled in their attempts to discern the masses, moles, molar masses, and stoichiometric proportions between chemical species and the roles these values serve within systematic approaches. One manner by which this challenge manifested was with confusion related to the use of coefficients from a balanced chemical reaction in the determination of molar masses (see Figure 27).

"First we have to find the amount of each compound. $3 \text{ H}_2 = 6 \text{ g}$, $\text{N}_2 = 28 \text{ g}$, and $2 \text{ NH}_3 = 34 \text{ g}$. We know that $6 \text{ g of H}_2 = 34 \text{ g of NH}_3$, so we have to find how the amount of $45.6 \text{ H}_2$. To find the mass of NH$_3$ produced we have to multiply $45.6 \times 34$ and divide it by 6 and we get $257 \text{ g}$ which is the correct answer." (👤 Stanley)

"Took $45.6 \text{ g}$ and multiplied it by 1mol over $6\text{ g of H}_2$. Then multiplied that by 2 moles of NH$_3$ over 1 mole of H$_2$. Then [sic] multiplied that by the total grams of 2 NH$_3$ and got 34. As a result, acquired 516. Then [sic] divided that by the 2 moles of NH$_3$ and got roughly 258. Not the correct way, just how I did it. Will use the correct way next time. I think I just lucked out with this answer…” (👤 Meredith)

"$45.6 \text{ g H}_2 \times (3\text{mol}/2.016\text{g H}_2) \times (1 \text{ N}_2/1 \text{H}_2) \times (1\text{mol}/28.02)……… and that's where I stopped because it was incorrect. I don't know how to get to the correct answer :( I cant [sic] figure it out" (👤 Holly)

As observed in all three student elaborations presented in Figure 27, the role of a coefficient was unclear. Reflected by Stanley and Meredith’s responses, students were
observed to apply stoichiometric ratios to their calculations of the molar mass of a compound. At-risk students manifested this misconception as to the utility of coefficients and were observed to achieve the correct answer via an inaccurate conception of the algorithm. There were other solution processes, however, wherein at-risk students applied stoichiometric coefficients to molar masses in a manner that did not result in the attainment of a correct answer (see Holly’s response). The stoichiometry assessment item presented with a low mean differential ($MD = 8.5\%$) and was observed to attract $11.3\%$ of the students to answer choice B in Figure 25. The rationale of this distractor reflected students’ use of coefficients in the balanced chemical reaction to calculate the molar mass of a substance (e.g., $H_2$ would have, not $2.016 \text{ g/mol}$, but $6.048 \text{ g/mol}$ following the multiplication of the molar mass by a coefficient of 3 from the balanced chemical reaction). This distractor was responsible for $4.5\%$ of the overall $8.5\%$ mean differential observed.

Challenges with the interpretation and use of molar mass were also observed throughout the assessment items on the second interim exam. The gas law item provides the values necessary to compute moles of molecular hydrogen gas using the ideal gas law and asks students to determine the moles of ammonia gas produced given excess molecular nitrogen gas (see Figure 28 on the next page).

The provided chemical reaction was the same as that of the stoichiometry item from the first interim test where a predicted solution path for the item could be depicted as follows.

$$n = \frac{(0.920 \text{ atm})(126 \text{ L})}{(0.08206 \text{ L} \cdot \text{atm} \over \text{mol} \cdot \text{K})(15 + 273.15 \text{ K})} = (4.90 \text{ mol } H_2) \left( \frac{2 \text{ mol } NH_3}{3 \text{ mol } H_2} \right) = 3.27 \text{ mol } NH_3$$
Figure 28. The gas laws interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.

Students were observed to calculate the moles of H₂ with little difficulty. Their ability to algebraically manipulate the ideal gas law equation for moles appeared unencumbered by having scored in the bottom quartile of math aptitude scores. Challenges arose when converting the moles of H₂ gas calculated to moles of NH₃. Often, this difficulty involved determining whether molar mass plays a role in discerning a final answer as demonstrated in Figure 29 (on the next page).

In parallel, the changes in energy assessment item is predicted to have a solution path that does not inherently require the use of the molar mass for ammonia, yet students demonstrate the use of this value as a prominent point-of-confusion. For this item, students are provided the following chemical reaction and heat of reaction (see Figure 30 on the next page).
"First, I found the molar mass of NH₃ 17.02 then divided the number of mols (4.902 mols) calculated by PV=nRT and found that there is 3.27 mol of NH₃." (Angela)

"Using PV=nRT formula, First you would first you would find the N of N₂. Once finding number of moles for N₂= 4.902, you will then find moles by getting molar mass of NH₃ (N)= 14.007 + (H)3= 1.008*3. Moles to grams, grams to moles to find NH₃ which = around 3.27 to 3.4 moles or n." (Darryl)

"They did not take into account the H₂ in the equation. You have to do another conversion using the molar mass. You have to use the molar mass of H₂... I am not exactly sure where or what to do after that." (Gabe)

"I didn’t know how to arrive to the final answer to be honest, i tried to find what what was missing, but i cant determine, i think mostly i got confused about determine which molar mass i should use in this question." (Robert)

**Figure 29. Misapplications of molar mass in response to the gas law assessment item.**

<table>
<thead>
<tr>
<th>The Changes in Enthalpy Item</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consider the following balanced chemical equation:</td>
</tr>
<tr>
<td>4 NH₃(g) + 5 O₂(g) → 4 NO(g) + 6 H₂O(l)</td>
</tr>
<tr>
<td>ΔH°rx = +1168 kJ</td>
</tr>
<tr>
<td>How much heat is absorbed/released when 2.05 mol of NH₃(g) reacts with excess O₂(g) to produce NO(g) and H₂O(l)?</td>
</tr>
<tr>
<td>A) 5.99 x 10² kJ of heat is released.</td>
</tr>
<tr>
<td>B) 5.99 x 10² kJ of heat is absorbed.</td>
</tr>
<tr>
<td>C) 2.40 x 10² kJ of heat is released.</td>
</tr>
<tr>
<td>D) 2.40 x 10² kJ of heat is absorbed.</td>
</tr>
<tr>
<td>E) 1.02 x 10² kJ of heat is absorbed.</td>
</tr>
</tbody>
</table>

**Figure 30. The changes in energy interim assessment item profile comparing the performance of not-at-risk (NAR) and at-risk (AR) chemistry students.**
Students were asked to calculate the amount of heat absorbed or released when 2.05 mol of NH\textsubscript{3}(g) reacts with excess O\textsubscript{2}(g) to produce NO(g) and H\textsubscript{2}O(l). A theoretical solution path to the item is depicted below.

\[
(2.05 \text{ mol } \text{NH}_3) \left( \frac{+1,168 \text{ kJ}}{4 \text{ mol } \text{NH}_3} \right) = +559 \text{ kJ}
\]

Here, confusions over the application of molar mass arose once again (see Figure 31).

"First you must \underline{find the moles of NH}_3 by dividing the mass of NH\textsubscript{3} by the molecular weight. Next, you want to find out how much heat is absorbed during ONE NH\textsubscript{3}, so we divide 1168 kJ by 4. You then take that number and multiply it by the moles to get 5.99 \times 10^2 kJ absorbed." (👤 Pete)

"I believe you would need to find \underline{grams per mole} of NH\textsubscript{3} then multiply that by the heat absorbed given and also divide the heat absorbed by the \underline{2 moles of NH}_3 in the equation, not just one mole." (👤 Phyllis)

"I got the answer by first \underline{finding the molar mass of NH}_3, multiply 17.03 g by 2.05 which = 34.9g, then I multiplied the 4 NH\textsubscript{3} by the molar mass of NH\textsubscript{3} (17.03 \times 4) = 68.12, then divide 34.9/68.12 = 0.512, 0.512 \times 1168 \text{ kJ} = 5.99 \times 10^2 \text{ kJ} (👤 Angela)"

"for me i didn't know how to get to the final answer yet, i was so confused. i think the error was in the \underline{molar mass, but all other number were in the right place." (👤 Robert)

"2.05 mol NH\textsubscript{3}(17.04 \text{ NH}_3/1 \text{ mol NH}_3)(4 \text{ mol NH}_3/1168 \text{ kJ})=5.99 \times 10^2

You multiply what is given by the \underline{molar mass of NH}_3, then you multiply that by the moles and divide it by kilojoules and get your final answer." (👤 Nellie)"
Solution processes similar to that of Pete and Angela (Figure 31 on the previous page) would result in an accurate response via an inaccurate conception of the process.

Our interpretation of Angela’s solution process is illustrated below and may demonstrate the degree to which students rely on algorithmic solution processes without conceptual understanding.

\[
2.05 \text{ mol } NH_3 \left( \frac{17.03 \text{ g}}{1 \text{ mol } NH_3} \right) = 34.9 \text{ g} \left( \frac{34.9 \text{ g}}{17.03 \text{ g}} \right) = 0.512 (1168 \text{ kJ}) = 5.99 \cdot 10^2 \text{ kJ}
\]

Angela’s application of the molar mass of ammonia twice was, presumably, to execute an algorithm by which the mass of NH\textsubscript{3} was required. This difficulty in determining whether or not to use the molar mass may be indicative of the difficulty associated with distinguishing moles of one chemical species from moles of another. Alternatively, the challenge could be related to the use of the moles of NH\textsubscript{3} to reconcile the connection between energy and mass or moles. The persistence of this challenge was commonly observed (see responses authored by Angela and Robert in Figures 29 and 31). These difficulties were most notable in the changes in energy assessment item as students struggled to differentiate between the stoichiometric coefficient (e.g., the 4 mol of NH\textsubscript{3} required for the balanced chemical reaction) and the amount in moles of ammonia consumed in the chemical reaction (see responses by Pete and Phyllis in Figure 31).

"In Excess" as a Source of Confusion

Inaccurate solution processes of at-risk chemistry students for the applied stoichiometry assessment items also involved the incorporation of chemicals prompted as “in excess” suggesting that this phrase did not inform students as to whether a theoretical yield or
discerning the limiting reactant algorithm should be executed. Exemplar responses demonstrate how students invoke algorithms that seem to emphasize the conservation of matter using excess reactants (Figure 32) or include excess reactants in the determination of theoretical yield (Figure 33).

“I think you would need to do 1 mol of N₂ times its molar mass and add that to 1 mol of H₂ and times that by its mass and convert it to liters and then plug it into the equation PV=nRT” (Charles)

“I don't know what I could do to get a. I tried converting the moles mass of NH₃ to liters to get the volume of NH₃ but that was too high, I tried getting a difference of the mass of NH₃ and H₂ but that gave 11.13 which is not an answer. I cannot understand why it is A.” (Kevin)

“Add the total moles of NO(g) and H₂O(l) which is 10. Then use the given moles of NH₃ as conversion factors. So…” (Roy who also responded with the equation below)

\[
2.05 \text{ mol } NH₃ \times \frac{10 \text{ mol NO and } H₂O}{4 \text{ mol } NH₃} \times \frac{1168 \text{ kJ}}{1 \text{ mol NO and } H₂O} = 5,986 \text{ kJ}
\]

“In order to arrive at the correct answer you must multiply 2.05 with 1168 in order to get the kJ once you get that you must add in the oxygen in order to finalize the reaction.” (Jo)

“I think you would have to take 4 mols of NH₃ and divide it by the 2.05 mols and then take that and add the 5 mols of O₂ and then multiply the total by 1168 kJ. I am not positive though because I got that one wrong on the test…” (Charles)

Figure 32. Interchangeable chemical identities manifested via uncharacteristic solution processes involving excess reactant to the gas laws (first two responses) and changes in energy (last three responses) assessment items.

Charles, an at-risk student whose solution processes are shown for both the gas laws and changes in energy items in Figure 32, demonstrates persistent difficulty with discerning whether or not to implement excess reactants toward arriving at an answer. “In excess” is a phrase commonly observed amongst the assessment items related to stoichiometry. The phrase is
intended to communicate an abundance of a reactant suggesting one need not consider the reactant present in excess as limiting to the theoretical yield of a chemical reaction. None of the four assessment items require students to discern the limiting reactant in a calculation of a produced mass, challenges with limiting reactants were not expected to emerge within students’ responses, yet challenges were prevalent amongst the stoichiometry, the gas laws, and changes in energy assessment items. Often, students included the excess reactant in their calculations (see Figure 33).

“To calculate how much NH₃ is present the individual will need to find how many moles of H₂ and N₂ are present in the problem. After solving for both, a molar ratio can be set up to solve for the amount of NH₃ produced. After which when moles of NH₃ are found one will need to multiply the moles of NH₃ by its molar mass because the answer is required to be in grams.” (Clark)

“I didn’t guess the right answer, but what i did was i multiply 45.6 by the the mole of H₂/ molar mass of H₂, then i multiply that by molar mass of N₂/1mole of N₂.” (Robert)

“So manipulate the equation to get n on one side and it becomes PV/RT. Plug in values and you get 4.90. I’m assuming this isn’t the final answer because you’re asked for excess N₂ gas and I am not sure how to get there or apply that to the problem. Thankfully, this is what studying is for…” (Meredith)

“Personally, I got answer B, and I think I got it wrong, because I did not factor in the excess amount of N₂. I am not sure how to arrive to the correct answer.” (Sam)

“2.05 moles needs to be converted since its reacted with excess oxygen. convert 2.05 mol of ammonia to moles of oxygen using mol to mol ratio and then use delta H= q/mol of oxygen. no, because I did the equation with the 2.05 moles given.” (Katy)

“The correct conversion factor would have been:
2.05 mol NH₃ x 5 mol O₂\4 mol NH₃ x 1168kJ/5 mol O₂ giving us 5.99x10² and since its positive we know it was absorbed.” (Oscar)

Figure 33. Incorporating a limiting reactants algorithm for stoichiometry (Clark and Robert), the gas laws (Meredith and Sam), and changes in energy (Katy and Oscar) assessment items describing all other reactants as in excess.
Manifestations of these errors are presented in each pair of exemplary student responses in Figure 33 representative of the stoichiometry, gas laws, and changes in energy assessment items, respectively.

Each response in Figure 33 demonstrates confusion as to how to incorporate the “excess” reactant, which was commonly observed throughout the students’ elaborations. If the instructors intend to measure proficiency with associating an abundant reactant as having no impact on the theoretical yield of a chemical reaction in addition to solving gas law stoichiometry or energy changes of a chemical reaction, then these assessment items are well placed. It may be more plausible that the intent was to include “excess” as a means of eliminating the additional challenge of discerning limiting reactants within the context of the stoichiometric calculations. This simplification does not seem to correspond with students’ elaborations of the item in which they indicate confusion over the phrase and attempt incorporation within the algorithms by which problems are solved.

A Comparison of the Challenges Observed between Not and At-Risk Chemistry Students

In evaluating potential sources for differential performance, the prevalence of challenges with these items were considered with respect to each code (see Table 14 on the next page). Overall, the most common challenges observed involve interchangeable chemical identity, unit conflation, and the incorporating excess reactants. Of the students struggling with these concepts, the challenges and the prevalence by which they affect students does not seem to explain the differential performance observed. In the process of qualitatively reviewing student responses, it was clear that some students were able to arrive at a correct answer for the prompt via a chemically implausible solution process. For example, the exemplar responses of at-risk students in Figure 33 demonstrates an error wherein students equal the moles of oxygen with the mass ratio of oxygen in potassium sulfate to arrive at a correct answer for the mole concept assessment item via an incorrect process.
Table 14. Codes (by theme) with prevalence measured as the percentage each code comprises of the total number of inaccuracies identified in students’ solution processes (NAR = 393, AR = 234).

<table>
<thead>
<tr>
<th>Code Descriptions (by theme)</th>
<th>Representational Competence</th>
<th>Systematic Approaches</th>
<th>Conceptual Understanding</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Atomic vs. Molecular</strong></td>
<td>Atoms comprising species are misrepresented as covalent compounds using subscripts (e.g., 4 O in potassium sulfate vs. O₄).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NAR: 4%</td>
<td>AR: 3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Misrepresented Proportions</strong></td>
<td>Subscripts and/or coefficients of a chemical formula are misapplied.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>4%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Elemental Symbols</strong></td>
<td>Atomic masses are incorrectly assigned.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1%</td>
<td>1%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nomenclature</strong></td>
<td>Chemical symbols are misinterpreted when determining the predominate state of matter or name of a chemical species.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Inaccurate Conversion Factors</strong></td>
<td>Selected conversion factors between units of measure (e.g., mass to moles; moles to mass) are incapable of resulting in a viable measure for the converted value (e.g., converting mass to moles using Avogadro’s number).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>5%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Misapplied Molar Mass</strong></td>
<td>Molar masses are used incorrectly and/or are unnecessary for the process described.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Conversion Arrangement</strong></td>
<td>Numbers and units are arranged improperly result in units incapable of achieving a coherent number and unit of measure (e.g., 12.1 moles O multiplied by 1 mol O/16.0 g O to achieve mass of O).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5%</td>
<td>7%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Sequence</strong></td>
<td>Steps in the predicted solution path(s) are present but are rearranged in a manner that does not reflect a solution path that is chemically plausible.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Coefficient and Molar Mass</strong></td>
<td>Stoichiometric coefficients are incorrectly incorporated in the calculation of molar mass (e.g., mass of H₂ with a coefficient of 2 is equal to 2*2.0158 g/mol or 2.0158 g/mol divided by 2).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>± Moles or Mass</strong></td>
<td>Often used for seemingly uncharacteristic solution pathways, the sum of some values of reactants to equal that of the products is used.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3%</td>
<td>3%</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interchangeable Chemical Identity</strong></td>
<td>Conversions between numerical values representative of different chemicals are conducted demonstrating a lack of identity reflective of the interrelation of one chemical to another (e.g., 15 g of K₂SO₄ divided by atomic mass of oxygen to find moles of oxygen).</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19%</td>
<td>15%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 14 (continued)

<table>
<thead>
<tr>
<th>Interchemical Unit Identity</th>
<th>[\text{Moles of a compound are not distinguishable from moles of a constituent element.}]</th>
<th>1%</th>
<th>0%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unit Conflation</td>
<td>[\text{Differing units of measure are used interchangeably (e.g., molar mass as mass).}]</td>
<td>11%</td>
<td>14%</td>
</tr>
<tr>
<td>Intrachemical Unit Identity</td>
<td>[\text{Terminologies inherent to particular units are misapplied (e.g., using the term “the number of atoms” of a chemical to describe the moles of the same chemical).}]</td>
<td>8%</td>
<td>9%</td>
</tr>
<tr>
<td>Incorporating Excess Reactants</td>
<td>[\text{Despite the expected cue of “excess” in the prompt of the item task, solution processes are described that include determining the limiting reactant or the inclusion of the excess reactant in the solution process.}]</td>
<td>13%</td>
<td>11%</td>
</tr>
<tr>
<td>Distinct Measures</td>
<td>[\text{Conversions between numerical values representative of the same chemical, a lack of identity toward conversions that would be reflective of the interrelation of one unit to another is demonstrated (e.g., equating the moles of oxygen in 15 g K_2SO_4 with the mass percent of oxygen in 15 g of K_2SO_4).}]</td>
<td>6%</td>
<td>9%</td>
</tr>
</tbody>
</table>

Responses were coded for “incorrect process, correct answer” if the student was observed to apply unnecessary (e.g., converting the moles of \(\text{NH}_3\) provided in the changes in enthalpy item into grams and back into moles) or chemically implausible (as described above with regard to the difference between mass ratio and moles of a constituent element) steps within their solution processes. Students who conflated units of measure (e.g., describe the molar mass of a substance) or referred to chemicals by inaccurate nomenclatures (e.g., describing \(\text{NH}_3\) as NH) but otherwise were observed to arrive at a correct answer were not included in this measure as these students merely represented aspects of their processes inaccurately and did not execute inaccurate processes to arrive at a correct numerical answer.

The quantitative results of these codes can be found in Table 15 (on the next page). Qualitatively, each assessment item presented a unique variety of pathways by which student arrive at accurate numerical answers via inaccurate processes.
Table 15. Code prevalence for students who arrived at accurate numerical values for the assessment items implementing an inaccurate process.

<table>
<thead>
<tr>
<th>Assessment Item</th>
<th>NAR</th>
<th>AR</th>
<th>MD (NAR - AR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mole concept item</td>
<td>24%</td>
<td>15%</td>
<td>9%</td>
</tr>
<tr>
<td>Stoichiometry item</td>
<td>15%</td>
<td>13%</td>
<td>2%</td>
</tr>
<tr>
<td>The gas laws item</td>
<td>13%</td>
<td>4%</td>
<td>9%</td>
</tr>
<tr>
<td>Changes in enthalpy item</td>
<td>31%</td>
<td>21%</td>
<td>10%</td>
</tr>
</tbody>
</table>

*The Mole Concept Item*

Exemplar responses of not-at-risk chemistry students for the mole concept item wherein students arrive at a correct numerical answer via an incorrect process (Figure 34).

How many moles of oxygen are in 15 g of K₂SO₄?
A) 0.022 mol   B) 0.043 mol   C) 0.086 mol   D) 0.17 mol   E) 0.34 mol

“what you do is divide the 15 grams given by the mass of K₂SO₄. then you would multiply the mass of O₂ (64 grams) by 15/174.27. you should get around 5.5 grams. then you divide 5.5 by the mass of 1 mole of oxygen and you would get 0.34. ((15/174.27)x64)/16=0.34” (👨‍🏫 Ricky)

“You first convert 15g of K₂SO₄ to its molar mass and convert that to the mole to mole ratio, altogether you should have (15x4)/174.27” (👨‍🏫 Alex)

“To get the correct answer you take the grams of O (16) and times that by 4 over the molar mass of K₂SO₄.

16*4
--------
174.27”
(👨‍🏫 Tim)

Figure 34. Common and chemically implausible solution strategies along with representational competence challenges observed amongst the not-at-risk chemistry students.
Here, Ricky and Alex describe their solution processes following their descriptions with the calculations by which they arrived at their answer to the prompts. Ricky adopts the mass percent solution path incorporating the mass of 4 moles of oxygen atoms (64 g). The process can be summarized as follows:

1. First, convert the provided mass of K$_2$SO$_4$ to moles by dividing the molar mass of the compound,
2. Then, multiply the moles of K$_2$SO$_4$ by the mass of 4 moles of oxygen atoms,
3. Finally, divide by the molar mass of 1 mole of oxygen atoms.

The step in Ricky’s solution process involving the multiplication of the moles and mass of two different chemical compounds demonstrates the conceptual challenge of interchangeable chemical identity. The instances in which students present all the steps of a predicted solution path but rearrange the order of these processes to one that is chemically implausible was coded in the Figure 34 as an error in sequence. Ricky, thereby, attains a solution to the assessment item that would be marked correct on the exam but was arrived at via a chemically implausible solution process that reflects a lack of conceptual understanding as to the meaning of the numerical values used in the algorithm.

Alex also demonstrates a similar error via the predicted mole ratio path. Alex’s solution process involves the multiplication of 15 grams of K$_2$SO$_4$ by the mole-to-mole ratio of 4 mol of oxygen atoms to 1 mol of potassium sulfate. The use of chemically implausible processes that result in the correct answers was common amongst not-at-risk chemistry students in solving the mole concept item (24% or 21 out of 86 inaccurate responses for this item). This compared to that of at-risk chemistry students who arrive at correct answers via inaccurate processes for the mole concept item (15% or 8 of 52 inaccurate responses) explains some of the differential performance observed between these two groups of students.
The Stoichiometry Item

One hundred fifty-four inaccurate solution processes were presented amongst the students (100 NAR, 54 AR) for the stoichiometry assessment item. While not-at-risk and at-risk cohorts manifested similar challenges with their solution processes, multiplying coefficients by the molar mass of chemicals and rearranging a chemically plausible sequence for the solution processes were two common routes by which students arose at a correct answer via an incorrect process (see Figure 35).

Calculate the amount of NH₃ gas produced by reacting 45.6 g of H₂ gas with an excess of N₂ gas according to the following chemical equation.

\[ 3 \text{H}_2(g) + \text{N}_2(g) \rightarrow 2 \text{NH}_3(g) \]

A) 85.6 g  B) 171 g  C) 257 g  D) 385 g  E) 579 g

“you do the molar mass of 2NH₃ divided by the molar mass of 3H₂ and multiply that by the given value of 45.6 of H₂.

Your final answer should be 258.4, which is close to C.” (Oliver)

“\( (45.6 \text{ grams of H}_2/2.018) (17.02 \text{ of NH}_3)(2/3)=257 \)
2.018=molar mass of H₂
17.02=molar mass of NH₃
2/3 is the ratio” (Trudy)

“The correct answer:

\( 45.6/2.02*17.04/1*2/3= 257 \text{g} \) with 2.02 the molar mass of H₂ and 17.04 the molar mass of NH₃.” (Donna)

“I still don’t fully understand what I’m doing wrong but I think I came up with a way to get 257 g.

What I did was take the molar mass of 2NH₃ which was 34 g and I multiplied it by 45.6 g of H₂ and then divided it by 6 g of H₂ which gave me 257 g.” (Shiela)

“You would divide 45.6 by the molar mass of H₂ and convert this number to the moles of NH₃. Then this number would be multiplied by two. The same process must be done with N₂ and this will give a limiting factor. We will then subtract the excess from the limit, giving us 257g.” (Ben)

Sequence  Interchangeable chemical identity

Coefficients and Molar Mass  Incorporating Excess Reactants

Figure 35. Inaccurate conceptions of algorithms communicated by not-at-risk chemistry students that result in accurate responses to the stoichiometry assessment item.
Sequence (or the rearrangement of a solution process that would otherwise be accurate) was observed as a common method by which students arrived at an accurate value utilizing a solution pathway that does not attribute meaning or identity to the numerical values involved. Consider the depicted solution process designed to reflect an error made by Trudy and Donna.

\[
(45.6 \text{ g } H_2) \left( \frac{1 \text{ mol } H_2}{2.02 \text{ g}} \right) \left( \frac{17.03 \text{ g}}{1 \text{ mol } NH_3} \right) \left( \frac{2 \text{ mol } H_2}{3 \text{ mol } NH_3} \right) = 257 \text{ g } NH_3
\]

While an accurate value is acquired, students apply the mole ratio following the multiplication of the attained moles of H₂ by the molar mass of NH₃ demonstrating both a lack of chemical identity to these measures. Another inaccurate solution path (as represented here by Oliver), resulted in accurate response via the incorporation of coefficients in the calculation of molar mass. The process functionally removes unit and chemistry identity for the numerical values used in the calculation.

\[
(45.6 \text{ g } H_2) \left( \frac{2 \cdot 17.03 \text{ g}}{3 \cdot 2.02 \text{ g } \text{ mol } H_2} \right) = 257 \text{ g } NH_3
\]

The lack of order to not-at-risk students’ algorithms suggest that their conceptual understanding of these algorithms may not be more advanced than their at-risk peers, but rather not-at-risk students are more familiar with the execution of a predictable series of calculations with little meaning attributed to the numerical values that arise. A lack of order relegates the process of sequential unit cancelation as the primary means by which to solve a prompt, regardless of the chemical plausibility of a solution processes. These occurrences serve as another example by which items designed to assess a proficiency with attaining a final numerical answer have a potential to reward chemically implausible algorithm execution.
The Gas Laws Item

Responses to the gas laws item demonstrate additional manifestations of not-at-risk students employing algorithms that arrive at the correct numerical value but fail to demonstrate a chemically plausible solution process (see Figure 36).

If 126 L of H₂(g) at 15°C and a constant pressure of 0.920 atm reacts with excess N₂(g) as shown below, how many moles of NH₃(g) are produced?

<table>
<thead>
<tr>
<th>Answer</th>
<th>Moles of NH₃</th>
</tr>
</thead>
<tbody>
<tr>
<td>A) 3.27 mol NH₃</td>
<td>B) 4.90 mol NH₃</td>
</tr>
</tbody>
</table>

A) 3.27 mol NH₃
B) 4.90 mol NH₃
C) 7.36 mol NH₃
D) 55.7 mol NH₃
E) 84.0 mol NH₃

“I think what had to be done was, yes plug everything into the formula, where the results equal the mistaken answer of 4.9. But then do not stop there; multiply that by (2/3), number of moles of NH₃ (2), over moles of H₂ or something’s molar mass. But either way 4.9 times 2/3 gives you 3.26667, which is close to 3.27, answer A. I just am not so sure where the 2 and 3 being multiplied come from exactly.”

( Rachel)

“You must use PV=nRT to figure out this question. There is a molar ratio of 3:2 which you must use in order to figure out how many liters of NH₃ is produced. After setting up the molar ratio, you get the value of 84, which you must then plug into PV=nRT. Plugging in all variables would result in: $\frac{84 \times 0.920 \times 0.08206 \times 288}{3}$, which would give us 3.27 mols of NH₃.”

( Malcolm)

“there are only two hydrogen atoms in H₂, but 3 in NH₃, so you have to multiply 4.9 by 2/3 to get the correct answer”

( Karen)

“I think what had to be done was, yes plug everything into the formula, where the results equal the mistaken answer of 4.9. But then do not stop there; multiply that by (2/3), number of moles of NH₃ (2), over moles of H₂ or something’s molar mass. But either way 4.9 times 2/3 gives you 3.26667, which is close to 3.27, answer A. I just am not so sure where the 2 and 3 being multiplied come from exactly.”

( Oliver)

First you must convert the 126L of H₂ to L of NH₃. You would take the 126L of H₂ x

\[
\frac{1 \text{molH}_2}{1 \text{LH}_2} \times \frac{2 \text{molNH}_3}{3 \text{molH}_2} \times \frac{1 \text{LNH}_3}{1 \text{molNH}_3} = 84.
\]

Then you would plug the numbers into $n= \frac{PV}{RT}$, $n= (0.920 \times 84)/(0.08206 \times 288)$ $n= 3.27 \text{mol NH}_3$”

( Phillip)

While both groups of chemistry students were described as commonly interchanging the chemical identities of numerical values (e.g., moles of H₂ vs. moles of NH₃ in the context of the
gas law assessment item, not-at-risk students who err in their solution processes tended to apply proportions used to interconvert chemical species in a variety of ways. For examples, the not-at-risk student responses in Figure 36 refer to the conversion of 126 L of H₂ to 84 L of NH₃ as the conversion necessary to relate molecular hydrogen and ammonia rather than applying a mole-to-mole ratio. In the case of students like Rachel, Malcolm, Karen, and Phillip, students demonstrate a conceptual understanding of chemical identity by applying a mole ratio in their solution processes but seem to assume that temperature is held constant as the reaction proceeds and thus the mole ratio can be applied to the volume.

A similar variety of this code can be observed in Karen’s solution process. Here, the mole ratio is not derived from the coefficients of the balanced chemical reaction but rather the subscripts of the formula wherein moles of hydrogen atoms in ammonia are 3 and in molecular hydrogen are 2. However, this error in the context of the problem reflects the challenges all students face with interpretation chemical symbolism in addition to conceptualizing how the depicted quantities interrelate. Of the inaccurate responses observed from not-at-risk students 13% arrived at a correct answer via a chemically implausible solution processes wherein only 4% of at-risk chemistry students were observed to do the same. These responses suggest a shortcoming in conceptual understanding of the processes by which answers are obtained.

The Changes in Energy Item

Students solving the changes in energy assessment item in Figure 30 appeared more comfortable converting the 2.05 mol of NH₃ provided into mass and then solving the problem (even recalculating moles) perhaps as a means to rationalize their responses (see Figure 37 on the next page). A solution process provided by a not-at-risk student reflects that students could be aware of the difference between the execution and conceptual understanding of a solution process:
“To arrive at the correct answer, you would divide 2.05 moles of NH₃ by 4 moles of NH₃. From there you would multiply by 1168 and you would come up with your answer of 5.99 ✕ 10². You would be able to distinguish that the heat is absorbed because the delta h reaction is positive. When I actually look at this problem, I can numerically see what I need to do to arrive at the correct answer, however I do not believe that I am conceptually understanding it at the moment.” (Jon)

Consider the following balanced chemical equation:

4 NH₃(g) + 5 O₂(g) → 4 NO(g) + 6 H₂O(l)  \quad \Delta H_{rxn} = +1168 \text{ kJ}

**How much heat** is absorbed/released when 2.05 mol of NH₃(g) reacts with excess O₂(g) to produce NO(g) and H₂O(l)?

A) 5.99 ✕ 10² kJ of heat is released.  B) 5.99 ✕ 10² kJ of heat is absorbed.
C) 2.40 ✕ 10³ kJ of heat is released.  D) 2.40 ✕ 10³ kJ of heat is absorbed.
E) 1.02 ✕ 10⁴ kJ of heat is absorbed.

“First convert, 2.05 moles of NH₃ in grams, which is around 34.91 grams of NH₃. Plug it into formula; 34.91 times 1 mol over molar mass of NH₃, which is 17.04, and then multiply that by the +1168 kJ divided by 4, which is the moles or coefficient from NH₃. Doing all the math of that gives us 598.22, which is close to answer choice B above. (correct answer) And we know it is absorbed heat.” (Oliver)

“I got the answer by first finding the molar mass of NH₃, multiply 17.03 g by 2.05 which = 34.9g then I multiplied the 4NH₃ by the molar mass of NH₃ (17.03*4) = 68.12 then divide 34.9/68.12 = 0.512
0.512 * 1168 kJ =5.99x10² kJ” (Brenda)

“The correct conversion factor would have been:
2.05 mol NH₃ x 5 mol O₂/4 mol NH₃ x
1168kJ/5 mol O₂ giving us 5.99x10² and since its positive we know it was absorbed.” (Devona)

“Convert the moles to kJ and then multiply by the coefficient of O₂…” (Lee)

---

Figure 37. Inaccurate conceptions of algorithms communicated by not-at-risk chemistry students that result in accurate responses to the changes in energy assessment item.
Similarly to the other assessment items, 31% of the not-at-risk students arrived at a correct answer via a chemically implausible solution process (compared to 21% observed amongst not-at-risk students).

Oliver’s responses were chosen as exemplary for the last three figures as conceptual challenges with molar mass and its application was a common challenge observed throughout students’ solution processes. This student demonstrates how some not-at-risk students successfully develop algorithms that, from the perspective of an instructor may demonstrate little conceptual proficiency yet result in accurate responses on multiple-choice exams designed to measure students’ ability to attain a numerical response. Further, Oliver’s repetition of this chemically implausible algorithm from test 1 to test 2 reflects the concern that assessment items designed to measure a student’s ability to attain a numerical value via a step-by-step process could encourage and reward conceptually implausible processes.

Discussion

Challenges observed amongst at-risk chemistry students commonly related to conceptualizing the unit and chemical identities of numerical values in the context of their solution processes. Descriptions of students’ challenges with the mole concept and stoichiometry in past works used as the conceptual framework for this study were summarized into themes of representational competence, systematic approaches, and conceptual understanding. Within these works, these themes often appear as three distinct challenges for the students; however, the authors posit a more pluralistic view of the challenges students face with the moles concept and stoichiometry. All three often manifested in the erred processes of students wherein students were observed to struggle with the conceptualization of the processes used to solve the problems rather than any one of the themes presented in the literature.
For example, students were often observed to describe molar mass as mass and applying gram units to these values but were able to apply these values properly to achieve an accurate and chemically plausible answer to the problem (see Chris, David, and Gareth’s responses to the mole concept prompt). In another research study, the impact of terminology and conflicting descriptions were thought to result from a lack of a consistent conception of the mole as presented in textbooks and instruction. It is possible that students’ challenges to conceptualize the mole begin with the difficulty the scientific community as a whole has had in communicating its meaning and utility to chemists.

Another plausible explanation for students’ challenges with the mole relates to how this topic is assessed. At-risk chemistry students were observed to struggle with attributing chemical identities to numerical values, particularly when units were shared between chemicals. For example, consider the stoichiometry assessment item (Figure 25) in which students were asked to convert 45.6 grams of molecular hydrogen to grams of ammonia. At-risk students were often observed to start the problem using the following process:

\[ 45.6 \text{ g } H_2 \left( \frac{1 \text{ mol } NH_3}{17.03 \text{ g}} \right) \]

Here, neither a terminological understanding of mass nor the algebraic arrangement of units for cancellation impacted a students’ ability to conduct this conversion accurately. What appears to be lacking is a conceptual understanding of the process in which the molar mass of one chemical cannot be used to convert a mass of another chemical to moles. While chemical interchangeability was more common amongst at-risk chemistry students, difficulties with conceptualizing algorithms were regularly observed throughout. Similar observations were described in a study conducted by Staver and Lumpe, wherein the extent of prior coursework in mathematics and chemistry was inconsequential concerning the lack of conceptualization amongst the students. Regardless of incoming preparation in mathematics or prior coursework
in chemistry, students did not attain the conceptual understanding intended following their assessment of this topic.

“This subject’s responses show a high level of awareness of the mole as a vehicle for moving back and forth between the macro and atomic/molecular levels, but he is unable to explain the numerical identity issue and does not correctly work either of the problems. Surmounting this barrier requires much more than an awareness; it requires that students overcome… insufficient understanding of the concepts and use of memorized algorithms, rules, or other information.” (p. 190)

Concerning representational competence, errors were observed in students’ interpretations of the chemical symbolism provided; however, these errors rarely influenced the successes of their algorithmic solution processes. For example, students solving the mole concept item often described the moles of oxygen as O, O₂, or O₄ while retaining the coefficient of 4 in their calculations (see Figures 24 and 26). The works of Gabel and Sherwood, Potgieter and Davidowitz, and Phillips suggest low-performing students and students who struggle with proportional reasoning could benefit via the provision of more visual instruction (e.g., graphic organizers and diagrams of the particulate nature of matter). The challenges in the context of this study rarely reflected issues with interpreting chemical symbolism (see nomenclature and elemental symbols in Table 14) and were more commonly concerning how to use these values in an algorithmic process. Thus, these visuals seem less likely to support students’ algorithmic success in solving these problems but may aptly address the more pertinent challenge in the students’ conceptual understanding of stoichiometry. Representational competence and conceptual understanding most commonly converged in students’ confusion concerning the distinct roles of coefficients and molar mass in calculating stoichiometric proportions (see Figures 27, 29, and 31). Here students apply coefficients to the molar masses of chemicals and can typically arrive at correct answers. While correct, these solution processes demonstrate the conflation of mass and moles.
The knowledge demonstrated enacts an algorithmic property where students execute a step-by-step process learned and practiced throughout the semester. Assessments, as observed in the research setting, can reinforce these algorithmic strategies by requesting students replicate these algorithms in the selection of a single numerical product. Thus, the algorithmic nature of assessments designed to measure the mole concept and stoichiometry may result in fostering chemically implausible algorithms that inequitably favor students of higher math aptitude and fail to measure conceptual understanding. A study concerning students’ approaches to solving stoichiometry problems concluded that while students with high proportional reasoning abilities used algorithmic reasoning strategies more frequently than their peers, students overall did not understand the chemical concepts on which the problems were based. This study found similar outcomes as the algorithmic processes of not-at-risk students more commonly resulted in the achievement of desirable assessment scores, some of the differential performance observed could be interpreted as students with higher math aptitude scores achieving greater comfort in applying and devising algorithms regardless of their chemical plausibility.

Overall the items involving mole-to-mole conversions, 70 (21%) of the 340 chemically implausible solution processes presented by not-at-risk chemistry students resulted in a correct answer for the items. Comparatively, 26 (13%) of the 199 at-risk chemistry students accomplished the same result. As students of both groups were discovered to share similar misconceptions related to these topics (see Table 14), it is possible how the mole concept and stoichiometry are assessed creates and reinforces the differential performance observed.

This implication is of consequence to instructors, researchers, and institutions interested in more equitable assessment design and closing performance gaps observed amongst students of variable incoming preparation. The combination of systematic approaches and conceptual understanding toward students’ conceptualization of the algorithms used in addition to the similarities observed amongst not-at-risk and at-risk chemistry students in a lack of
conceptual understanding of stoichiometry were informative. However, the inequitable successes following the implementation of chemically implausible algorithms observed as a potential source for the differential performance observed between these group is a concerning finding of this study. Ultimately, the work presented furthers the literature concerning at-risk chemistry students (or those with low math aptitude scores) and in the challenges of these students experience with the mole concept and stoichiometry.

Practical adaptations to the assessment items designed to measure proficiency with these topics could reduce the inequity observed by reducing algorithmic reliance and emphasizing the conceptualization of the processes by which students engage in solving stoichiometry problems. Such adaptations to assessment items could 1) improve the learning experiences of students by providing more detailed feedback related to their performance as a result of the distractors they’ve selected on a multiple-choice assessment items, 2) provide instructors with tractable data facilitating responsive changes to instruction following assessments, and 3) a means to potentially reduce the differential performance observed on these topics amongst cohorts of diverse student preparation. These objectives may be attained via the alignment of assessment prompts with the tasks they elicit. Engaging in this practice could improve the confidence in cognitive validity, the relationship between what an assessment aims to measure and what it elicits from the student, of the data collected from assessment items.\textsuperscript{38,39} Practical suggestions for adaptations to assessment items to promote conceptual understanding of the processes used to solve stoichiometry problems response to the findings of this study are discussed below and present opportunities for future exploration.

**Implications for More Equitable Assessment Design**

The processes used to solve the items were far more conceptual than expected and involved processes that intertwine students’ representational competence and conceptual aspects related to the unit and chemical identity of numerical values used. This skill set is highly
process-oriented suggesting that one way to align the task to the item and improve the cognitive validity of data acquired from multiple-choice questions on the mole concept or stoichiometry is to assess the process and not the product. For example, consider the mole concept item to which students’ solution processes were analyzed (Figure 38, left tile) in this study redesigned to items that emphasize process-over-product (Figure 38, center and right tiles) in addition to clarifying the interpretation of the stem by replacing the phrase “moles of oxygen” with “moles of oxygen atoms”.

![Figure 38. Potential adaptation of distractors for a product-oriented assessment item to that of a process-oriented assessment item.](image)

In the original assessment item (left), not-at-risk students were often observed to engage in solution processes that, although chemically implausible, could result in an accurate response (see responses authored by Alex and Ricky in Figure 34). The data collected from such a product-oriented assessment item may be biased in favor of not-at-risk chemistry students who have shown an ability to arrive at accurate responses via inaccurate conceptions of an algorithm. The adaptations of the mole concept item (left tile of Figure 38) to more process-
oriented representations (center and right tiles of Figure 38) could have a number of benefits. Process-oriented assessment items can elicit unique aspects of students’ solution process. The center tile focuses on misidentifying the mass percent of oxygen atoms in potassium sulfate as the moles of oxygen atoms in the compound (choice D) as observed in Figure 33 and interchangeable chemical identities (A and C) as observed in Figure 32. The right tile focuses on the atomic versus molecular representations of oxygen observed in Figure 26. The distractors of process-oriented assessment items can also present a variety of answer choices that achieve a shared numerical value. Note the right tile where answer choices B and E would both result in the correct numerical value, yet choice B would identify the number incorrectly as moles of O₄. The items in the center and right tiles promote student attention to solution processes.

From the instructor perspective, consider the data received regarding student proficiency with this concept across the three items. Instructors would be able to explicitly identify errors in the processes by which students engage including whether students confuse molecular with atomic oxygen, struggle with interchangeable chemical identities, or are unsure of the sequence by which chemically plausible solution processes occur via the distractors selected by their students. Pyburn et al., describe the impact distractors have on learning experiences of students taking multiple-choice tests and the benefit that follows the purposeful alignment of distractors with common alternative conceptions (or in the case of this study, chemically implausible solution process) when provided to students as feedback. Instructors could insert process-related challenges they have observed amongst their students or those observed in the presented study and gauge from the data the degree to which these challenges impact students’ conception of the topic. Further, by assessing process, the potential for reinforcing incorrect strategies that arrive in the correct numerical answer, as noted in students’ response processes, is avoided along with the inequitably this introduces to cohorts of students with variable preparation in mathematics.
Other Implications for Improved Assessment Design

One error prevalent amongst inaccurate responses of the students overall was confusion regarding the interpretation of “in excess” or “excess.” The stoichiometry, gas laws, and changes in energy assessment items uses the phrase “excess” to indicate that one of the two reactants involved in the chemical reaction is not limiting and thereby the limiting reactants algorithm is not necessary for determining the theoretical yield of the reaction described. Students often did not attribute “excess” with “the other reactant is limiting” and instead expressed solution pathways that at best were inefficient but resulted in accurate responses and at worst confused students to the point where they were unable to respond.

Should the intent be to measure students’ proficiency with determining theoretical yield and not connecting the phrase “excess” to the algorithm then eliminating the use of “excess” would be beneficial. Ultimately, researchers and instructors may consider the inclusion of “excess” as a confounding variable that poses a threat to the validity data collected on assessments of theoretical yield. Three possible options for adapting items to reduce this threat to the validity data collected are shown in Figure 39.

<table>
<thead>
<tr>
<th>Calculate the amount of NH₃ gas produced when 45.6 g of H₂ gas with reacts with a stoichiometric equivalent of N₂ gas according to the following chemical equation.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 H₂(g) + N₂(g) → 2 NH₃(g)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>If 126 L of H₂(g) at 15°C and a constant pressure of 0.920 atm reacts with N₂(g) as shown below, how many moles of NH₃(g) are produced? H₂(g) is the limiting reactant.</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 H₂(g) + N₂(g) → 2 NH₃(g)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Consider the following balanced chemical equation:</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 NH₃(g) → N₂(g) + 3 H₂(g)</td>
</tr>
<tr>
<td>ΔH_rxn = +91.8 kJ</td>
</tr>
<tr>
<td>How much heat is absorbed/released when 2.05 mol of NH₃(g) decomposes to produce N₂(g) and H₂(g)?</td>
</tr>
</tbody>
</table>

Figure 39. Suggestions for adapting assessment items in Figures 25, 28, and 30 toward items to avoid eliciting a limiting reactants algorithm from students as a result of reading the term “excess”.

142
The left and center tiles in Figure 39 provide options for maintaining the reaction without the use of the word excess. Additionally, researchers and educators may replace the chemical reactions related to calculations of theoretical yield with that of decomposition reactions (Figure 39, right tile) such that no competing reactants are present that could influence students’ solution processes.

One final consideration for the improvement of assessment design where the chemical identity and units associated with a numerical value are of importance is the inclusion of units with chemical identity in the distractors. The mole concept assessment item, for example, lists numerical values with corresponding answer choices without a description of chemical identity. It may change the manner in which students respond. For example, 23.7% of students selected distractor C (Figure 22) of “0.086 mol” but may have changed their response if the answer choice read “0.086 mol of O”. In particular, this is possible for students who reference that this value is indicative of the moles of potassium sulfate and not oxygen. Such a change would be reflective of their proficiency with the material and could serve as a learning opportunity for the student. Further, students may learn from the assessment and the data collected regarding their proficiency should common numerical values be paired with different units and chemicals identities. For example, should the distractors in Figure 22 be replaced with: “0.022 mol of O”, “0.086 mol of O₄”, “0.086 mol of O₂” and “0.34 mol of O₂” with the correct answer “0.34 mol of O” it would require a student to consider both the numerical responses and the chemical representation.

**Pedagogical Implications for the Instruction of Stoichiometry**

Adaptations to assessment design offer potential advantages to improving the accuracy of data generated from the assessment and ameliorating differential performance. Response processes observed in the dataset also provide guidance for pedagogy and preparation of instructional materials. In light of the interchanging of chemical identity that was observed,
instruction and associated materials may benefit by explicitly incorporating chemical identity along with units (e.g., grams / mol NH₃) during the presentation of worked problems. By explicitly noting chemical identity students may avoid the interchange of chemical identity and form a stronger conceptual link between the numerical values and their relevance toward the chemical phenomenon. While presenting worked problems, multi-step factor-label representations could be replaced with solution processes that pause after each conversion to describe the chemical meaning of the numerical value (see below).

Multi-step factor-level representation

\[
11.27 \text{ g } N_2 \times \frac{1 \text{ mol } N_2}{28.01 \text{ g}} \times \frac{2 \text{ mol } NH_3}{1 \text{ mol } N_2} \times \frac{17.03 \text{ g}}{1 \text{ mol } NH_3} = 13.70 \text{ g } NH_3
\]

Each step presented with units and chemical

\[
11.27 \text{ g } N_2 \times \frac{1 \text{ mol } N_2}{28.01 \text{ g}} = 0.4024 \text{ mol } N_2
\]

\[
0.4024 \text{ mol } N_2 \times \frac{2 \text{ mol } NH_3}{1 \text{ mol } N_2} = 0.8047 \text{ mol } NH_3
\]

\[
0.8047 \text{ mol } NH_3 \times \frac{17.03 \text{ g}}{1 \text{ mol } NH_3} = 13.70 \text{ g } NH_3
\]

A plausible explanation for the prevalence of the code “intrachemical unit identity” observed commonly throughout students’ solution processes is the conflation between moles as a coefficient in a chemical equation and as a starting value in the prompt of an item. Another way to further support students in delineating these quantities may be to present tabular representations of the values determined when presenting worked problems and encouraging students to do the same in their work (see Table 16 for an example).
Table 16. Example tabular representation of worked problem.

<table>
<thead>
<tr>
<th></th>
<th>N₂(g)</th>
<th>H₂(g)</th>
<th>NH₃(g)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mass</strong></td>
<td>11.27 g</td>
<td>25.19 g</td>
<td>0 g</td>
</tr>
<tr>
<td><strong>Change in Moles</strong></td>
<td></td>
<td></td>
<td>0 mol</td>
</tr>
<tr>
<td><strong>Moles</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If the stoichiometry assessment item was presented as a worked problem, instructors could model creating and completing the following table while working the problem. Subsequent student assignments could then be scaffold to first complete instructor-provided tables and later for students to generate and complete their own tables.

Didactic proposals for the use of tables in stoichiometry calculations can be found in the chemistry education, and engineering education research literature bases. Further research concerning students’ response processes to table stoichiometry may be a helpful start for evaluating whether or not this strategy could promote conceptual understanding of student’s solution processes when solving stoichiometry problems.

Finally, instruction may benefit by addressing students’ use of molar mass in the context of chemical reactions. Students’ processes described multiplying molar mass by the stoichiometric coefficient. Conventionally, introduction of molar mass calculations takes place without the context of chemical reactions. Later when stoichiometry is introduced, student proficiency with molar mass is presumed without explicit mention for how chemical coefficients fail to alter the molar mass. Instead stoichiometry offers an opportunity to reinforce the definition of molar mass as a conversion factor that is not influenced by the number of moles via its operationalization as the mass of one mole.

**Limitations and Future Works**

The data collected in this study reflect the responses of chemistry students within a single semester at a research setting, which consists of a single institution. While the intent of
the study is not to provide generalizable characterizations of the challenges chemistry students face when solving mole concept and stoichiometry questions, the authors sought to generate hypotheses regarding the nature of at-risk student struggles and the causes of differential performance observed with these topics. The information elicited from students regarding mole concept and stoichiometry assessment items were (as argued in the study) highly dependent on the items selected for review. All of the questions selected may be considered “traditional” items and are product-oriented, multiple-choice items that rely on chemical formulas and balanced chemical reactions to convey information as to the proportions by which chemicals react. As representational competence played a role in the lack of conceptual understanding conveyed, future explorations of students’ representational competence and other forms of chemical representation (e.g., submicroscopic diagrams) may be analyzed to discern whether this skill set contributes to differential performance or the overall difficulty of an assessment item.

The consistency among assessment items in both representation and the use of the particular chemical reaction depicting the synthesis of ammonia was intentional to reduce conflating variables when reviewing students’ performance on a subset of tasks. This decision could also limit the information observed regarding challenges students’ experience with these items and sources of differential performance observed. Items were designed naturalistically by the course instructors and were chosen a priori to students’ performance and differential performance. Thus, items intentionally selected as a result of high mean differential could provide a different viewpoint of the challenges that separate these two groups of students. Given the information collected, researchers could seek to design assessment items with the intent of identifying a particular challenge with a conceptual model in mind for how at-risk chemistry students struggle and explore differential performance based on assessment design. Other limitations of the study include the role of motivation and implicit assumptions made regarding students’ responses. Students exhibited a great range of responses to the items, and one could argue that some of the errors observed were the result of students’ apathy toward
completing the survey with greater detail. While the consistency of the errors observed and the nature of detail given for the responses suggest students’ misconceptions were not necessarily derived from motivation, these concerns informed decisions made by the authors in the treatment of the data. Only students who responded to all three surveys following interim exams were included in the analysis presented above. Additionally, student responses too vague to evaluate students’ efforts such as “I don’t know” or “just use stoichiometry” were not considered in the description of the challenges students faced with their solution processes.

A final limitation concerning the data elicited by the survey prompts was the frequency by which researchers had to make implicit assumptions about the unit, chemical, or numerical identity of a value provided by the students. For example, a student could start their solution process to the mole concept item as “take 15 grams and divide by the molar mass” which was interpreted by the researchers to mean “divide 15 grams of K₂SO₄ by the molar mass of K₂SO₄”. While the data collected more efficiently provided the solution processes for hundreds of students in the course, in-depth interviews carried with a subset of students are a target for future exploration as therein exists the capability for researchers to ask follow-up questions regarding the identities of the values they present in their solution processes. Suggestions for adaptations of assessment items including an emphasis in process observed in Figure 39, adaptations to prompts presented in Figure 40, and table stoichiometry as exemplified by table 15 have not been explicitly tested by the researchers. Response processes for these changes in addition to reports on student outcomes are intended in future works but do not encompass the scope of the presented works intended to identify how students struggle with the stoichiometry assessment items on which at-risk chemistry students were identified to perform differently. Ultimately, the authors hope to issue a call for future research as to the impact of assessment design on differential performance. Whether by in-depth interviews or a compiled database of items presenting with variable differentials, greater nuance as to moderators of difficulty and differential performance could support assessment design and inform interventions or
prerequisite practices issued for students of variable preparation entering this critical point in the STEM coursework.

**Conclusion**

Despite comparable proficiencies with the representational competency, systematic approaches, and conceptual understanding required for the mole concept and stoichiometry assessment items (see Table 14), students’ success at implementing chemically implausible algorithms appears the most apparent source of differential performance between chemistry students of lower and higher math aptitude scores. Not-at-risk chemistry students were observed to make similar errors to those of their at-risk peers often regarding the conceptual understanding of the algorithms used to solve mole concept and stoichiometry assessment items. However, a difference was observed in the commonality by which not-at-risk students used chemically implausible solution paths to arrive at the correct numerical value elicited by the prompt of these assessment items when compared to their at-risk peers. As the differential performance observed does not seem to reflect differences amongst the students’ knowledge of chemistry and instead favors students entering with more preparation in the rote execution of algorithms, current practices involving the assessment of stoichiometry is inequitable. The primary challenges observed in students’ solution processes with the mole concept and stoichiometry assessment items were: 1) interchanging chemical identities of the numerical values used in stoichiometry, 2) distinguishing representational differences between atomic and molecular representations of elements capable of forming diatomic molecules, 3) applying mass percentages to moles or molar mass, 4) applying stoichiometric coefficients in the determination of molar mass, 5) inserting molar mass in algorithms requiring only mole-to-mole ratios, and 6) applying solution processes related to identifying limiting reactants where all other reactants are in excess. To reduce the inequity concerning how this topic is assessed, instructors and researchers should actively seek out processes that invoke a conceptual understanding of
stoichiometry. Potential changes to instruction and assessment of stoichiometry including the use of table stoichiometry as a scaffold for students, replacement of product-oriented assessment items (e.g., calculate x) with process-oriented assessment items (e.g., which process describes how to calculate x) and other proposed multiple-choice formats are presented with an intent to support future research aimed at addressing the observed performance gaps in this gateway STEM course.

References


(31) IBM SPSS Statistics for Macintosh, Version 25.0; IBM Corp.: Armonk, NY, 2017.

(32) MAXQDA 2018; VERBI Software: Berlin, Germany, 2018.


(42) Serafin, M. Engineering’s Approach to Teach Stoichiometry. 2006, 7.
CHAPTER 6:
IMPACT OF REPRESENTATIONS IN ASSESSMENTS ON STUDENT
PERFORMANCE AND EQUITY

Note to Reader

Portions of this chapters have been previously published in *Journal of Chemical Education* (DOI: 10.1021/acs.jchemed.9b01058). These works have been reproduced for this dissertation with permission from the American Chemical Society (see Appendix B for Copyright Permissions).

Introduction

Representations comprise the language by which scientists communicate,¹ and can be used to model, predict, and explain unique aspects of a physical phenomenon.² ³ Emphasized throughout the Next Generation Science Standards,⁴ and prevalent in assessments of biology, chemistry, earth science, and physics,⁵ competence in the use of representations is a prominent aspect of engaging with the sciences. While the role of representations in teaching and learning science is evident, practical implications for the use of representations in assessments is less prominent in the literature. Consider a sample of the representations that may be used in an assessment of students’ proficiencies with stoichiometry (see Figure 40 on the next page).

The prompt in Figure 40 could be paired with a single representation (or combinations of multiple representations). However, assessment designers will find limited evidence as to how their choice of representation will influence student outcomes, particularly for students most at-risk for achieving unfavorable outcomes.
Prompt
How many mol of HCl(aq) are required to completely react with 3.7 mol of Cu(s)?

Representation

Symbolic

Cu(s) + 2 HCl(aq) → CuCl₂(aq) + H₂(g)

Particulate

Table

<table>
<thead>
<tr>
<th>Chemical</th>
<th>Ratio</th>
<th>Moles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu(s)</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>HCl(aq)</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Figure 40. Examples of representations for use in the assessment of stoichiometry.

Differentiating Representation (Symbolic vs. Particulate) from Task (Conceptual vs. Algorithmic)

Explorations of students’ differential performance on conceptual vs. algorithmic assessment tasks,⁶⁻¹³ provided the foundation for this investigation on the impact of representations on assessments. Throughout these works, a paired problems methodology was enacted wherein the same group of students completed pairs of assessment items related to a common chemical phenomenon (e.g., Boyle’s law, limiting reactants) but differing task as algorithmic or conceptual (characterized as requiring and not requiring mathematics, respectively). In these works, the compared items differed in both task, what the students is asked to do, and representations used as algorithmic tasks most commonly used symbolic representations (e.g., chemical reactions and formulas) and conceptual tasks often used particulate representations (e.g., pictorial representations of chemical species at the particulate level). For example, Nurrenbern and Pickering compared student performance on gas law items that required calculating the volume or pressure of gas (algorithmic) with an item that required identifying the distribution of gas molecules in a container (conceptual).⁶

The evidence generated was inconsistent. Some studies observed students’ algorithmic proficiency exceeding that of the conceptual,⁶,⁷,⁹ one study reports inconsistent trends for pairs
spanning various chemical phenomenon,\textsuperscript{8} other studies reported non-significant differences between students’ algorithmic and conceptual proficiencies,\textsuperscript{11,13} and a final study reported greater conceptual proficiency than algorithmic.\textsuperscript{12} These inconsistencies may reflect challenges interpreting differential student performance from paired problems differing in both task (conceptual vs. algorithmic) and representation (symbolic vs. particulate). For example, Coștu designed a pair of gas law items communicated using similar representations (depictions of pistons and chemical formulas) to assess students’ proficiency with Boyle’s law approached conceptually (will pressure increase or decrease if the volume of the container is increased) and algorithmically (calculate the pressure after the volume of the container is decreased).\textsuperscript{12} The results suggest students were more successful with the conceptual task than the algorithmic task for paired problems in which representations was not a confounding variable.

The methodology of paired problems introduces other threats to the validity of the data generated as differences in student performance could be explained by differences in answer choice viability. For example, Sawrey reports frequencies for the distractors selected on the conceptual item suggesting that two answer choices (the correct answer and one of the distractors) were selected by 98.5\% of the students.\textsuperscript{9} While the distractor selection frequencies of the algorithmic assessment item were not reported, if the task, representations used, and ease of eliminating distractors differs between the paired problems, conclusions as to what is driving these students’ differential success can be difficult to deduce.

Just as the problems used by Coștu differ only in task and not representation, one could devise a pair of problems differing only in representation and not task. Further, the answer choices of the paired problems could share a common rationale to facilitate cross-examination of the impact representations may have on the distractors to which students attend. This substitution of representations (without a change in task) can provide novel implications for assessment designers including instructors and education researchers.
Investigating the Role of Representations in Chemistry Assessments

Previous works sought to diagnose students’ proficiencies with interpreting representations,14–16 but do not inform how the use of one representation over another (substitution) or inclusion of more than one representation (supplementing) may moderate student performance on similar assessment tasks. For example, Irby et al.,16 describe a card sorting task diagnosing students’ ability to sort practice problems by concept (dilution, stoichiometry, and mass present) when presented with macroscopic, particulate, and symbolic representations. Participants were observed to sort the cards on a progression from sorting by representation to sorting by concept and were not tasked with solving the practice problems. Nyachwaya et al.,14 describe an open-ended drawing tool to contrast differences in students’ ability to balance symbolic representations of chemical reactions and illustrate particulate representations (pictorial diagrams of chemical species at the particulate level as shown in Figure 40). Posited as a tool for formative assessment, the article echoes critiques as to the reliance of symbolic representation in summative assessment,14,17,18 and identifies instruments designed to incorporate more variable representation.13,15,18–21 While both studies advance works seeking to diagnose and assess students’ ability in translating representations, neither presents evidence as to the use of the activities in summative assessments nor compares student performance following the application of differing representations to similar assessment tasks.

In the research literature, assessments have been developed that incorporate varying representations. In the chemistry concept inventory advanced by Mulford and Robinson, students’ ability to translate symbolic and particulate representations is assessed using a multiple-choice format with distractors designed to reflect the misconception literature.19 The chemistry concept inventory applies a variety of representations. However, the content of these assessment tasks changes across items and representation, limiting explorations of the impact representations have on student performance. A strength of this inventory is the use of common
misconceptions as distractors in multiple-choice instruments. These distractors were used to examine the prevalence of these misconceptions in chemistry students and the degree to which instruction lessened their prevalence. This approach presents a means by which to examine the role of representations in eliciting unique aspects of students’ solution strategies not previously investigated.

Another example of studies evaluating the role of representations in summative assessments can be found in the work of Bunce and Gabel who developed the Symbolic, Application, Particulate (SAP) instrument wherein 3 items spanning 10 topics (30 items total) were represented using symbolic, particulate, or macroscopic representations and applied the instrument to an investigation of gender differences following a treatment of three, 2-week modules presented with or without particulate representations. The evidence presented involves cross-sections of students’ mean performance on the SAP before and after the modules. Similar to the chemistry concept inventory, the content of the assessment tasks changes for each of the three representations shown, which serves as a confounding variable that prevents investigating the impact of representations. Chandrasegaran et al., developed a two-tier, multiple-choice instrument assessing students’ use of macroscopic and particulate-level reasoning to justify answers to a symbolically represented prompts. The study presents evidence of students’ difficulty with translating between macroscopic and particulate representations within the same assessment task but was not designed to explore the impact of representations between assessment tasks. These three studies, detail the development of summative assessments that incorporate various representations emphasizing the need to make explicit the relationships among varying representations. However, these works do not address the central question of this study concerning the impact of substituting one representation for another or supplementing with various representations on students’ assessment performance or selection of answer choices.
Advancing Research Concerning Equitable Assessment Design Practices

In investigating the impact changes to the representations used in chemistry assessments can have on overall student performance, we acknowledge calls to promote research examining how reforms impact diverse groups of students.\textsuperscript{22–27} While there are many forms of equity research, the focus of this study was to promote equity of individuality, a model of equity examining performance within a marginalized group. For the purposes of this paper, chemistry students predicted as at-risk as a result of differential preparations in mathematics will comprise the group of interest in an investigation of equity.\textsuperscript{23,25}

In undergraduate general chemistry courses, a strong relationship has been observed between students' precollege math test scores (e.g., SAT and ACT) and outcomes in general chemistry courses.\textsuperscript{26,28–37} Accounting for 20-40\% of the variance observed amongst chemistry students,\textsuperscript{26,30,32,37–40} performance in chemistry courses is related to precollege math test scores with students scoring in the bottom quartile of math tests scores at-greatest-risk for withdrawing or achieving D's and F's when compared to their peers with higher math test scores.\textsuperscript{26,29,31,33,36,37,41}

Decades of research predicting \textit{which} students are at-risk for unfavorable academic outcomes in chemistry have yielded limited understandings as to \textit{how} at-risk chemistry students struggle with the content of the course. Ralph and Lewis identified assessment of the mole concept and stoichiometry as most differentially challenging to at-risk chemistry students,\textsuperscript{37} and analyzed students' response processes to these assessments to identify why these differentials may persist.\textsuperscript{41} The results suggest the differential between students not-at-risk and at-risk were most commonly the result of not-at-risk students' executing chemically implausible algorithms to arrive at accurate numerical values (see Figure 41 for examples of these solution processes).
Consider the following balanced chemical equation:

\[ 4 \text{NH}_3(g) + 5 \text{O}_2(g) \rightarrow 4 \text{NO}(g) + 6 \text{H}_2\text{O}(l) \quad \Delta H_{rxn} = +1,168 \text{kJ} \]

How much heat is absorbed/released when 2.05 mol of NH\(_3\)(g) reacts with excess O\(_2\)(g) to produce NO(g) and H\(_2\)O(l)?

\[
2.05 \text{ mol NH}_3 \times \frac{17.03 \text{ g NH}_3}{1 \text{ mol NH}_3} \times \frac{1 \text{ mol NH}_3}{17.03 \text{ g NH}_3} \times \frac{+1,668 \text{kJ}}{4 \text{ mol NH}_3} = 5.99 \times 10^2 \text{kJ}
\]

Figure 41. Examples of chemically implausible solution processes presented by not-at-risk students in response to a task involving stoichiometry.

These findings were concerning as the performance gap was attributable to differential success in students’ rote execution of algorithms to arrive at a numerical answer despite deficient understandings of the chemistry context in which the problems were situated. For example, the first response process in Figure 41 depict a solution process that involves converting from moles of NH\(_3\) to grams of NH\(_3\), then back to moles of NH\(_3\) before executing the calculation. The second response reflects students who multiplied the molar mass of NH\(_3\) to both the mole quantity and stoichiometric coefficient of NH\(_3\) provided in the prompt. These erroneous processes resulted in the correct answer suggesting assessments of this design may generate an invalid measure of chemistry understanding and inequitably favor students with higher standardized math scores. The response processes potentially reflect a lack of conceptualization as to the meaning of units, numerical values, and the relation to the chemical substances represented therein. These works,\(^{37,41}\) acknowledge a flaw in the measures used to assess chemistry knowledge and reflect the critiques of other researchers on the summative,
multiple-choice instruments reliant on symbolic representations that are commonly used in chemistry assessments.\textsuperscript{14,17,18}

**Purpose**

Assessment designs comprised predominately of items communicated symbolically to express algorithmic tasks resulting in the calculation of numerical values may generate invalid measures of chemistry understanding and perpetuate inequity for at-risk chemistry students. The use of particulate representations and tables to support students’ organization of the units and chemical identities involved in their solution processes has been posited as a potential reform.\textsuperscript{41} However, research as to the impact of representations on students’ performance on assessments sharing a common task was not identified. This study seeks to provide an exploratory evidence base as to these impacts of representations on assessments to provide practical implications for broader assessment design practices that promote the use of representations beyond that of the symbolic level, provide multi-modal learning experiences for students, and promote equity in the performance of students with diverse preparations. With these objectives in mind, the work was guided by the following research questions:

1. How does the *substitution* of symbolic and particulate representations and *supplementation* of symbolic with tables or particulate representation in assessments impact overall and at-risk student performance?

2. What can be inferred about the impact of representations on students’ distractor selections?

**METHODS**

**Research Setting**

Data were collected at a large, public research institution in the southeastern United States. In this setting, multiple sections of first-semester chemistry classes were held each term
and were attended by 200 to 240 students. The syllabus describes the content of the course as related to the principles and applications of chemistry, including properties of substances and reactions, thermochemistry, atomic-molecular structure and bonding, and periodic properties of elements and compounds. Classes were coordinated using a shared textbook, learning objectives, syllabus, grading scheme, and assessments. The textbook used for the classes in these semesters was “Chemistry 2e” by OpenStax at Rice University. Students attended lecture twice per week and problem-solving sessions facilitated by peer leaders once per week.

Data Collection

The data presented herein are of student responses to multiple-choice assessments administered as exams typically collected in the instructional setting. Four exams (three interim and a final, cumulative exam) were administered throughout the semester, comprising 70% of each student’s final grade. Students’ assessment responses were collected in a controlled setting wherein exams were proctored by teaching assistants and instructors. Students received a cover sheet with a periodic table, were permitted to bring a nonprogrammable, scientific calculator, and were offered scratch paper. Interim exams included twenty, multiple-choice items (with five answer choices) worth 7 points each. The final, cumulative exam included forty-five, multiple-choice items of the same format worth 5 points each. Assessments were designed by a committee of the instructors teaching the course and were common across all classes within each semester.

The remaining 30% of a student’s grade was comprised of three, effort-based measures worth 10% each. These measures involved a student’s performance on in-class clicker questions used in the lecture setting, attending and participating in peer leading sessions, and completing online homework assignments. All data collection adhered to guidelines set by the
institutional review board at the research setting in the evaluation of the effectiveness of curricular approaches on student success in chemistry.

Substituting Representations

To explore the impact of representations, assessment items with the same task (e.g., predicting ionic dissociation) but different representations, symbolic or particulate, were given to students within the same exam. For the first interim exam, students were tasked to predict the quantity and charge of ions that arise from the dissolution of an ionic compound (see Box 3).

**Box 3. Substituted representations for the dissolution of ionic compounds.**

*Symbolic Representation*
In water, 1 mol of Li₂CO₃(aq) will dissociate into which ions?
A) 2 mol of Li⁺ and 2 mol of CO₃²⁻
B) 1 mol of Li²⁺ and 1 mol of CO₃²⁻
C) 1 mol of Li²⁺ and 2 mol of CO₃²⁻
D) 2 mol of Li⁺ and 1 mol of CO₃²⁻
E) Li₂CO₃ will not dissociate in water.

*Particulate Representation*
Which diagram represents the dissolution of K₂SO₃(aq)?

Since students would see both items within the same exam, the chemical context was changed (e.g., Li₂CO₃ versus K₂SO₃) to reduce the likelihood students would recognize the similarity between the items. The correct answers to these items were “D” for the symbolically
represented item and “A” for the particulate item. Both items shared a common set of distractor rationales which included:

- **Insoluble** - inaccurately assigning a soluble compound as insoluble (E, both items).
- **Incorrect quantity** - selecting an incorrect quantity for the anion (A, symbolic; D, particulate).
- **Cation charge** - conflating the charge for the quantity of the cation (B, symbolic; C, particulate).
- **Charge and quantities** - conflating the charge and quantities of both ions (C, symbolic; B, particulate).

The second pair of items related to stoichiometry were administered to students during the second interim examination of the semester for the calculation of theoretical yield (see box 4).

**Box 4. Substituted representations in the calculation of a theoretical yield.**

*Symbolic Representation*
Consider the balanced chemical equation: Cu(s) + 2 HCl(aq) → CuCl2(aq) + H2(g)

How many grams of Cu(s) is required to completely react with 57.2 g of HCl(aq)?

A) 49.9 g of Cu  B) 51.3 g of Cu  C) 99.7 g of Cu  D) 105 g of Cu  E) 199 g of Cu

*Particulate Representation*
The following picture represents a balanced chemical reaction between N2(g) and H2(g) to form NH3(g).

How many grams of N2(g) is required to completely consume 15.9 g of H2(g)?

A) 36.8 g of N2  B) 73.7 g of N2  C) 89.6 g of N2  D) 221 g of N2  E) 663 g of N2
The correct answers to these items were “A” for the symbolic item and “B” for the particulate item. Both items shared a common set of rationales in the design of the distractors, which included:

- *Skipped mole ratio* - failing to apply a mole ratio between reactants (C, symbolic; D, particulate).
- *Incorrect molar mass* – applying the molar mass of a constituent element; Cl for HCl and N for N₂ (B, symbolic; A, particulate).
- *Product mass* - calculating the mass of a product instead of the intended reactant described by the prompt (D, symbolic; C, particulate).
- *Flipped mole ratio* - applying the inverse of the mole ratio (E, both).

A similar investigation of items in the third interim exam was not conducted as topics assessed at this stage of the learning sequence had few explicit links to the mole concept and stoichiometry.

The enactment of substitution representations featured some threats to internal validity in the form of potential alternative explanations. First, the items differed in chemical context, which may have an unforeseen impact on student success. Second, since the items appears in the same exam, there are possible priming effects where the item order impacts student success. To address these concerns, a different methodology was used in supplementing representations.

**Supplementing Representations**

To evaluate the impact of *supplementing representations*, students were randomly assigned to one of three versions of a final exam item sharing an identical prompt, chemical context, distractors, and symbolic representation (see Figure 42).
Figure 42. Three versions of a stoichiometry item involving percent yield (correct answer in bold).

The particulate version was supplemented with a particulate representation, and the table version was supplemented with a table. The items were randomly distributed amongst the students using a form series wherein forms A, B, and C each presented a different version of the item, thereby removing the potential priming effect.

Out of concern for the one-third of students who would not receive supplementary representations with the implementation of a form series, a triad of stoichiometry items were designed. These items ensured each student received one item presenting the balanced chemical reaction alone (symbolic only), one with a supplemental particulate representation (+particulate), and one with a table (+table) as shown in Table 17 (see next page).
Table 17. Each student was randomly assigned to a version of a stoichiometry item with each representation.

<table>
<thead>
<tr>
<th>Test Form</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molarity</td>
<td>+ Particulate diagram</td>
<td>+ Table</td>
<td>Symbolic only</td>
</tr>
<tr>
<td>Percent Yield</td>
<td>Symbolic only</td>
<td>+ Particulate diagram</td>
<td>+ Table</td>
</tr>
<tr>
<td>Gas Laws</td>
<td>+ Table</td>
<td>Symbolic only</td>
<td>+ Particulate diagram</td>
</tr>
</tbody>
</table>

All three items involved a mole-to-mole conversion in the context of percent yield, reactions in solution, or the gas laws (see Figures 42, 43, and 44).

Figure 43. Three versions of a stoichiometry item involving the molarity of a solution (correct answer in bold).
Figure 44. Three versions of a stoichiometry item in the context of gaseous solutions (correct answer in bold).

All three items referenced the same chemical reaction between magnesium and hydrochloric acid and were presented to students in the order of percent yield, molarity, and then gas law, regardless of test form to reduce potential priming effects. Answer choices (and by extension, rationales) were kept constant throughout the versions of each item. Across the triad of stoichiometry items, two distractor rationales remained constant. Numerical values for these distractors reflect failing to apply \textit{(skipped mole ratio)} or inverting \textit{(flipped mole ratio)} the mole ratio during the solution process.

Instructors had full discretion to edit the assessment items as they saw fit, resulting in the items presented in this work. For example, the percent yield item with a table in 42 originally
had a column for “moles”. This column was removed as the instructors were concerned with providing too much of a scaffold with which students could reason the answer. Instructors also had the option to award full credit to students for all three items of the form series if the differences in performances among representations were considerable.

Identifying At-Risk Students

For this study, changes in student performance overall were not sufficient in an exploration of the impact representations have on students’ assessment performance on these topics given the strength of the relationship between students’ precollege math aptitude scores and academic performance in chemistry, and the differential performance commonly observed on items of the mole concept and stoichiometry topics. As such, the impact of these changes in representation to at-risk students, previously identified as having scored in the bottom quartile of math aptitude scores for the student cohort, were considered in this study. There were 2,035 students enrolled in the course during data collection. Of these students, 1,881 (or 92%) registered a precollege math aptitude score with the university. The distribution of these scores, concorded to the new SAT, have a median score of 610 and a bottom quartile of 580, the latter score serving as a cutoff below which identified students as at-risk (see Figure 45 on the next page).

Applying this cutoff, 432 students were identified as at-risk. Differences in the academic performance of students predicted as at-risk and not-at-risk throughout the semester support the strength of the relationship between precollege math aptitude scores and students’ performance in chemistry courses at this setting (see Table 18 on the next page).
Figure 45. A histogram of the distribution of math test scores for students enrolled in first semester, general chemistry course at the research setting during data collection.

Table 18. Of students who completed all four exams and had a math aptitude score (n = 1,469), the differential performance observed between those at-risk (scoring in the bottom quartile of math aptitude scores), and not-at-risk is evident.

<table>
<thead>
<tr>
<th>Group</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Final Exam</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>ABC (%)</td>
<td>DF (%)</td>
<td>ABC (%)</td>
</tr>
<tr>
<td>Not-at-risk</td>
<td>1176</td>
<td>77</td>
<td>23</td>
<td>74</td>
</tr>
<tr>
<td>At-risk</td>
<td>293</td>
<td>39</td>
<td>61</td>
<td>43</td>
</tr>
</tbody>
</table>

The majority of at-risk students were observed to achieve a “D” or “F” (or below 60% of the points possible) across the exams administered throughout the semester, whereas the majority of not-at-risk students achieved an “A”, “B”, or “C”.

Results for the remainder of this report relate to the total sample size (N = 2,035 students) toward evaluating overall performance and the subset of students (n = 432) at-risk to
investigate how these changes to the assessment impact chemistry students predicted to perform less favorably. As the semester progresses, the total number of students completing each assessment decreased (i.e., 1,948 students took the first interim exam, 1,851 took the second, and 1,800 took the final exam).

**Measures and Evaluation**

Chi-square tests of independence were used to evaluate the impact of substituting and supplementing representations in assessments. These tests were used to compare differences in student performance (correct or incorrect, \( df = 1 \)) and distractor selections (selecting each of the answer choices, \( df = 16 \)) between the reference (symbolic representation) and experimental groups (supplemented with either a table or particulate representation). The null hypothesis for these tests can be described as:

- Evaluating success (correct = 1, incorrect = 0) \( H_0 \): the representations shown in these assessment items have no detectable relationship with students’ success in answering the item correctly.
- Evaluating distributions of distractor selection frequencies (A = 1, B = 2, C = 3, D = 4, and E = 5) \( H_0 \): the representations shown in these assessment items have no detectable relationship with students’ distractor selections.

To control for group-wise error rates, lower critical values (\( \alpha = \frac{0.05}{2} = 0.025 \)) were used in the determination of statistical significance determined via a Bonferonni correction (the quotient of the critical value and the number of tests conducted).\(^46\) For estimations of substantive significance, Cohen’s \( w \), was calculated using Equation 4 below.\(^47\)

Equation 4. Cohen’s \( w \) formula where \( N \) is equal to the total sample size or the number of students who answered the items for which proportions are being compared and \( \chi^2 \) is equal to the chi-square statistic.

\[
w = \sqrt{\frac{1}{N} \chi^2}
\]
Effects were interpreted as small when $w$ is close to 0.1, medium when $w$ is close to 0.3, and large when $w$ is close to 0.5.\textsuperscript{48}

A power analysis was conducted post hoc to evaluate the power of the chi-square tests evaluating these differences. Power estimates were acceptable (1-$\beta > .80$) for tests conducted with students overall and at-risk.\textsuperscript{48}

**Results**

**Substituting Symbolic with Particulate Representations had Negative Impacts on Student Performance and Equity**

Substituting symbolic representations with particulate diagrams on assessments had a statistically significant, negative impact on student performance for both assessment tasks with effect sizes ranging from small ($w = 0.19$) to medium ($w = 0.37$) and reductions in overall student performance from 3.8-4.1\% (see Table 19).

**Table 19.** Chi-square tests comparing the number of students who selected the correct answer for assessment tasks substituting representation\textsuperscript{a}

<table>
<thead>
<tr>
<th>Item Representation</th>
<th>Dissolution (Box 3)</th>
<th>Item Representation</th>
<th>Stoichiometry (Box 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Symbolic</td>
<td>Particulate</td>
<td>Symbolic</td>
</tr>
<tr>
<td>$n$</td>
<td>1948</td>
<td>1948</td>
<td>1851</td>
</tr>
<tr>
<td>$P$ (%)</td>
<td>82.3</td>
<td>78.5</td>
<td>84.1</td>
</tr>
<tr>
<td>$X^2$ (w)</td>
<td>73.38 (0.19)</td>
<td>408</td>
<td>$X^2$ (w)</td>
</tr>
<tr>
<td>$n_{AR}$</td>
<td>408</td>
<td>408</td>
<td>$n_{AR}$</td>
</tr>
<tr>
<td>$P_{AR}$ (%)</td>
<td>73.3</td>
<td>68.1</td>
<td>70.3%</td>
</tr>
<tr>
<td>$X^2$ (w)</td>
<td>19.25 (0.22)</td>
<td>45.14 (0.34)</td>
<td></td>
</tr>
</tbody>
</table>

\textsuperscript{a}Where $P$ is the percent of students who answered the item correctly with $P_{AR}$ as the percent of at-risk students who answered the item correctly.

\textsuperscript{b} $p < 0.025$.  

---

170
The substantive significance of these differences in representation was more pronounced for the stoichiometry assessment tasks \((w = 0.37)\) than the dissolution tasks \((w = 0.19)\). Similarly, the performance of at-risk chemistry students declined (by 5.2%) with the substitution of particulate representations with differences measured at medium effect sizes \((0.22 \leq w \leq 0.34)\).

**Supplementing Items with Tables had a Positive Impact on Student Performance and Equity, whereas Supplementing with Particulate Representations Presented Variable Outcomes**

No significant impacts on overall student performance were detected for supplementing with particulate diagrams \((0.01 \leq w \leq 0.05; \text{ see Table } 20)\). A significant impact was detected for tables including a column for “moles” as observed for the molarity (Figure 43) and gas laws items (Figure 44).

Table 20. Chi-square tests comparing the number of students who selected the correct answer for assessments supplementing symbolic representation (S) with either a particulate diagram (P) or table (T)

<table>
<thead>
<tr>
<th>Item</th>
<th>Percent Yield (Fig. 42)</th>
<th>Molarity (Fig. 43)</th>
<th>Gas Laws (Fig. 44)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rep</td>
<td>S</td>
<td>P</td>
<td>T</td>
</tr>
<tr>
<td></td>
<td>Rep</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>605</td>
<td>595</td>
</tr>
<tr>
<td></td>
<td>P (%)</td>
<td>68.8</td>
<td>67.9</td>
</tr>
<tr>
<td></td>
<td>(X^2 (w))</td>
<td>\text{Ref.}</td>
<td>0.103 (0.01)</td>
</tr>
<tr>
<td></td>
<td>nAR</td>
<td>139</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>PAR (%)</td>
<td>57.6</td>
<td>53.3</td>
</tr>
<tr>
<td></td>
<td>(X^2 (w))</td>
<td>\text{Ref.}</td>
<td>0.465 (0.04)</td>
</tr>
</tbody>
</table>

Where \(P\) is the percent of students who answered the item correctly with \(P_{AR}\) as the percent of at-risk students who answered the item correctly.

\(p < 0.025\).

Small effect sizes ranged from 0.08 to 0.15 but were undetectable \((w = 0.02)\) for tables without a column for “moles,” as observed with the percent yield problem shown in Figure 42. Similar
results were observed with at-risk students wherein particulate diagrams had no detectable effect on their performance \((0.03 \leq w \leq 0.04)\). Tables with a column for “moles” improved at-risk student performance at +7.2% for the molarity item and +22.2% for the gas laws item resulting in small to medium effects \((0.07 \leq w \leq 0.23)\).

**Representations in Assessments had a Considerable Impact on the Distractors Selected**

**Substituted Representations**

For the items of substituted representation (see Table 21), the distribution of distractors students selected significantly differed with medium to large effect sizes detected \((0.28 \leq w \leq 0.50)\).

Table 21. Chi-square tests comparing the number of students who selected each of 5 distractors for items of substituted representation.

<table>
<thead>
<tr>
<th>Item Representation</th>
<th>Dissolution (Box 3)</th>
<th>Item Representation</th>
<th>Stoichiometry (Box 4)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Symbolic</td>
<td>Particulate</td>
<td></td>
</tr>
<tr>
<td>( n ) Correct answer (%)</td>
<td>1948</td>
<td>1948</td>
<td>( n ) Correct answer (%)</td>
</tr>
<tr>
<td></td>
<td>82.3</td>
<td>78.5</td>
<td></td>
</tr>
<tr>
<td>Incorrect quantity (%)</td>
<td>2.6</td>
<td>2.8</td>
<td>Incorrect molar mass (%)</td>
</tr>
<tr>
<td></td>
<td>8.4</td>
<td>6.0</td>
<td>Skipped mole ratio (%)</td>
</tr>
<tr>
<td>Cation charge (%)</td>
<td>3.5</td>
<td>1.5</td>
<td>Product mass (%)</td>
</tr>
<tr>
<td>Charge &amp; quantities (%)</td>
<td>3.1</td>
<td>11.1</td>
<td>Flipped mole ratio (%)</td>
</tr>
<tr>
<td>Insoluble (%)</td>
<td>149.9 (0.28)a</td>
<td></td>
<td>454.9 (0.50)a</td>
</tr>
<tr>
<td>( X^2 (w) )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\(^a p < 0.025.\)

A strength of the paired problem approach wherein answer choices share a common set of rationale is the ability to compare the distractors students selected when answering both items. In responding to the substituted dissolution problems, the majority (1319 or 67.7%) of students answered both items correctly. Of the 133 students who answered both dissolution problems
incorrectly, 45 students selected distractors of the same rationale (i.e., students selected the “insoluble” distractor, or any two of the same distractor rationales, for both items). For the substituted stoichiometry problems, 1345 students (72.7%) answered both items correctly and of the 159 who answered both incorrectly, 73 selected distractors of the same rationale (i.e., students selected the “skipped mole ratio” distractor, or any two of the same distractor rationales, for both items). As such, more students selected distractors of a common rationale for the paired stoichiometry problems (73 of 159 or 45.9%) than for the paired dissolution problems (33.8%).

One difference likely to be reflective of the representation shown was in the percent of students selecting the “insoluble” distractor for the paired dissolution problems. However, it is unclear the extent to which differences in students’ selections of the distractors within a pair of assessment tasks can be attributed to 1) the representations shown, 2) differences between the assessment tasks within each pair, or 3) priming effects that arise wherein answering one item influences students’ performance on the other item.

Supplemented Representations

Items of supplemented representation were designed to evaluate changes in the performance of students randomly assigned to one of three versions of an assessment item wherein a symbolic representation (S) alone is presented, or supplemental representations such as a particulate diagram (P) or table (T) are also provided. Negligible differences in students’ distractor selection were detected when supplemented with particulate diagrams. Small, substantive effects on the distractors students selected were detected for supplementing with tables in the molarity and gas laws items (where a column for “moles” was included; see Table 22 on the next page).
Table 22. Chi-square tests comparing the number of students who selected each of 5 distractors for items of supplemented representation.

<table>
<thead>
<tr>
<th>Item</th>
<th>Percent Yield (Fig. 42)</th>
<th>Molarity (Fig. 43)</th>
<th>Gas Laws (Fig. 44)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>P</td>
<td>T</td>
</tr>
<tr>
<td>n</td>
<td>605</td>
<td>595</td>
<td>600</td>
</tr>
<tr>
<td>Correct answer (%)</td>
<td>68.8</td>
<td>67.9</td>
<td>70.2</td>
</tr>
<tr>
<td>Skipped mole ratio (%)</td>
<td>13.4</td>
<td>15.0</td>
<td>14.0</td>
</tr>
<tr>
<td>Flipped mole ratio (%)</td>
<td>6.1</td>
<td>4.9</td>
<td>5.2</td>
</tr>
<tr>
<td>Other (%)</td>
<td>5.8</td>
<td>5.9</td>
<td>5.7</td>
</tr>
<tr>
<td>Other (%)</td>
<td>6.0</td>
<td>6.4</td>
<td>5.0</td>
</tr>
<tr>
<td>$\chi^2$ (w)</td>
<td>Ref. 1.493 (0.04)</td>
<td>1.153 (0.03)</td>
<td>Ref. 4.977 (0.06)</td>
</tr>
</tbody>
</table>

*p < 0.025.

The greatest difference observed in students’ distractor selections pertained to the “skipped mole ratio” distractor, wherein students fail to apply a stoichiometric ratio between two chemical species. For the “percent yield”, “molarity”, and “gas law stoichiometry” problems, the percent difference in students who selected the distractor was +0.4%, -4.3%, -18.5%, respectively. This suggests 1) the inclusion of a table with a column for “moles” reduces this selection of a distractor reflective of a common error, and 2) the impact of these tables on students’ performance may increase with the difficulty of the stoichiometry assessment item provided.

Discussion

The Representation used in an Assessment Task has a Significant Impact on the Distractors to which Students Attend

Pairwise comparisons suggest a significant relationship between students’ distractor selections and whether particulate representations were substituted for symbolic representations in enacting similar tasks (items shown in boxes 3 and 4). Effect sizes for these items ranged from medium, $w = 0.28$, to large, $w = 0.50$ (see Table 19) and represented a
decrease in percent correct, an increase in the selection of the insoluble distractor with the dissolution question, and the skipped mole ratio distractor with the stoichiometry question. A possible explanation of this result is the difficulty students face when translating between symbolic and particulate representations. Gkitzia et al. recently published an investigation of the student proficiency with translations between representations at the macroscopic, submicroscopic, and symbolic levels and found that students struggle most with translating between symbolic and particulate representations. The article concludes that descriptive and explanatory assessments of the translation between representations present a valuable tool in teaching and can refine students’ mental models of the particulate nature of matter.

While student performance and equity were negatively impacted by the substitution of particulate for symbolic representations, the inclusion of particulate diagrams in assessment tasks may still be warranted. Evidence supporting the use of particulate representations for promoting visual literacy and problem-solving in chemistry courses, also suggest translations between particulate representations could support students’ conceptual models of the particulate nature and states of matter, stoichiometric quantities, and charges that are often communicated using symbolic representations. The inclusion of particulate diagram representations as a supplement in addition to symbolic (Figures 42-44), particulate diagrams had little to no detectable impact on students’ selection of distractors (w = 0.04 to 0.06; see Table 21) and may be a means to promote these translations without impacting student performance.

As such, instructional support with the illustration, interpretation, and application of particulate representations in addition to the assessment of these skillsets could promote students’ development of conceptual models of chemical phenomena. A modeling framework presented in physics education suggests students rely upon representations to interpret models of phenomena stating, “no single representation is sufficient, and students (and experts) must continually translate amongst these representations in their attempts to explain phenomena.” It
is plausible different representations elicit different aspects of these students’ conceptual models, suggesting pertinent implications for researchers and practitioners to consider the role of representation in alignment to an assessment task. Further investigation as to changes in students’ responses processes when solving problems that vary in representation could prove informative in describing the various aspects of students’ conceptual models these representations may elicit.

**Supplementing Assessment Tasks with Tables can Improve At-Risk Student Performance**

The results of this study suggest at-risk student performance can be enhanced by as much as 19.2% when supplementing stoichiometric assessment tasks with tables including a column for “moles” (as observed in the molarity and gas laws items in Figures 43 and 44 but not for the percent yield item in Figure 42). While the relationship between students’ proficiency with mathematics and academic performance in chemistry is well-established,26,28–35,52–54 how at-risk chemistry students struggle with the material and what can be done to support their success is not yet understood. Donovan and Wheland, while investigating the impact of institutional implementations for mathematics prerequisites on the percent of at-risk chemistry students passing first-semester general chemistry question whether the performance of at-risk students is dependent on actual mathematics knowledge.55

As the data presented in this study suggest, at-risk student performance could be improved by embedding opportunities by which students could better conceptualize the quantities, units, and chemical identities with which they solve problems. This supports the hypothesis presented in a prior study that students overall struggle to identify meaningful relationships between stoichiometric quantities and the chemical identities to which these quantities are applied.41 By supplementing representations with tables, students may consider and organize these relationships more meaningfully, reducing students’ reliance on rote algorithmic execution. The application of tables for this purpose could be extended to other
representations (e.g., graphs and graphic organizers) intended to depict relationships between units, variables, and physical phenomena. Prior works evaluating student reasoning with graphs in the contexts of chemical kinetics,\textsuperscript{56,57} could inform these and other extensions of this work to novel investigations concerning the role of mathematical representations in chemistry education research.

**Modeling and Assessing Translations between Representations of Chemical Phenomena**

As a hypothesized extension of this work beyond stoichiometry, instruction modeling the solutions to chemistry problems using translations between symbolic and particulate representation could promote student understanding. Consider the use of the symbolic and particulate representations in Figure 46 (next page) for a classroom discussion regarding spectator ions. A procedure for determining the net ionic equation of a reaction could be described as 1) ensure the chemical reaction is balanced, 2) rewrite as a complete ionic reaction, 3), cross out ions that appear similarly in the reactants and products such as “2 Cl\(^-\)(aq)”, and 4) rewrite as a net ionic equation without spectator ions. If the instruction of this skillset is left only to the algorithm, assessments would likely measure students’ ability to recognize the prompt as a “execute the net ionic procedure” and enact the algorithm by rote.

\[
Zn(s) + 2 \text{HCl}(aq) \rightarrow Zn\text{Cl}_2(aq) + H_2(g)
\]

\[
Zn(s) + 2 \text{H}^+(aq) + 2 \text{Cl}^-(aq) \rightarrow Zn^{2+}(aq) + 2 \text{Cl}^-(aq) + H_2(g)
\]
Students could benefit from an instructor using both representations to model discussion as to
1) how the neutral Zn(s) in the reactants yields Zn^{2+}(aq), 2) differences in the rates of
consumption of HCl(aq) and Zn(s), 3) how the production of gaseous H_{2} from HCl(aq) occurs
and what this could mean for concentration, pressure, and other extensive properties of matter,
4) differences in the manner with which gaseous, aqueous solutions and solids are depicted in
these representations, and 5) the inaction of the spectator ion Cl^{-}(aq) between the reactants
and products. As students matriculate through the course, these translations between the
symbolic and particulate nature of matter could inform a students’ conceptual model for electron
transfer in chemical reactions supporting their reasonings’ with oxidation-reduction reactions
and later coursework in organic and biochemistry.

**Limitations and Future Works**

This study was exploratory in nature, using a small number of assessment items
representative of a few topics. While promising for the advancement of student performance
and equity when applied across assessments, more investigations with the impact of
representations are needed to evaluate best practices for the use of representations to develop
useful, meaningful learning experiences. While paired problems is a commonly enacted
methodology in chemistry education research,^{6-10,58-60} it is subject to sources of noise including
potential priming effects (wherein students’ responses to one item are influenced by having
responded to another) and differences in task, prompt, or distractor rationales. Substitution and
supplementation offer a means to minimize these sources of noise observed in the paired
problem approach and could be enacted in future works examining implications of assessment
design.
Conclusion

A primary conclusion of the work is the role representations can have on students’ assessment performance and the distractors to which they attend. Each of the representations evaluated (particulate diagrams and tables) impacted student responses distinctly suggesting representations may elicit unique aspects of students’ understanding of chemical phenomena. For example, particulate diagrams promoted student attention to the solubility of symbolically represented compounds in addition to the differences in the roles of subscripts, coefficients, and superscripts towards denoting quantity and charge.

Secondly, assessment designers can leverage the use of representations to promote more equitable achievement among chemistry students of diverse preparations in mathematics. Tables were potentially helpful in scaffolding students’ stoichiometric reasonings wherein the moles of one chemical is explicitly identified as different from the moles of another chemical, helping to prevent a common error in which students apply moles of substances interchangeably. This may be reflective of a need for students to reorganize and reconceptualize algorithmic solution processes that may otherwise be learned by rote. Representations provide opportunities for varied assessments that may promote equity and reward students who have pursued more explicative conceptions of chemistry concepts rather than promoting algorithmic proficiency with repetitive task execution for assessment tasks predominately represented symbolically.

Finally, the topic and task communicated in an assessment item may moderate the utility of representations. For these predominately stoichiometric tasks, particulate representation either had no impact or proved detrimental to student success. This may again reflect a reliance on rote learning facilitated by algorithmic assessment tasks communicated via symbolic representation. Alternatively, these results could encourage novel perspectives concerning which representations are best suited for particulate topics and assessment tasks. Replication
to determine the generalizability of these results to other research settings and to extend research concerning the role of representations in assessment design to include response processes and other novel representations appears warranted.

References


CHAPTER 7:
CONCLUSIONS AND FUTURE DIRECTIONS

Relevance of the Work Presented to Chemistry Education Research

Addressing the “Describe” Gap in the Literature

In chapter 1 of this dissertation, past research exploring the inequity observed of chemistry students with lower standardized math test scores was mapped on to Lynch’s equity schema for scientific reform (see Figure 2).\(^1\) There was a surfeit of articles effectively identifying the relationship between students’ math test score and academic chemistry performance. These articles were nested in the “investigate” stage of the framework. There were also many articles sharing the efficacy of interventions based entirely on the existence of this relationship between math in chemistry (the “adapt” and “evaluate” stages of the equity schema). A dearth of research was identified for the “describe” stage wherein empirical works seek to understand why and how these disparities persist between chemistry students differing in math test score.

The efforts of this dissertation have contributed unique insights to the “describe” component of Lynch’s framework providing necessary information for designing interventions based, not on the relationship identified, but on the sources of differential performance and what these may mean for the systems in place potentially propagating these inequities (see Figure 47 on the next page). The contribution of these works extends beyond prior works identifying the relationship between students’ standardized math test scores and academic chemistry performance, by describing the challenges students experience with the content material and providing an explanatory basis for differences in student performance. While providing
information that can be used for the design of more effective prerequisite, remedial, and supplemental coursework, these works also present considerable resources for research in practice to support the advancement of equity and rigor in general chemistry courses.

Figure 47. Citations related to each stage of Lynch’s equity schema demonstrating a progression from investigate to adapt and evaluate with limited research describing the nature of the disparities observed and the papers published of these works contributing to the literature base.

**Implications for Research and Practice**

**Randomization Tests (Chapter 3)**

Chapter 3 describes a methodology not previously identified in discipline-based education research known as Randomization Tests wherein the effect of an instructional intervention can be quantified with sample sizes as small as an $N = 1$, without the need for statistical software beyond spreadsheets, and that relies on the participants baseline behavior to serve as the control. This experimental methodology provides the strengths of randomized control trials to educational settings and serves to improve insights in past research examining the impact of evidence-based practices through the lens of methodological pluralism (the need
for efficacy research to be conducted rigorously using a variety of methodologies to determine the staying power of an intervention). Instructors could implement this methodology to evaluate trends in the performance of their student(s) over time. Likewise, researchers could evaluate the effects of changes enacted on a course-wide, or institutional, level.

The work presented in Chapter 3 applied Randomization Tests to investigate the impact of collaborative learning pedagogies (peer-led team learning, specifically) on student performance and equity between students of higher and lower standardized math test scores. Large effect sizes for student performance in favor of collaborative learning were observed in three of the four instructors’ classrooms spanning multiple semesters (see Table 9). From an equity perspective, two instructors observed similar differences in student performance between groups of higher and lower standardized math test score independent of the pedagogy enacted. The other two instructors, however, observed a reduction in the differential performance suggesting a collaborative learning pedagogy paired with instructor-level differences resulted in a more equitable learning environment for general chemistry students (see Figure 11). Descriptions of these instructor-level differences could not be identified with the data collected, providing promising avenues for future research.

*Incommensurate Topics (Chapter 4)*

Chapter 4 identifies the mole concept, often involving conversions between macroscopic quantities (e.g., volumes, masses) and microscopic quantities (e.g., moles), and stoichiometry, commonly requiring conversions between chemical species using microscopic quantities (e.g., moles of one chemical to moles of another), as the topics on which chemistry students with lower math test scores most differentially perform (see Figure 14).

Implications for practice are detailed in the context of prior research toward promoting students’ conceptualization of stoichiometry as advanced by other researchers. The chapter
ends with a call for future research characterize the tasks on which at-risk students struggle with beyond that of the over-arching topic.

An Explanative Basis (Chapter 5)

In chapter 5 of the dissertation, the call issued in chapter 4 for characterizing the tasks at-risk students differentially struggle on was initiated. Despite comparable proficiencies with the representational competency, systematic approaches, and conceptual understanding required for the mole concept and stoichiometry assessment items (see Table 14), students’ success at implementing chemically implausible algorithms was the most apparent source of differential performance between chemistry students of lower and higher math aptitude scores (see Table 15).

These findings suggest assessment design is of great consequence, not only in the equity of chemistry courses, but in developing assessments likely to elicit more explanatory models from students that could be applied in future coursework throughout the sciences. Students’ response processes were compartmentalized into three categories (representational competence, systematic process, and conceptual understanding and further disaggregated into codes listed in Table 13). In the discussion of this chapter, an acknowledgement is made as to how this compartmentalization reflects other efforts in research and practice but criticizes this approach as what was of challenge to students was representative of enacting all three skillsets in concert. A number of potential avenues for supporting students’ effective conceptualizing of stoichiometric algorithms are described:

- Scaffolding and modeling for students the conceptual differences between moles of one chemical and moles of another,
- adopting a process-oriented (rather than product-oriented) approach in assessment design (see Figure 38),
• removing (or providing explicit, conceptual instruction on interpreting) the phrase “in excess” (see Figure 39),
• ensuring modelled solution processes are not collapsed into a dimensional analysis “ladder” but into individual steps with explorations of what each numerical product means before it is inputted to another calculation (see pgs. 200-203),
• adapting what are commonly referred to as ICE (initial, change, equilibrium) tables to BCA (before, change, after) tables to teach stoichiometric reasoning to serve as both a scaffold (rows and columns facilitating different calculations) and opportunities for conceptualization (the moles of reactants and products may not be conserved, but mass is always conserved; see Table 16).

These approaches warrant (and ideally would elicit) future research but could prove promising recommendations for practice.

*Representations in Assessment (Chapter 6)*

Chapter 6 describes a study introducing novel explorations as to the impact of representations in assessment design. The work identifies a relationship between the representations shown, students’ performance, and the distractors to which students attend (see tables 17-21). Implications call for further research as to the impact of multi-model, or more than one type of, representation on students’ assessment performance in chemistry (see p. 242-245). Tables 18 and 21 provide evidence as to the use of tables to promote student success (overall) and support the successes of chemistry students with lower math test scores. On pgs. 245-246, a hypothesis made concerning the role of tables in assessments, related to scaffolding and organizing student conceptualizations, that could be applied more broadly to the physical sciences.
A Dissertation, Synthesized

The necessity for understanding common sources and content-related challenges of students experiencing inequitable outcomes these early STEM courses is underscored by these works. Previous efforts (vide supra) have sought to identify a disparity and immediately intervene, however, these works seek to expand on the importance of describing the nature of a disparity prior to seeking intervention.

These investigations of the pedagogies, topics, skillsets, and representations impacting differential performances observed of chemistry students with higher and lower standardized math test scores suggest a viable option for promoting more equitable general chemistry courses is to assess proficiencies with the material beyond recursive plug-and-chug. While algorithms are a useful aspect of the physical sciences intended to simplify complex calculations, so too should algorithms be grounded in conceptual understanding. Conceptualizing algorithms is a novel concept advanced by these works progressing beyond a dichotomous perspective of whether an assessment is predominately algorithmic or conceptual in nature and recognizing that students can approach algorithmic tasks with varying degrees of conceptualization. By assessing the degree to which a student has conceptualized algorithms, particularly in the context of chemical systems, it is possible one could improve both the equity and rigor observed in students’ early STEM coursework.

Scaffolding and organizing student responses, particularly in the form of tables, seems a promising endeavor for research and practice. Explorations as to why these tables support student performance and how students interact with these tables in authentic assessment settings warrants further description with insight through relevant theoretical frameworks for reasoning and conceptualization in the physical sciences. For example, physics education research has produced two potential theoretical frameworks for how tables may support student performance: model-based reasoning,\(^2\) and epistemic games as an extension of the resource model.\(^3–7\)
Future works could seek to expand upon this research exploring why and how tables support student cognition, develop interventions intended to improve proficiency with the mole concept and stoichiometry for chemistry students with lower math test scores, and seek parallel opportunities through which students can learn to conceptualize the algorithms chemists (and other physical scientists) find so helpful in mastering the course content and understanding physical phenomenon.

Impacts Beyond Chemistry Education Research

Diversity and Inclusion in the STEM Workforce

To reduce the inequity concerning how these topics are assessed, instructors and researchers should actively seek out processes that invoke a conceptual understanding of stoichiometry. As students of the at-risk cohort were also comprised of a greater number of students having self-identified as Black/African American and Hispanic/Latino and thus studies aimed at improving at-risk student outcomes have potential for ameliorating both differences observed amongst students of differential math preparation and towards to diversification of STEM careers.

Prerequisites and Remedial Coursework are Insufficient Strategies, Calls for Interdisciplinary Cooperation

The communicated need for more interdisciplinary cooperation between mathematics and chemistry instruction has been advocated, as the role of mathematics in supporting quantitative elements of chemistry becomes increasingly relevant. One suggestion observed with regard to ensuring incoming chemistry students possess necessary skills in mathematics is the assignment of prerequisite courses in mathematics, based on incoming metrics such as high school GPA, math ACT, or SATM scores. In a study investigating math reasoning with and
without a chemistry context, chemistry students were provided paired sets of chemistry and mathematics assessment items.\textsuperscript{12}

Chemistry assessment items were related to the topics of mole concept and stoichiometry and mathematics assessment items were designed to measure student ability in the quantitative reasoning skills necessary to solve the chemistry assessment items without a chemistry context. The study found that success on the mathematics assessment items often did not transfer to paired, chemistry items in which similar quantitative reasoning skill sets are utilized. As the author states:

“Since the mathematics questions are analogous to the chemistry questions, any practice at one should transfer some improved ability at the other; however, this does not appear to translate from mathematics to chemistry.” (p. 336)\textsuperscript{12}

In consideration of these works, the simple placement of a prerequisite course in mathematics seems unlikely to ameliorate differential performance within chemistry, though it is possible that a collaboratively designed preparatory course in math and chemistry may resolve the lack of transferability between the content matter of the two disciplines.

**Advancing Understandings as to How Tables Impact Students’ Conceptualization of Algorithms Applied throughout the Physical Sciences**

Whether applied as a formative learning exercise or supplementary representation on a summative assessment, tables appear to be a meaningful strategy for scaffolding students’ solution processes and can be designed to promote attention to essential features of a problem. For stakeholders that seek to promote student success, a plausible and cost-effective intervention could involve providing tables wherein the rows or columns separate measures (moles, mass, molar mass, etc.) and the corresponding chemical identities. These tables could serve as a schematic for students to use throughout instruction and assessment. This approach
could be used more broadly in STEM courses outside of chemistry that are particularly reliant on mathematics (e.g., physics and engineering).

References

APPENDIX A:

INSTITUTIONAL REVIEW

6/18/2014

Scott Lewis, Ph.D.
USF Department of Chemistry
4202 E. Fowler Ave. CHE205
Tampa, FL 33620

RE: Expedited Approval for Initial Review
IRB#: Pro00017861
Title: Improving Large Lecture Gateway Chemistry Courses through Flipped Classes with Peer-Led Team Learning (NSF #1432085)

Study Approval Period: 6/18/2014 to 6/18/2015
Dear Dr. Lewis:
On 6/18/2014, the Institutional Review Board (IRB) reviewed and APPROVED the above application and all documents outlined below.

Approved Item(s):
Protocol Document(s):
IRB Research Protocol Gateway Courses.docx
Consent/Assent Document(s)*:
IRB Gateway Informed Consent.docx.pdf

*Please use only the official IRB stamped informed consent/assent document(s) found under the "Attachments" tab. Please note, these consent/assent document(s) are only valid during the approval period indicated at the top of the form(s).
It was the determination of the IRB that your study qualified for expedited review which includes activities that (1) present no more than minimal risk to human subjects, and (2) involve only procedures listed in one or more of the categories outlined below. The IRB may review research through the expedited review procedure authorized by 45CFR46.110 and 21 CFR 56.110. The research proposed in this study is categorized under the following expedited review category:

(5) Research involving materials (data, documents, records, or specimens) that have been collected, or will be collected solely for nonresearch purposes (such as medical treatment or diagnosis).

(6) Collection of data from voice, video, digital, or image recordings made for research purposes.

(7) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

Your study qualifies for a waiver of the requirements for the informed consent process for records review, as outlined in the federal regulations at 45CFR46.116 (d) which states that an IRB may approve a consent procedure which does not include, or which alters, some or all of the elements of informed consent, or waive the requirements to obtain informed consent provided the IRB finds and documents that (1) the research involves no more than minimal risk to the subjects; (2) the waiver or alteration will not adversely affect the rights and welfare of the subjects; (3) the research could not practicably be carried out without the waiver or alteration; and (4) whenever appropriate, the subjects will be provided with additional pertinent information after participation.

As the principal investigator of this study, it is your responsibility to conduct this study in accordance with IRB policies and procedures and as approved by the IRB. Any changes to the approved research must be submitted to the IRB for review and approval by an amendment.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

[Signature]
John Schinka, Ph.D., Chairperson
USF Institutional Review Board
2/4/2015

Li Ye, M.S.
USF CITRUS - Center for the Improvement of Teaching and Research in Undergraduate STEM Education
4202 East Fowler Ave., CHE205
Tampa, FL 33620

RE: Expedited Approval for Initial Review
IRB#: Pro00020840
Title: Investigating Evidence for the Validity of Chemistry Assessments Methods


Dear Ms. Ye:

On 2/4/2015, the Institutional Review Board (IRB) reviewed and APPROVED the above application and all documents outlined below.

Approved Item(s):
Protocol Document(s):
IRB Protocol.pdf

Consent/Assent Document(s)*:
Informed Consent B.docx.pdf

*Please use only the official IRB stamped informed consent/assent document(s) found under the "Attachments" tab. Please note, these consent/assent document(s) are only valid during the approval period indicated at the top of the form(s).

It was the determination of the IRB that your study qualified for expedited review which includes activities that (1) present no more than minimal risk to human subjects, and (2) involve only
procedures listed in one or more of the categories outlined below. The IRB may review research through the expedited review procedure authorized by 45CFR46.110 and 21 CFR 56.110. The research proposed in this study is categorized under the following expedited review category:

(5) Collection of data from voice, video, digital, or image recordings made for research purposes.

(6) Research on individual or group characteristics or behavior (including, but not limited to, research on perception, cognition, motivation, identity, language, communication, cultural beliefs or practices, and social behavior) or research employing survey, interview, oral history, focus group, program evaluation, human factors evaluation, or quality assurance methodologies.

As the principal investigator of this study, it is your responsibility to conduct this study in accordance with IRB policies and procedures and as approved by the IRB. Any changes to the approved research must be submitted to the IRB for review and approval by an amendment.

We appreciate your dedication to the ethical conduct of human subject research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

Kristen Salomon, Ph.D., Vice Chairperson
USF Institutional Review Board
RE: Exempt Certification
IRB#: Pro00040459
Title: Exploring the Role of Chemistry Representation Translations in General Chemistry Success

Dear Mrs. Ralph:
On 5/20/2019, the Institutional Review Board (IRB) determined that your research meets criteria for exemption from the federal regulations as outlined by 45 CFR 46.104(d):

(5) Research, conducted in established or commonly accepted educational settings, that specifically involves normal educational practices that are not likely to adversely impact students’ opportunity to learn required educational content or the assessment of educators who provide instruction. This includes most research on regular and special education instructional strategies, and research on the effectiveness of or the comparison among instructional techniques, curricula, or classroom management methods.

As the principal investigator for this study, it is your responsibility to ensure that this research is conducted as outlined in your application and consistent with the ethical principles outlined in the Belmont Report and with USF HRPP policies and procedures.

Please note, as per USF HRPP Policy, once the exempt determination is made, the application is closed in ARC. This does not limit your ability to conduct the research. Any proposed or anticipated change to the study design that was previously declared exempt from IRB oversight must be submitted to the IRB as a new study prior to initiation of the change. However, administrative changes, including changes in research personnel, do not warrant an Amendment or new application.

We appreciate your dedication to the ethical conduct of human subjects research at the University of South Florida and your continued commitment to human research protections. If you have any questions regarding this matter, please call 813-974-5638.

Sincerely,

Kristen Salomon, Ph.D., Chairperson
USF Institutional Review Board
APPENDIX B:
COPYRIGHT PERMISSIONS

Royal Society of Chemistry

RSC Licenses, Copyright & Permissions

The Royal Society of Chemistry does not require a formal request of permissions to reuse a whole article in a thesis. "If you are including an article or book chapter published by us in your thesis please ensure that your co-authors are aware of this." Checked 12/15/2016.

RSC Journal List

List of journals published by the Royal Society of Chemistry.

ed. Policies are subject to change; make note of the date the policy was checked.

A screenshot of the Royal Society of Chemistry Publication copyright policy taken 01/24/2020:

Author reusing their own work published by the Royal Society of Chemistry

You do not need to request permission to reuse your own figures, diagrams, etc, that were originally published in a Royal Society of Chemistry publication. However, permission should be requested for use of the whole article or chapter except if reusing it in a thesis. If you are including an article or book chapter published by us in your thesis please ensure that your co-authors are aware of this.

Reuse of material that was published originally by the Royal Society of Chemistry must be accompanied by the appropriate acknowledgement of the publication. The form of the acknowledgement is dependent on the journal in which it was published originally, as detailed in 'Acknowledgements'.
Question 15: Do I have any rights to use material that I have authored, but for which ACS holds copyright? establishes that use in a thesis is permitted subject to the conditions in the Journal Publishing User Guide. Checked 12/12/2016.

- ACS Publications

Full-text access to 26 journals of the American Chemical Society. Includes illustrations, structures, figures, and supporting information.

A screenshot of the ACS Publication copyright policy taken 01/24/2020:

15. Do I have any rights to use material that I have authored, but for which ACS holds copyright?  

The current ACS Journal Publishing Agreement covers several permitted uses by authors. Permitted uses are subject to certain conditions (see the Journal Publishing Agreement User Guide [PDF]) and may vary according to the version of the paper in production (see question 14). Permitted uses of all versions include:

- Use in theses and collections of your own work
- Teaching and training
- Presentations at conferences and seminars
- Sharing with colleagues
- Posting on websites and repositories
- Sharing links to free eprints of the published work
- Use and authorizing the use of Supporting Information
- Other permitted uses by authors that are delineated in the Agreement
APPENDIX C:

TOPICS CODING SCHEME FOR FIRST-SEMESTER GENERAL CHEMISTRY (GCI)

The following major topics were the result of the iterative coding scheme and include the learning objectives that were affiliated with each major topic. The learning objectives were used to craft the exams.

1) Structure of the Atom (SA):
   a) Describe the structure of the atom in terms of the placement and charge of protons, neutrons and electrons.
   b) Describe the structure of the atom in terms of the number of protons, neutrons and electrons given mass number and chemical identity or atomic number.
   c) Define isotopes and ions in terms of the structure of the atom.
   d) Relate isotope abundance to the average atomic mass of an element.

2) The Mole Concept (MC):
   a) Define mole and relate it to number of units.
   b) Differentiate empirical formula and molecular formula.
   c) Define and solve the formula mass of a given compound.
   d) Convert between mass, mole and number of atoms for any compound.
   e) Given a chemical formula, solve for the mass percent of each element in a compound.

3) Nomenclature & Models of Bonding (NM):
   a) Describe the reason chemical bonds are stable and differentiate covalent and ionic bonds in terms of electron placement.
   b) Classify chemicals as atomic elements, molecular elements, molecular compounds or ionic compounds.
   c) Predict the ratio that cations and anions combine in an ionic compound.
   d) Name covalent and ionic compounds including ionic compounds with transition metals and polyatomic ions.

4) Stoichiometry (ST):
   a) Balance a chemical equation, given an unbalanced chemical equation.
   b) Given the mass of any compound and given a chemical reaction, solve for the mass of any other compound in the reaction.
   c) Given the mass of two reactants in a chemical reaction, solve for the mass produced of any product, determine which reactant is limiting and determine the mass remaining of the excess reactant.
   d) Given two of the three: percent yield, theoretical yield and actual yield, or a means to determine two of the three, solve for the third.

5) Molarity (MR):
   a) Given two of the three: molarity, mol, volume of solution, be able to solve for the missing variable.
   b) Perform calculations for solution dilutions using $M_1V_1 = M_2V_2$.
   c) Perform stoichiometric calculations for reactions in aqueous solutions.

6) Reactions in Solution (RS):
a) Define strong electrolyte, weak electrolyte, and nonelectrolyte.
b) Classify ionic and molecular compounds as strong, weak, or nonelectrolytes.
c) Identify the ions formed when an ionic compound is dissolved in water.
d) Determine the solubility of ionic compounds in water.
e) Given two ionic compounds that are dissolved in water, predict the possible products and identify if a precipitate forms.
f) Given two ionic compounds that are dissolved in water, write a molecular equation, ionic equation and net ionic equation and identify spectator ions.

7) Oxidation-Reduction (or Redox) Reactions (RR):
   a) Define and identify: reduction, oxidation, reducing agent, oxidizing agent and redox reactions.
b) Assign oxidation numbers to any compound.

8) The Gas Laws (GL):
   a) Define pressure in terms of molecular collisions.
b) Use the ideal gas law to solve for any missing variable.
c) Use the ideal gas law to determine molar volume, density, and molar mass of a gas.
d) Define and use mole fraction and partial pressure for a mixture of gases (Dalton’s law).
e) Relate stoichiometry calculations to the ideal gas law.
f) Define standard temperature and pressure (STP) and molar volume at STP.
g) Identify the three central parts of kinetic molecular theory and relate the theory to the observations in the simple gas laws.
h) Interpret graphical representations of distributions of molecular speeds.
i) Demonstrate the conceptual relationships between molar mass, temperature and the root mean square velocity of molecules.
j) Given any two of molar mass, temperature, and root mean square speed, be able to calculate the third.
k) Define mean free path, diffusion, and effusion.
l) Use Graham’s Law and the effusion rate or time of a known substance to solve for the rate, time, or molar mass of another substance.

9) Changes in Enthalpy (CH):
   a) Relate mass of a compound in a reaction, enthalpy change of a reaction and energy change of a reaction.
b) Describe the changes in ΔH when manipulating a chemical reaction (reversing, multiplying by a constant) as per Hess’s Law.
c) Use Hess’s law and ΔH of chemical reactions to solve for the ΔH of a different chemical reaction.
d) Relate ΔH of a reaction to heats of formation, ΔHf.
e) Write or identify a formation reaction for a compound from elements in the standard state (e.g. corresponds to the value for ΔHf).

10) Changes in Energy (CQ):
    a) Define kinetic energy, potential energy, chemical energy and state function.
b) Describe the first law of thermodynamics and use it to model energy changes.
c) Given three of the four: energy, mass, specific heat and change in temperature, or a means to solve three of the four, determine the value for the missing variable.
d) Define the concept heat capacity and relate it to specific heat capacity and molar heat capacity.
e) Define enthalpy, exothermic reaction, and endothermic reaction.
f) Perform calorimetry calculations relating mass of a reactant to change in temperature of surrounding water.

11) Properties of Light (PL):
   a) Define the term quantum mechanical model.
b) Characterize the different regions of the electromagnetic spectrum.
c) Describe the evidence for the wave-particle dual nature of light.
d) Relate energy, frequency and wavelength conceptually and mathematically.
e) Relate amplitude to intensity conceptually and mathematically.
f) Calculate and relate the concepts of threshold frequency, binding energy and kinetic energy of an ejected electron in the photoelectric effect.
g) Define the terms emission spectrum and absorption spectrum.
h) Relate deBroglie wavelength to mass and velocity conceptually and mathematically.
i) Explain the term complementary properties and the specific example in Heisenberg’s uncertainty principle.
j) Define probability density and contrast deterministic with indeterminacy.
k) Solve for the energy and wavelength associated with electron transitions in a Hydrogen atom and explain the relationship with the Bohr Model.

12) Electron Configurations and Quantum Numbers (EC):
   a) Describe the purpose of each of the four quantum numbers and use the rules that define allowable sets of quantum numbers.
b) Know the shapes of s, p, d, and f orbitals and the relationship to quantum numbers.
c) Define and apply the Pauli exclusion principle.
d) Define and apply the Aufbau principle and Hund’s rule.
e) Describe an orbital filling diagram for any element on the periodic table.
f) Relate orbital filling diagrams, electron configurations and quantum numbers.
g) Determine number of valence electrons and core electrons for any atom on the periodic table.
h) Determine the expected electron configuration for any atom on the periodic table (complete configuration and noble gas abbreviation).
i) Know and understand that an electron configuration shows the number of electrons that occupy particular orbitals in atoms and is the basis for chemical reactivity.
j) Write electron configurations of ions.
k) Define and make predictions for diamagnetic and paramagnetic.

13) Periodic Trends (PT):
   a) Define the term periodic property.
b) Define the term degenerate as it applies to orbitals.
c) Indicate the roles of Coulomb’s Law, shielding and penetration in sublevel splitting.
d) Describe the trends in atomic radii on the periodic table and relate the observed trends to the structure of the atom.
e) Relate the radius of an atom to an ion of the same element.
f) Describe the trends in ionization energy on the periodic table and relate the observed trends to the structure of the atom.
g) Predict the expected trends in successive ionization energies.
h) Define electron affinity.
i) Describe what is meant by metallic character and relate it to trends on the periodic table.

14) Lewis Structures (LS):
   a) Define and provide examples of ionic, covalent and metallic bonds and differentiate between them based on physical properties.
   b) Represent any atom with a Lewis structure.
   c) Use Lewis structures to represent covalent compounds or ions.
   d) Use Lewis structures to represent ionic compounds containing main group elements.
   e) Relate bond order to bond energy and bond length.
   f) Know that Lewis structures are simple predictors of how atoms combine to form ionic compounds and molecules.
   g) Define and describe the trends in electronegativity.
   h) Determine if a bond is considered covalent, polar covalent or ionic, given values for electronegativity and indicate the direction of the dipole.
   i) Use the formula for dipole moment and percent ionic character.
   j) Understand the resonance concept and relate it to relative bond strength and length.
   k) Solve for the formal charge of any atom in a Lewis structure and use formal charge to determine plausibility of a Lewis structure.
   l) Describe structures that are exceptions to the octet rule including odd-electron species, incomplete octets and expanded octets.
   m) Understand trends in bond length and bond energy and the relationship between bond length and bond energy.

15) Molecular Geometry (MG):
   a) Understand the premise to VSEPR theory, particularly the role played by electron groups.
   b) Determine the electron geometry and molecular geometry for any Lewis structure.
   c) Determine the bond angle among any three atoms in a Lewis structure.
   d) Determine the polarity of a bond, molecule, or ion given electronegativity values.

16) Valence Bond Theory (VB):
   a) Describe the principles of valence bond theory.
   b) For any Lewis structure, predict hybridization and number of sigma and pi bonds.
   c) Describe the orbitals that contribute to each hybridization scheme.
   d) Describe the relationship between hybridization and bond type.
   e) Define bonding, antibonding and nonbonding orbitals.
   f) Use molecular orbital theory to determine the bond order for diatomic molecules given the MO diagram.
APPENDIX D:

EFFECT SIZES OF TOPIC SCORES BY SEMESTER

Effect sizes comparing mean topic scores of not at-risk and at-risk students as determined from interim and ACS exams by semester (spanning Spring 2016 through are provided in Figures A1-4.

Figure A1. Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Spring 2016.

Figure A2. Effect sizes comparing mean topics scores of not at-risk and at-risk students on interim and ACS exams for Fall 2016.
Not all topics were included in analyses for each semester’s interim exams as a few topics did not have 3 or more assessment items with which to complete the analysis. In the U.S. system, the traditional academic year has a fall and spring semester. For the course first-semester General Chemistry the fall and spring semester cohorts have notably different profiles. In comparison to fall semester cohorts, spring semester tends to have lower overall SATM scores and lower differential performance (Figure A1 and A3).

Across each semester, stoichiometry (ST) has the highest differential performance among all topics on the overall average. Demarcating by interim and ACS shows a consistent trend of stoichiometry having among the highest differential with the interim tests in Spring 2017 as the only exception. Mole concept (MC) also features a consistent relative peak for three of the four semesters, with the exception of Spring 2016 owing to the interim test items from that semester. The exploration by semester offers an indication of the consistency of the finding for differential performance on stoichiometry and mole concept.
APPENDIX E:

RESULTS OF ADDITIONAL MISSING DATA ANALYSES

Missing data were analyzed to determine the extent possible the key finding was applicable to those with different sources of missing data. A frequency count of the different sources of missing data are presented in Table A.1. The first data row in the table, SATM with All Tests, comprise the original cohort analyzed. The fourth data row in the table represents students without SAT scores but with ACT scores and complete test scores and these students were analyzed as a separate group. The second and fifth data rows, combined, represent students with SAT or ACT who took the summative assessment but missed an interim assessment; this group was analyzed separately as well in the text and below in table A1.

Table A1. Frequencies of students within each subgroup of the overall cohort.

<table>
<thead>
<tr>
<th>Semester</th>
<th>Spring 2016</th>
<th>Fall 2016</th>
<th>Spring 2017</th>
<th>Fall 2017</th>
<th>Total</th>
<th>Percent of Cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Tests</td>
<td>416</td>
<td>1189</td>
<td>269</td>
<td>1105</td>
<td>2979</td>
<td>52.9%</td>
</tr>
<tr>
<td>Missed Test; Took ACS</td>
<td>60</td>
<td>148</td>
<td>85</td>
<td>147</td>
<td>440</td>
<td>7.8%</td>
</tr>
<tr>
<td>No ACS Score</td>
<td>143</td>
<td>125</td>
<td>35</td>
<td>67</td>
<td>370</td>
<td>6.6%</td>
</tr>
<tr>
<td>No SATM; Took ACT-M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Tests</td>
<td>99</td>
<td>304</td>
<td>122</td>
<td>365</td>
<td>890</td>
<td>15.8%</td>
</tr>
<tr>
<td>Missed Test; Took ACS</td>
<td>17</td>
<td>40</td>
<td>29</td>
<td>49</td>
<td>135</td>
<td>2.4%</td>
</tr>
<tr>
<td>No ACS Score</td>
<td>37</td>
<td>28</td>
<td>55</td>
<td>40</td>
<td>160</td>
<td>2.8%</td>
</tr>
<tr>
<td>Neither SAT nor ACT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All Tests</td>
<td>82</td>
<td>98</td>
<td>78</td>
<td>148</td>
<td>406</td>
<td>7.2%</td>
</tr>
<tr>
<td>Missed Test; Took ACS</td>
<td>16</td>
<td>15</td>
<td>41</td>
<td>35</td>
<td>107</td>
<td>1.9%</td>
</tr>
<tr>
<td>No ACS Score</td>
<td>32</td>
<td>36</td>
<td>43</td>
<td>38</td>
<td>149</td>
<td>2.6%</td>
</tr>
<tr>
<td>Total Students Enrolled</td>
<td>902</td>
<td>1983</td>
<td>757</td>
<td>1994</td>
<td>5636</td>
<td>100.0%</td>
</tr>
</tbody>
</table>
As mentioned, 575 students (10.2% of the cohort) with SAT or ACT who took the summative (ACS) exam but missed an interim exam were analyzed separately. At-risk and not at-risk were identified using SAT or ACT (if SAT was not available) and their performance was compared on the ACS exam by topic. The results are presented as the dashed line in Figure A5 and are consistent with those who have complete data (dashed line in Figure A5).

Figure A5. Mean effect sizes between students with complete data and students missing tests.
ABOUT THE AUTHOR

Vanessa Rosa Ralph attended school in the Hillsborough County school system in Florida, attaining a Bachelor’s in Science in Biomedical Sciences at the University of South Florida, a Master's in Science in Pharmaceutical Chemistry at the University of Florida, and her Doctorate in Philosophy in Chemistry at the University of South Florida. Ralph has taken instruments of suspect validity to evaluate aspects of her personality to determine that she is an INTJ (Introverted, Intuitive, Thinking, Judging), Type 8: Protector, Architect who is often sorted into either Slytherin or Ravenclaw house. Ralph looks forward to serving as a Research Associate for her post-doctoral research with Dr. Ryan Stowe at the University of Wisconsin-Madison and hopes to establish a lab at a doctorate granting institution.