Evaluation and Application of Instruments Measuring Spatial Ability and Attitude for College Chemistry Students

Xiaoying Xu
University of South Florida, xxu4@mail.usf.edu

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

Part of the Chemistry Commons

Scholar Commons Citation

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.
Evaluation and Application of Instruments Measuring Spatial Ability
and Attitude for College Chemistry Students

by

Xiaoying Xu

A dissertation submitted in 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: Jennifer E. Lewis, Ph.D.
Xiao Li, Ph.D.
Jeffrey R. Raker, Ph.D.
Robert F. Dedrick, Ph.D.
Robert Potter, Ph.D.

Date of Approval:
November 20, 2014

Keywords: reliability, validity, factor analysis, chemistry performance, higher education

Copyright © 2014, Xiaoying Xu
Acknowledgments

First, I would like to thank my major advisor, Dr. Lewis, for her patience and encouragement for me to develop the knowledge and skills in educational research and evaluation, and her tremendous amount of effort to help me accomplish this dissertation work. Without her support, my life during the past years might not have been so meaningful, and this dissertation work would be not possible. I also greatly appreciate the instrumental roles of the committee members, Drs. Robert Dedrick, Xiao Li, Robert Potter, and Jeffrey R. Raker, for the academic advice, feedback, and the commitment for my graduation work and manuscript preparation. Similarly I acknowledge the support and recommendations from my collaborators, Drs. Kim, Alhooshani, Southam, Brandriet, Bretz, Jiang, Garcia and Sachel Villafañe.

A special thanks to former and current chemical education research group members Alicia, Teresa, Keily, Yujuan, Janelle, Benjamin, Li, Todd, Adrian, and Scott for their support and friendship. Also I appreciate the help from many people including but not limited to: Drs. Sandi-Urena, Guida, Ferron, Kromrey, general chemistry classroom TAs, chemistry office staffs and my class instructors.

Last but not least, I am in gratitude to my family, Charley, Feng, Wangfeng, Lachun, Ying and Zhirong for the love and encouragement during the dissertation.
Table of Contents

List of Tables  v

List of Figures vii

Abstract viii

I. Introduction 1
  The Need for Transforming STEM Education 1
  Important Factors in STEM Education 2
  Spatial Ability 2
  Attitude toward Science 3
  Role of Measurement Tools in STEM Education 5
  Purposes of this Work 5
  References 10

II. Methods 15
  Instruments 15
  Data Collection and Participants 16
  Data Analysis Strategies 18
    Psychometric Analysis 18
      Internal consistency reliability 19
      Internal structure 20
    Relations to other Variables by Structural Equation Modeling 21
  References 22

III. Sex Difference in Spatial Ability for College Students and Exploration of
    Measurement Invariance 26
  Note to Reader 26
  Introduction 26
  Research Methods 30
    Instruments 31
    Participants and Data Collection 32
    Statistical Data Analysis 33
  Results and Discussion 35
    Descriptive Statistics 35
    Test-retest Relationship for Temporal Stability 36
    Confirmatory Factor Analysis 37
    Measurement Invariance between the Sexes 42
    Sex Difference in Spatial Ability 43
  Conclusion 45
References

IV. The Role of Spatial Ability in Students’ Progression through Organic Chemistry
   Note to Reader
   Why Spatial Ability?
   Research Questions
   Research Methods
   Results
      Descriptive Statistics
      Comparison of Students Who Can Or Cannot Mentally Visualize
      SEM for the Effect of Mental Rotations on Organic Chemistry Course Grade
   Discussion
   References

V. Attitude toward the Subject of Chemistry in Australia: An ALIUS and POGIL
   Collaboration to Promote Cross-national Comparisons
   Note to Reader
   Introduction
      Instrument
      Sample
      Data Analysis Methods
   Results
      Descriptive Statistics
      Validity Results from CFA
      Reliability
      Factor Score Comparison
   Conclusions
   Acknowledgements
   References

VI. Gathering Psychometric Evidence for ASCIv2 to Support Cross-Cultural Attitudinal
    Studies for College Chemistry Programs
   Note to Reader
   Introduction
      The Need for Attitude
      Measurement for Attitude
      Research Questions
      Settings
   Research Methods
      Instrument
      Participants and Data Collection
      Data Analysis
   Results
      Descriptive Statistics for Item Scores for KU
      Two-Factor CFA Model Fit
      Internal Consistency Reliability for ASCIv2
VII. College Students’ Attitudes toward Chemistry, Conceptual Knowledge and Achievement: Structural Equation Model Analysis

Introduction

Student Achievement in College Chemistry
Attitude toward Chemistry and Relationship with Achievement
Prior Conceptual Knowledge and Relationship with Achievement
Math Ability and Relationship with Achievement
Conceptual Relationship among the Four Factors in Chemistry
Structural Equation Modeling (SEM)
Present Study

Method

Instruments

Attitude
Prior conceptual knowledge
Math ability
Chemistry achievement

Data Collection and Participants

Model Specification for SEM Analysis

Data Analysis

Descriptive Statistics and Assumption Checking for SEM
Structural Equation Model (SEM)

Results

Descriptive Statistics and Assumption Checking for SEM
Structural Equation Model Analysis

Model fit
Decomposition of relationships
Measurement model for ASC Iv2
Measurement part for TMI
Relationships among the three predictors
Relationships with student achievement in college chemistry

Discussion and Implications

References

VIII. Conclusions and Implications

Summary Conclusion
Limitations and Implications
References

Appendices

Appendix A: Administration instruction for ROT
<table>
<thead>
<tr>
<th>Appendix</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Administration instruction for ASCIv2</td>
<td>174</td>
</tr>
<tr>
<td>C</td>
<td>Supplement for Chapter 3</td>
<td>175</td>
</tr>
<tr>
<td>D</td>
<td>Supplement for Chapter 4</td>
<td>177</td>
</tr>
<tr>
<td>E</td>
<td>Supplement for Chapter 7</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Figure notes for Figures 7.2-7.6</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Additional analysis</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Measurement Invariance for ASCIV2 to Measure Attitude toward Chemistry</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Unique Contribution of Spatial Ability When Attitude Is Controlled</td>
<td>178</td>
</tr>
<tr>
<td></td>
<td>References for Additional Analysis</td>
<td>179</td>
</tr>
</tbody>
</table>
List of Tables

Table 2.1 Summary of instruments administration 17
Table 3.1 Demographic Information for the Study Sample: SAT Math by Sex 31
Table 3.2 Cronbach’s Alpha and Descriptive Statistics for ROT by Semester 36
Table 3.3 Descriptive Statistics of ROT Scores for Retakers 37
Table 3.4 Time Interval for Retaking the ROT Test 37
Table 3.5 Fit Statistics of Three Models for ROT in Spring 2010 38
Table 3.6 Fit Statistics of ROT for Three Models for Three Semesters 42
Table 3.7 Measurement Invariance Result across Sexes for CFA Bi-factor Model 43
Table 3.8 Effect Size from MCFA Bi-factor Strong Invariance Model and Cohen’s d 43
Table 3.9 Sex Difference by Degree of Angular Disparity for Two Item Types 45
Table 4.1 Descriptive of SAT Math, ROT, ACS, and Organic Chemistry Grade by Sex 60
Table 4.2 Pearson Correlations among SATM, ROT, and ACS Scores, Course Grade for Organic Chemistry I and II 61
Table 4.3 Mental Rotation vs Progression through Organic Chemistry Course Sequence 62
Table 4.4 Standardized Path Coefficient and Effect Size for Chemistry Performance 66
Table 5.1A Descriptive Statistics for Curtin after Recoding Items 1, 4, 5 and 7 81
Table 5.1B Descriptive Statistics for the US University for Comparison 81
Table 5.2 CFA Model Fit for the 2-Factor Solution 82
Table 5.3 CFA Item Loadings for the 2-Factor Solution 83
Table 5.4  Internal Consistency Reliability by Cronbach's Alpha for ASCIv2  
Table 5.5  Factor Scores for each Subscale for US and Curtin Students  
Table 6.1  Descriptive Statistics of Item Scores for KU  
Table 6.2  CFA Item Loadings for the 2-Factor Solution and Item-total Correlation within Each Intended Subscale  
Table 6.3  Inter-rater Agreement Data For Whether an Interpretation of Item 6 Is a Good Indicator of The Intended Scale  
Table 6.4  CFA Model Fit for the 2-Factor Solution Without Item 6  
Table 6.5  Factor Scores for each Subscale for KU, WU and SE  
Table 7.1  Variables in ASCIv2 Instrument  
Table 7.2  Items in TMI Instrument  
Table 7.3  Demographics: Number and Percentage of Students by Sex and Race/Ethnicity  
Table 7.4  Descriptive Statistics for each Measured Variable  
Table 7.5  Fit Indices and Predicted Variance for Models  (N = 963)  
Table 7.6  Standardized Correlation Coefficients  
Table 7.7  Standardized Path Coefficient and Effect Size for the Regression on Achievement  
Table A3.1  Item Difficulty, Item-total Correlation, and Discrimination Index by Semester  
Table A4.1  Descriptive Analysis for SEM Variables  
Table A7.1  Fit Result for Multiple-group CFA Models with Measurement Equivalence Test across Sexes
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 2.1</td>
<td>Evidence Sources for Psychometric Quality</td>
<td>19</td>
</tr>
<tr>
<td>Figure 2.2</td>
<td>An Example of a Full SEM Model</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3.1</td>
<td>Rot Item 9</td>
<td>30</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Rotation Pattern for Item 9</td>
<td>39</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Standardized Parameter Estimates for One-Factor (3A), Second-Order (3B) And Bi-Factor (3C) Alternate Models</td>
<td>42</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Assign the Configuration of 2-Butanol Using Visualization</td>
<td>54</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Comparing Students Who Progressed through Organic Chemistry Or Not</td>
<td>62</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>SEM Model with Unstandardized Parameter Estimate</td>
<td>64</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>SEM Model with Standardized Parameter Estimate</td>
<td>65</td>
</tr>
<tr>
<td>Figure 6.1</td>
<td>Box-and-Whisker Plot for KU Students Overlapped with Asciv2 Items</td>
<td>102</td>
</tr>
<tr>
<td>Figure 7.1</td>
<td>Conceptual Model of Chemistry Achievement and Three Predictors (Model 1)</td>
<td>121</td>
</tr>
<tr>
<td>Figure 7.2</td>
<td>SEM Model 2 with Three Predictors</td>
<td>129</td>
</tr>
<tr>
<td>Figure 7.3</td>
<td>SEM Model 3 with Two Predictors</td>
<td>129</td>
</tr>
<tr>
<td>Figure 7.4</td>
<td>Standardized Parameters for the Path Analysis Model 1</td>
<td>135</td>
</tr>
<tr>
<td>Figure 7.5</td>
<td>Path Diagram with the Standardized Parameter Coefficients for Model 2</td>
<td>135</td>
</tr>
<tr>
<td>Figure 7.6</td>
<td>Path Diagram with the Standardized Parameter Coefficients for Model 3</td>
<td>135</td>
</tr>
<tr>
<td>Figure A4.1</td>
<td>Alternate Models</td>
<td>176</td>
</tr>
</tbody>
</table>
ABSTRACT

Student performance in college chemistry courses remains a concern for educators seeking to help improve the future STEM workforce. Besides math ability and prior chemistry knowledge, spatial ability and attitude toward chemistry have been considered as important factors influencing college chemistry performance. This work includes five studies and uses data collected from instruments to examine the relationships of these two factors – spatial ability and attitude – with student chemistry performance, and provides psychometric evidence for using the Purdue Visualization of Rotations test (ROT) and Attitude toward the Subject of Chemistry Inventory (ASCIv2) to measure these two factors, respectively, in college chemistry classrooms.

The first two studies are about the evaluation and application of the ROT to measure the spatial ability of college chemistry students. Study 1, Sex difference in spatial ability for college students and exploration of measurement invariance, examines whether ROT items function in the same way between sexes. This study provides evidence that the ROT is an appropriate tool to measure the mental rotation component of spatial ability for college students. Results indicate the bi-factorial structure of ROT items, with each of the five items sharing the same rotation pattern. Thus, the item score is influenced by the general construct of mental rotation ability and the item rotation pattern. In addition, ROT items function in the same way across sexes; therefore, they can support the use of ROT for examining sex difference with less concern about test bias. A sex gap favoring males persists in the data.
The second study, *The role of spatial ability in students’ progression through organic chemistry*, examines whether students who scored low on the ROT in general chemistry are less likely to advance through organic chemistry, and the relationship of spatial ability to organic chemistry course grades for students who completed organic chemistry courses. The findings indicate that the ability to perform mental rotations, as measured by the ROT, does not present a barrier for students seeking to advance and succeed in the organic chemistry course sequence in the current setting after other factors, such as math ability, are held constant. For students who were at the low end of ROT, alternate approaches can be used to compensate for the lack of mental rotation ability to solve spatial-related chemistry problems, as suggested by the literature.

The next three studies are about the evaluation and application of the ASCIv2 to measure student attitude toward chemistry. The ASCIv2 is based upon a previous publication, *Refinement of a chemistry attitude measure for college students*. Study 3, *Attitude toward the subject of chemistry in Australia: An ALIUS and POGIL collaboration to promote cross-national comparisons*, and Study 4, *Gathering psychometric evidence for ASCIv2 to support cross-cultural attitudinal studies for college chemistry programs*, gather psychometric evidence for the use of ASCIv2 internationally. The purpose for these two pilot studies is to investigate how ASCIv2 functions at multiple sites where the Process Oriented Guided Inquiry Learning (POGIL) is implemented with one of the objectives being to foster a positive attitude toward the subject in general. One student cohort is from a university in Australia, and the other is from a Saudi Arabian institution. The findings suggest ASCIv2 can be an appropriate tool to measure attitude in an Australian context; however, Saudi Arabian student cohorts responded differently to the item (chemistry is challenging vs. unchallenging), which raises concern about using these items to measure the attitude subscale for students in Saudi Arabia. As compared to Australian and
Saudi Arabian students, US students at SE exhibited a negative attitude towards chemistry. Special attention should be paid to ways of getting students more interested in Chemistry.

The last study, *College students’ attitudes toward chemistry, conceptual knowledge and achievement: structural equation model analysis*, examines the contribution of attitude to chemistry performance when math ability and prior chemistry conceptual knowledge are controlled. Results reveal that a full SEM model using three predictors at the beginning of the semester can explain a very high percentage (69%) of the variance in chemistry achievement at the end of course. Both prior conceptual knowledge and attitude toward chemistry contribute a significant unique portion to the prediction of chemistry achievement when controlling for math ability. Therefore, we should not ignore the importance of student conceptual understanding and attitude.

Additional analysis was performed using both ROT and ASCIv2. Results from multiple regression analysis indicate that the unique contribution of the ROT score to the ACS general chemistry exam score is not significant when other variables, such as attitude and math ability, are controlled.
I. Introduction

The Need for Transforming STEM Education

The importance of science, technology, engineering, and mathematics (STEM) education to have a qualified workforce has been well documented in the literature and educational reports. They include *Linking evidence and promising practices in science, technology, engineering, and mathematics (STEM)* (Fairweather, 2008), *STEM, STEM education, STEMmania* (Sanders, 2009), *Advancing STEM Education: A2020Vision* (Byhee, 2010), and *Engage to excel: Producing one million additional college graduates with degrees in science, technology, engineering, and mathematics* (Olson & Riordan, 2012). According to the New Bureau of Labor Statistics, employment in STEM occupations grew 7.9% during 2000-2010, compared to 2.6% for non-STEM. Further, this trend of faster growth in STEM is projected to continue for the next 10 years (Langdon, McKittrick, Beede, Khan, & Doms, 2011). However, the U.S. is still very likely to lose jobs in STEM areas to the global workforce because the U.S. is not preparing for a sufficient local workforce. For example, China has recently seen a sharp increase in their share of degrees in STEM, and awarded five times as many of those degrees than the U.S. did each year (Anderson & Kim, 2006). Therefore, there is an ongoing need to transform the US STEM education system to attract and retain a greater number of students in these fields (Olson & Riordan, 2012).

Currently great attention is being paid to improving STEM education in the U.S. in order to meet the global challenges of economic and societal development (National Academy of
Sciences, National Academy of Engineering, & the Institute of Medicine, 2010). Other countries, whether similar to the U.S. like Australia or less similar to the U.S., such as Saudi Arabia, also are emphasizing the importance of STEM education (Tytler, Osborne, Williams, Tytler, & Cripps-Clark, 2008). The Discipline-Based Education Research (DBER) integrates college educators across multiple disciplines in the development and implementation of evidence-based teaching curricula for fundamental STEM courses. The aim of DBER is to provide evidence of the effectiveness of instructional strategies, methods, pedagogies and assessments that apply to STEM (Fairweather, 2008; Singer, Nielsen, & Schweingruber, 2012).

Both cognitive and non-cognitive factors have been shown to be related to student academic performance and their long-term persistence in the study of STEM areas (House, 1995; Creemers, & Kyriakides, 2010). Spatial visualization and the attitude toward science are among the current focus areas in educational research and practice (Wai, Lubinski, & Benbow, 2009). However, many problems still remain unsolved, such as the lack of sufficient evidence on the reliability and validity of measurement tools, as well as the use of robust instruments to support research and practice (Arjoon, Xu, & Lewis, 2013; Blalock et al., 2008). This dissertation continues the effort to evaluate the suitability of two instruments for spatial ability (Voyer, Voyer, & Bryden, 1995) and attitude (American Association for the Advancement of Science, 1989), and to investigate the relationship of these two factors, as measured by the instruments noted above, to actual college student performance, respectively.

**Important Factors in STEM Education**

*Spatial Ability*

Spatial ability has been studied in the literature since the 1940s and recently achieved a new momentum in STEM education (Garrett, 1946; Michael, Guilford, Fruchter, & Zimmerman, 1989).
Spatial ability is defined as the capacity to “mentally manipulate, rotate, twist, or invert a pictorially presented stimulus object” (McGee, 1979). A sex gap in spatial ability is consistently reported as favoring males, especially in the domain of mental rotations (Voyer et al., 1995). While academic performance across sexes is closed at the school level, females are highly underrepresented at the professional level (Burke & Mattis, 2007; Xu, 2008). It is not yet clear whether this lack of spatial ability is an important barrier that prevents students, especially females who do tend to underperform in this regard, from entering STEM fields, and then failing to fulfill their potential even when they have received equally sufficient academic preparation (Hill, Corbett, & St. Rose, 2010; Xu, 2008).

Currently, there are ongoing efforts to develop the spatial skills of college students to improve their academic performance in STEM areas. Miller and Halpern (2013) found in a one-year longitudinal study of college students that 12 hours of training in spatial ability could improve student introductory physics scores. Sorby and Baartmans (1996) developed a ten-week course to train spatial thinking skills to students early on in the college engineering curriculum. More studies in this area will help understand, evaluate, improve, and implement effective training programs and help students excel at visually-intensive subjects, such as engineering and organic chemistry.

**Attitude toward Science**

Attitude can be defined as “a learned, positive, or negative feeling about science that serves as a convenient summary of a wide variety of beliefs about science” (Koballa & Crawley, 1985). Attitude, together with knowledge and skills, should be fostered in science education in the U.S. (AAAS, 1989; Kolsto, 2001). The literature indicates that attitude can affect student academic performance, and affect students’ long-term persistence in science-related endeavors (Koballa &
Crawley, 1985; M. Walczak & D. Walczak 2009; Richardson, Abraham et al., 2012). However, because student attitude measures are mostly self-reported and also subjective, it is hard for attitude to be emphasized in an equal way to science content knowledge. With the competition pressure to attain high scores on high-stakes content knowledge assessments, student attitude can be affected. The research has indicated that student interest in science declines from middle school and high school (George, 2003; R. Yager & S. Yager, 1985). Negative attitude can harm student performance in the long run as well as student willingness to pursue science-related careers (Tai, Liu, Maltese, & Fan, 2006). Unfortunately, society at large is too often left unaware of the critical nature of this particular issue and its effects (Wooden, 2006). Thus, more attention should be given to the level students’ attitudes toward science, so they will more likely develop a lifelong interest in science and eventually science literacy.

Similar to the science arena, student attitude toward chemistry plays an important role in a person’s persistence in seeking a chemistry-related career. The literature indicates that attitude positively relates to a student’s chemistry performance (Brandriet, Xu, Bretz, & Lewis, 2012; Widanski & McCarthy, 2009). Chemistry educators have developed numerous activities to address this issue and focus on student attitudes. Rather than attempting to rely on indoctrination, educators have developed active learning pedagogies, such as small- group peer learning (Amaral & Vala, 2009; S. Lewis & J. Lewis, 2005; Wamser, 2006), inquiry-based lab training, science writing heuristics (Greenbowe & Hand, 2005), clickers (Smith et al., 2009), problem-based learning (Horton et al., 1993), concept mapping (Horton et al., 1993; Pendley, Bretz, & Novak, 1994), and computer simulations (Ozmen, 2008; Winberg & Berg, 2007) to place an emphasis on both meaningful learning and positive student attitudes.
The Role of Measurement Tools in STEM Education

Measurement tools in education (e.g., quizzes, exams, surveys, and questionnaires) play a critical role, as they provide feedback that informs instruction. Just as a steam engine can only be improved through the measures of power and coal consumption, the importance of assessment tools cannot be overemphasized for the process of transforming STEM education (Gates, 2013). Other than the obvious potential benefits, these measurement tools, especially high-stake tests, can have a negative effect on instruction if not used appropriately, such as increasing test stress, distraction from effective instructional time, and excessively narrowing the curriculum (Mehrens, 1998). Due to the complexity of today’s educational settings, abstract variables, such as spatial ability, attitude, or performance, are often loosely defined and hard to quantify. Special concern should thus be taken to avoid negative effects from intended instrument use, as exemplified by the Standards for Educational and Psychological Testing (the Standards hereinafter) (AERA, APA, & NCME, 1999, 2014).

In the chemistry education context, a framework based on the Standards has been used to examine the instruments published in the Journal of Chemical Education, one of the key journals in the field, with respect to the presence of evidence. The results indicate that psychometric evidence is not being reported at the level of detail most appropriate to inform future research and satisfactorily meet the Standards (Arjoon, Xu, & Lewis, 2013). There is thus a need for more instruments and accumulated evidence that aligns well with the criteria found in the Standards.

Purposes of this Dissertation

This dissertation has three purposes. The first is to provide psychometric evidence on the appropriateness of using two instruments, one being the Purdue Visualization of Rotations Test.
(ROT) that measures the spatial ability of college students (Bodner & Guay, 1997), and the other, a shortened version of the Attitude toward the Subject of Chemistry Inventory (ASCIv2) on students’ attitude toward chemistry (Bauer, 2008; Xu & Lewis, 2011). The second purpose is to examine whether a lack of spatial ability, as measured by the ROT, can hinder students’ progression from general chemistry through organic chemistry. The third purpose is to examine the unique proportion of variance in chemistry achievement that can be attributable to other factors, especially student attitude, when other variables, such as math ability, are controlled.

College chemistry courses are examined in this dissertation because chemistry is a pivotal subject that serves many students who are interested in STEM areas. Most universities require students to pass general chemistry to be able to enroll in more advanced courses, such as organic chemistry, and graduate with a STEM degree. Therefore, student performance in college chemistry courses is a concern for educators (Lewis & Lewis, 2008; Zusho, Pintrich, & Coppola, 2003). Helping students to improve their chemistry performance will give them a better chance to persist in STEM education.

This dissertation also has implications for instruction by providing solid knowledge of the links of certain factors to student performance. If the unique contribution of spatial ability or attitude is confirmed, then instructors can better help students in that regard by taking advantage of existing tools or creating innovative new interventions. In addition, this research seeks information on appropriate measurement tools that can be used to assess student profiles and produce instruction that is tailored to the student level. Further still, this work will have value for informing other discipline-based educational research areas that utilize well-established instruments to support research findings and curriculum evaluation. With the integration of
measurement, evaluation and instruction, this work can collectively produce valued progress toward the most effective educational intervention and a better education system overall.

This dissertation includes five studies. The first study, *Sex difference in spatial ability of college students and the exploration of measurement invariance*, gathers evidence, using the *Standards*, on the validity and reliability of data gathered using the ROT instrument, so teachers can use this instrument in large chemistry classrooms. One focus of this study is to examine whether the ROT functions in the same way across sexes, so scores will offer a fair comparison between males and females. Currently, there are multiple instruments, including the ROT, that are available to measure the spatial ability of college students (Harle & Towns, 2011).

Regardless of the persistent sex gap, however, no measurement of spatial ability has been carefully enough investigated to determine whether the test scores reflect a true sex difference or simply are an instrument bias against females. Because of the ongoing emphasis on spatial ability in STEM education (Black, 2005; Liben, Kastens, & Christensen, 2011; Lord, 1985; Pribyl & Bodner, 1987; Wai et al., 2009), there is a need to gather further psychometric evidence on these measurement tools to support the research and also the actual practice in STEM areas. Findings from this study can provide valuable evidence for finding an appropriate assessment tool that that produces less bias against females.

Study 2, *the role of spatial ability in students’ progression through organic chemistry*, examines whether the lack of spatial ability represents a barrier for students’ progression from general chemistry through the organic chemistry learning sequence. Students who do poorly on ROT in general chemistry may be wonder whether they should make a decision to stay away from organic chemistry, which is assumed to be highly demanding in spatial ability. No longitudinal study has yet been reported to track students from general chemistry through
organic chemistry. Therefore, it is important to fill this gap to guide student decisions. In chemistry, because students can use alternate strategies to solve structure-related problems, student performance may not necessarily be associated with spatial ability (Devetak & Glažar, 2009). Accordingly, whether students actually require a high spatial ability to succeed in later chemistry courses needs further investigation, especially for females, who tend to do worse on ROT. Findings from this study will shed light on any concerns for whether the lack of spatial ability is a hurdle to overcome to progress successfully through organic chemistry.

Study 3, *Attitude toward the subject of chemistry in Australia: An ALIUS and POGIL collaboration to promote cross-national comparisons*, gathers psychometric evidence for using ASCIv2 to measure the attitude toward chemistry internationally. When an instrument is designed for cross-national use, there is a validity concern because students from various backgrounds can interpret items differently according to the standards criteria. The purpose for this pilot study, therefore, is to investigate how ASCIv2 functions in a Process-Oriented Guided Inquiry Learning (POGIL) context at a university in Australia. If ASCIv2 functions reasonably well, it may serve as a tool to assess the impact of implementation of POGIL in the classroom. The importance of student attitude in Australia is clearly demonstrated in the reports by the Australian Department of Education, which indicates a decline in the number of STEM degrees conferred due to student negative attitude toward STEM (Tytler, Osborne, Williams, Tytler, & Cripps-Clark, 2008). POGIL was actually introduced as one of the student-centered strategies to foster positive student attitude toward the discipline (POGIL, 2012). This study then seeks to demonstrate the utility of ASCIv2 in an Australia POGIL context that is based on internationally accepted psychometric standards.
Study 4, *Gathering psychometric evidence for ASCIv2 to support cross-cultural attitudinal studies on college chemistry programs*, examines how student samples at a Saudi Arabian institution respond to the ASCIv2 items and their attitude status compared to those found in an Australian and a U.S. institution. The Saudi Arabian population is of particular interest in chemistry education because of its unique cultural, religious and educational features that stand apart from Western countries such as the U.S. POGIL has also been implemented in a university in Saudi Arabia, but there is only limited knowledge of how ASCIv2 functions in the POGIL context in Saudi Arabia. The findings of Study 4 will contribute to a better understanding of how ASCIv2 functions in a Saudi Arabian context and also provide a preliminary attitudinal profile of students from universities in three countries.

Study 5, *College student attitudes toward chemistry, conceptual knowledge, and achievement: A structural equation model analysis* uses math ability, prior conceptual knowledge in chemistry, and attitudes toward chemistry measures to predict student achievement in chemistry. The purpose of this study is to explore to what degree multiple variables are directly associated with student chemistry performance. In chemical education, many factors have been reported to influence student achievement in college chemistry courses. Mathematical ability is important to solve multiple chemistry problems, such as unit conversion and equilibrium, and thus that ability is a criterion used for student placement into specific levels of chemistry courses at the college level (Cooper & Pearson, 2012; Pienta, 2003; Scott, 2012). Prior conceptual knowledge is critical for students in order to build new conceptual understandings, and this knowledge has been found to affect subsequent student achievement (Seery, 2009; Wagner, Sasser, & DiBiase, 2002). Student attitude toward chemistry has also been reported to influence student achievement (Bauer, 2008; Cukrowska, Staskun, & Schoeman, 1999; House,
1995; Xu & Lewis, 2011). Although each of the factors discussed above is clearly related to achievement, no study has been done that examines the relationship between these variables simultaneously while also taking into consideration measurement error. This study fills that research gap by using Structural Equation Modeling (SEM). The findings of this study can reveal a full SEM model to explain the substantial percentage (69%) of variance in student chemistry achievement.

In this dissertation, Chapter Two introduces general methods and techniques, including validity evidence based on the Standards and the SEM technique used to examine multivariate relations. Chapters Three and Four are manuscripts on spatial ability studies. Chapters Five through Seven are published studies on students’ attitudes toward chemistry. Each chapter has been reproduced verbatim with kind permission to be included in this dissertation from both publishers and coauthors. References for each chapter are formatted for the publication/journal. Lastly, Chapter Eight provides an overall discussion of the entire dissertation and its implications and limitations and opportunities for further research.

References


II. Methods

This chapter presents general methods and critical data analysis approaches that were applied more than once during this dissertation.

Instruments

Two instruments were evaluated and applied for a large student sample size in a college chemistry context in this dissertation work. Both are paper and pencil instruments and each only takes about 10 minutes to administer. Students were asked to bubble in answers on scantron, and then data were read by scanning techniques. Therefore, these instruments are quick and easy for instructors to administer and grade. This instrument format fits well the college chemistry classrooms, which usually have large enrollment of around 200 students. In addition, using instruments can also provide useful student information to inform instruction.

One instrument is the Purdue Visualization of Rotations (ROT) test to measure a spatial visualization component, mental rotations (Pribyl & Bodner, 1987). Mental rotations skill is important in chemistry because both chemistry experts and students may use this strategy to solve problems relating to visualizing molecules, such as stereochemistry and space groups, etc (Stieff, 2011, 2013; Stieff & Raje, 2010). The ROT was chosen because it has been frequently used not only in chemistry but STEM education areas (Bodner & Guay, 1997; Brownlow, McPherson, & Acks, 2003; Carter, LaRussa, & Bodner, 1987; Morgil, Yavuz, Oskay, & Arda, 2005). In addition, the ROT is shown to be suitable to measure college level student’s spatial ability according to the standards (AERA, APA, & NCME, 1999; Arjoon, Xu, & Lewis, 2013).
Please find the student instruction and proctor instruction for administering the ROT in the Supplement. The test takers were required to finish all 20 items in 10 minutes, they may use the instruction sheet as scratch paper if needed but they were not allowed to make any mark on the booklet; this minimizes the possibility of using analytical reasoning strategies. However, studies on the possible test bias against females are very limited. In the first two studies of this dissertation, ROT scores were used to examine evidence of temporal reliability, validity based on internal structure, and relationship with other variables such as chemistry performance.

The other instrument is a shortened version of the Attitude toward the Subject of Chemistry Inventory, ASCIv2 (Bauer, 2008; Xu & Lewis, 2011). Student positive attitude in science area is important because it is a goal of science education, and remains as one of the important research areas (American Association for the Advancement of Science, 1989; Koballa, 2006). The ASCIv2 was chosen because it has reported the most reliability and validity evidence among attitudinal measures specific to chemistry and student attitude fits the research interest in this work (Bauer, 2008; Brandriet, Xu, Bretz, & Lewis, 2012; Xu, Southam, & Lewis, 2012). It is short, with only eight items that can fit into a half piece of paper. Therefore, it is very convenient for large classroom use. My master’s thesis concerns the refining of this instrument (Xu, 2010). The last three studies of this dissertation continue to effort to use the ASCIv2 to examine the relationship of attitude with chemistry achievement, and how students at multiple sites with different backgrounds responded to instrument items.

**Data Collection and Participants**

The administration of instruments can be summarized in Table 2.1 in this dissertation. Participants were intact classes enrolled in college chemistry courses. Instruments were given during class time together with other instruments for purposes other than this dissertation.
Students were told that their answers were very important to provide information to improve the chemistry program, but would not harm their grades, and their instructors would only know the summary results but not their individual responses. Because the information gathered was intended to improve the chemistry curriculum, the local IRB declared that work not research, with one exception. Data collected at the Australian university was collected as per Human Research Ethics Committees (HREC).

Table 2.1

Summary of instruments administration

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Time</th>
<th>ROT/ASCIv2 &amp; instruments given at the same time (# of items)</th>
<th>Institution, course</th>
</tr>
</thead>
<tbody>
<tr>
<td>3,4</td>
<td>Fall 2008</td>
<td><strong>ROT</strong> (#1-20), Study Process Questionnaire SPQ$^d$ (#21-52)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>3,4</td>
<td>Spring 2009</td>
<td><strong>ROT</strong> (#1-20), CHEMX$^b$ (#21-67)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>3,4</td>
<td>Fall 2009</td>
<td><strong>ROT</strong> (#1-20), the Colorado Learning Attitudes about Science Survey</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>3,4</td>
<td>Fall 2009</td>
<td>CLASS (#1-17),$^{cd}$ Particulate Nature of Matter and Chemical Bonding CCP$^c$ (#18-37)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>3,4</td>
<td>Spring 2010</td>
<td>CLASS (#1-17), CCP (#18-37), <strong>ROT$^d$</strong> (#1-20)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>5,6</td>
<td>Spring 2009</td>
<td><strong>ASCIv2</strong> (#1-8)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
<tr>
<td>5,6</td>
<td>Fall 2011</td>
<td><strong>ASCIv2</strong> (#1-8)</td>
<td>WA in Australia, I: Gen Chem II; II: Organic &amp; Biological Chemistry II</td>
</tr>
<tr>
<td>6</td>
<td>Spring 2011</td>
<td><strong>ASCIv2</strong> (#1-8)</td>
<td>KU in KSA, Gen Chem I</td>
</tr>
<tr>
<td>7</td>
<td>Fall 2010</td>
<td><strong>ASCIv2</strong> (#1-8), the Test of Science Related Attitudes TOSRA$^f$ (#1-30), TMI$^g$ (#31-48)</td>
<td>SE in USA, Gen Chem I</td>
</tr>
</tbody>
</table>

NOTE: Data collected from instruments other than ROT and ASCIv2 are used for other purposes. SE, WA and KU are pseudonyms for each university.
a SPQ was developed by Biggs (1987); b CHEMX was developed by Grove & Bretz; c CLASS was modified by Barbera and colleagues (2008); d Separate scantron was used to make sure students work on ROT only for 10 minutes; e CCP was developed by Othman, Treagust, Chandrasegaran, (2008); f TOSRA was developed by Fraser (1978) g TMI was adapted by Heredia. CCP and TMI can be found in Heredia’s dissertation. Administration instructions for ROT and ASCIv2 can be found in the Appendices A and B.

Data Analysis Strategies

The data analysis strategies can be classified into two broad types. One was a psychometric evaluation of each instrument, with a focus on evidence based on internal structure using factor analysis and evidence based on internal consistency using Cronbach’s alpha (Arjoon et al., 2013). The other was relational analysis between studied construct with other variables, such as chemistry performance, using statistical inferences, e.g., Pearson correlation, Analysis of Covariance (ANCOVA), and Structural Equation Modeling (SEM).

Psychometric Analysis

In order to provide criteria for evaluation of test practice and consequence of instruments for educational community, American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) have jointly published and revised the Standards for Educational and Psychological Testing (hereafter referred to as the Standards) (AERA et al., 1999, 2014). The Standards is written for the test professionals and educated people to address professional and technical issues of test development and use in education, psychology, and employment. A framework based on the Standards, as shown in Figure 2.1, has been applied in the chemistry context to examine the state of art regarding the presence of evidence for instruments published in the Journal of Chemical Education (Arjoon et al., 2013). Because it is a long and iterative process to develop and validate an instrument, this dissertation started with existing instruments, ROT and ASCIv2, and to examine whether they functioned well in the study settings and produced valid data for research purposes. Because the community has been using these two instruments in multiple
studies in which ROT or ASCIv2 scores are interpreted (e.g., Maeda & Yoon, 2013, Xu, Southam, & Lewis, 2012), this use constitutes some evidence of content validity and response processes. Therefore, this dissertation work was focused on evidence of internal consistency reliability, temporal stability, and internal structure for multiple college student samples, with a small amount of work on response process for ASCIv2.

**Figure 2.1** Evidence Sources for Psychometric Quality

**Internal consistency reliability.** This work reported Cronbach’s alpha for each subscale when a factor score would be interpreted. Typically internal consistency is more reported in literature than temporal stability because of the convenience of avoiding replicate administrations. Cronbach’s alpha is the most frequently used approach to estimate the level at which multiple test items hold together and consistently measure one construct (Cronbach, 1951). It is calculated using the equation below:

\[ \alpha = \left[ \frac{K}{K-1} \right] \times \left[ 1 - \left( \sum \delta_k^2 \right) / \delta_{\text{total}}^2 \right] \quad \text{Equation 3.1} \]

\( K \) is the number of items. Because Cronbach’s alpha tends to be underestimated from sample data for a small number of items and needs to be adjusted. \( \sum \delta_k^2 \) is the sum of variance in each \( K \) individual item score, \( \delta_{\text{total}}^2 \) is the variance in the total score. If item scores are perfectly
positively related, the $\sum \delta_k^2 / \delta_{\text{total}}^2$ will be 1. The closer the Cronbach’s alpha value is to +1, the greater the probability that the score is reliable. Alpha greater than 0.7 is considered to be acceptable reliability for research (Nunnally, 1978). If the alpha is close to 0, the items scores are random and uncorrelated. In addition, because Cronbach’s alpha for all items is required but not sufficient to support the unidimensionality of the whole instrument and the use of the total score (Gardner, 1995), there is a need to establish internal consistency evidence for each subscale if factor scores other than the total scores are interpreted.

**Internal structure.** Confirmatory factor analysis (CFA) is more appropriate than other multivariate approaches, such as exploratory factor analysis, to support internal structure when the instrument is designed based on a theoretical and conceptual model. The preference for a CFA approach is endorsed by measurement professionals as shown in statements, e.g., “the most rigorous test of factorial structure is through CFA techniques” (Greenbaum & Dedrick, 1998), and “CFA is a commonly accepted method to test/confirm dimensionality” (Netemeyer, Bearden, & Sharma, 2003). CFA has the advantage of providing fit indexes, e.g., chi-square, Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA) to examine the overall goodness of fit of a model to empirical data (Bentler, 1990; Hu & Bentler, 1999; Yu, 2002). Because ASCIv2 and ROT items were designed to measure intended underlying constructs of attitudes and spatial ability respectively, CFA was used to support the internal structure and score interpretation.

In order to provide evidence whether the test items function in the same way to measure the same construct across all groups, measurement invariance testing was performed, so that we can use the factor score from the instrument to compare groups. When measurement invariance does not hold for a variable, the measurement could be biased against one or more groups.
Mathematically, the relationship of observed variable $y_g$ and latent variable $f_g$ for group $g$ is represented in the equation below:

$$y_g = \alpha_g + \Lambda_g f_g + e_g$$

Where $\alpha_g$ is the intercept for group $g$, $\Lambda_g$ represents the factor loading, and $e_g$ is the error term. To be able to make a fair comparison on the latent variable $f$ using observed variable $y$ across groups, at least intercept $\alpha_g$ and factor loading $\Lambda_g$ should be the same across groups (Raykov, Marcoulides, & Li, 2012).

**Relations to Other Variables by Structural Equation Modeling**

Structural Equation modeling (SEM) is defined “as a class of methodologies that seeks to represent hypotheses about summary statistics derived from empirical measurements in terms of a smaller number of ’structural’ parameters defined by a hypothesized underlying model” (Kaplan, 2009). Simply put, all univariate and multivariate parametric analyses can be considered as a special case of SEM, e.g., correlation, regression, ANOVA, and CFA (Fan, 1997). SEM analysis typically needs a large sample size, but has the advantage of evaluating overall model fitness to the data, variance explained, effect of each predictor on outcome, and relationship among all these variables simultaneously (Kline, 2011).

Before performing SEM analysis, a priori theoretic models need to be specified. SEM analyzes the covariance matrix of observed variables, fits the model to get the estimated parameter estimates, and compares the model-implied covariance matrix reproduced from parameters with the actual matrix. When the fit function follows the $\chi^2 (df)$ distribution, the criteria of $p > .05$ was used not to reject the null hypothesis, which indicates model fits the data well. The degree of freedom, $df$, is the number of information pieces known from the variance-covariance matrix minus the free parameter estimated in the model. In general, the $df$ has a
fixed value for the same model specification as shown in Table 3.7 in Chapter 3. However, $df$ can be different for different data set with the same model specification. Note that $df$ is calculated in a different approach for mean and variance adjusted weighted least squares (WLSMV) estimator to accommodate categorical indicators. and $df$ is adjusted based on each data set to give a trustworthy chi-square-based $p$ value. A full SEM model can have many variables (whether latent or observed) and estimate many parameters. For this situation, SEM requires large sample sizes, to have more power to find the true relationship (Kline, 2011). SEM was used in this work only when the sample size was appropriate for the analysis.

In this dissertation, multiple SEM models were performed to take into account measurement error and a large number of relationships simultaneously. A full model as shown Figure 2.1 and also in the second study, *The role of spatial ability in students’ progression through organic chemistry*, incorporates both a measurement part and a regression part. By using SEM, one can better evaluate overall model fit data, variance explained, effect of each predictor on outcome, and relationship among all these variables and rule out the measurement error.

![A Full SEM Model](image)

**Figure 2.2** An Example of a Full SEM Model
The ovals represent latent variables, the rectangles indicate observed variables, and the small circles are residual variances. Single headed arrow designates a causal relationship (from cause to effect). Double headed arrow defines a relationship without causality interpretation. The numbers on the line (path or correlation coefficients) can be used to test the statistical significance of null hypothesis. ROT items were set to load on their assigned factors only, and predictors (math ability, sex and mental rotations ability) are set to correlate with one another. In total, there are seven latent variables and 27 observed variables in this Figure.

References


III. Sex Difference in Spatial Ability for College Students and

Exploration of Measurement Invariance

Note to Reader

This chapter presents a manuscript submitted to the journal Learning and Individual Differences coauthored with Drs. Kim and Lewis, reproduced by permission of the Elsevier and coauthors.

Introduction

Spatial ability has been proposed as important for higher education including astronomy, geology, chemistry and engineering (Black, 2005; Liben, Kastens, & Christensen, 2011; T. R. Lord, 1985; Pribyl & Bodner, 1987; Wai, Lubinski, & Benbow, 2009). For example, geologists need to think about the configuration and transformation of rock layers, and to read a plethora of graphs and maps. Students without sufficient spatial skills need aids or extra effort to understand space-related tasks. Currently, there are ongoing efforts to train college students’ spatial skills to improve performance in the Science, Technology, Engineering, and Mathematics (STEM) areas at the college level. Miller and Halpern (2013) found the training of spatial ability improved student introductory physics scores in a one-year study for college students. Sorby and Baartmans (1996) developed a ten-week course to train spatial thinking skills in the early college engineering curriculum. In the development of spatial ability training programs, researchers need
robust measurement tools that align with instructional goals in order to better understand the effectiveness of those programs.

Spatial ability has been defined as the ability to “mentally manipulate, rotate, twist, or invert a pictorially presented stimulus object” (McGee, 1979). In studying spatial ability, researchers use terms that could mean either overall spatial ability or one aspect of spatial ability, such as spatial visualization (Salthouse, Babcock, Skovronek, Mitchell, & Palmon, 1990; Sorby, 1999), mental rotation (Shepard & Metzler, 1971; Stieff, 2007), or spatial orientation (Merchant et al., 2012; Tartre, 1990). Therefore, there is a need to distinguish between specific spatial ability skills, so that researchers can describe more clearly what is being studied. Lohman (1979) proposed a framework with major and minor components such as spatial relations, spatial orientation, visualizations and transformations (Harle & Towns, 2011; Lohman, 1979). In their review of the literature, Linn and Petersen (1985) instead suggested three categories: spatial perception, mental rotation and spatial visualization. The spatial relations component in Lohman’s framework and the mental rotation category proposed by Linn and Petersen both concern tasks requiring mentally rotating an object in space. When such a task involves the object as a whole without any change in shape, there is really no distinction between the two (Sorby, 1999). The present study can therefore be said to relate to spatial relations or mental rotations. We will use the latter term.

Because the mental rotation component of spatial ability is one of the cognitive areas with the most persistent sex gap favoring males (Linn & Petersen, 1985; Voyer, Voyer, & Bryden, 1995; Weckbacher & Okamoto, 2014), it has been posited as a potential reason for women’s underperformance and underrepresentation in STEM areas (Ceci, Williams, & Barnett, 2009; Hill, Corbett, & St. Rose, 2010). Recently, meta-analytic techniques have revealed
significant national variability in performance differences between female and male middle school students in math (Else-Quest, Hyde, & Linn, 2010) and science (Nosek et al., 2009), implicating societal factors rather than innate ability. Females are still underperforming in college level science in the United States, even with interventions intended to address cultural factors (Miyake et al., 2010) and are underrepresented in STEM careers worldwide (Blickenstaff, 2005), so there remains a need for researchers to understand whether the observed differences in spatial ability between the sexes are significant with respect to this problem. Unfortunately, studies on possible test bias with respect to sex have been rarely reported for spatial ability instruments, with some notable exceptions (e.g., Arendasy & Sommer, 2010; Bors & Vigneau, 2011), although some interesting recent work with respect to mental processing speed has ruled out this factor as a contributor to the persistent sex difference in performance on a timed mental rotation task (Debelak, Gittler, & Arendasy, 2014). Investigating possible test bias is critical for determining whether an observed sex difference stems from test bias (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999). The Purdue Visualization of Rotations Test (ROT) is chosen in this study because it is frequently used in STEM areas (Bodner & Guay, 1997; Brownlow, McPherson, & Acks, 2003; Carter, LaRussa, & Bodner, 1987; Morgil, Yavuz, Oskay, & Arda, 2005), there is no existing study of possible test bias for this instrument, and the instrument provides a way to measure mental rotation ability under normal classroom conditions. While computerized mental rotation tests have been developed (Arendasy & Sommer, 2010), the need for paper and pencil tests, such as ROT, that can be administered under normal classroom conditions remains.
ROT is chosen also because it has accumulated some psychometric evidence. Educators need measurement tools that are both practical and robust in order to facilitate better understanding and training of spatial ability in undergraduate courses. According to the *Standards for Educational and Psychological Testing*, one source of validity based on response processes is whether test takers apply the intended strategy or use unintended approaches to solve the items (AERA et al., 1999). The administration conditions for ROT are designed to minimize the use of alternate analytical strategies. The test takers are required to finish all 20 items in 10 minutes, and they are not allowed to make any mark on the booklet; this reduces the possibility of using analytical reasoning strategies. A meta-analysis of studies using ROT and similar instruments has revealed these tests usually favor males and has suggested that the time limit plays a role (Maeda & Yoon, 2013). Confirmatory factor analysis has been used to support a one-factor structure (Yoon, 2011), but there is no report on alternate possible models or measurement invariance. When a test is utilized to compare groups, it is important to gather evidence based on measurement invariance, i.e., the test should measure the latent variable in the same way across groups (AERA et al., 1999; Arjoon et al., 2013; Kim & Yoon, 2011). Without this information, it is premature to draw a conclusion regarding whether ROT’s tendency to favour males is related to a true performance difference or simply to test bias. Additionally, internal consistency reliability has been reported for the ROT at multiple sites (Battista, Wheatley, & Talsma, 1982; Sorby & Baartmans, 1996; Yoon, 2011), but little is known about the temporal stability of test scores.

One goal of this study is to gather evidence according to the *Standards* for validity and reliability of ROT scores so that instructors can safely use the measure in college classrooms (AERA et al., 1999; Arjoon et al., 2013). The other is to establish measurement invariance for
the group mean comparison across sexes for a student sample enrolled in a college chemistry
course. Specifically, three research questions guide this study:

(1) How stable is the ROT score when students retake the test?
(2) How similarly do ROT items function between the sexes?
(3) What is the sex difference, if any, as measured by the ROT test?

Research Methods

Instruments

In this study, ROT (Bodner & Guay, 1997) was used to measure the mental rotations component
of spatial ability. It is a paper-and-pencil test with 20 items, which asks students to discern how
the block in the first row is rotated, and choose from the five drawings how the object in the
second row looks when the given rotation is applied. The ROT is derived from the Purdue
Spatial Visualization Tests: Visualization of Rotations, which was based on the Shepard-Metzler
test as adapted by Vandenberg (Shepard & Metzler, 1971; Vandenberg & Kuse, 1978). An
example of a ROT test item is illustrated in Figure 3.1. The correct answer is A.

Figure 3.1 ROT Item 9
Participants and Data Collection

ROT was given to students enrolled in a college chemistry course for science majors during the second week of four separate semesters (Fall 2008, Spring 2009, Fall 2009, and Spring 2010) at a large Southeastern public research university in the United States. The administrators followed the test procedure described in the literature (Bodner & Guay, 1997). There are 4168 ROT records in total during the four semesters. The ROT data were examined for each semester separately to establish psychometric evidence concerning whether the test functioned similarly across time and sex. The data were also combined only to track students who retook the course and the ROT. Demographic information such as sex, major and SAT math scores was obtained via the registrar. The top five majors in which students ultimately graduated (n = 986) were biomedical sciences (36%), biology (19%), health sciences (11%), psychology (5%) and public health (4%).

Table 3.1

Demographic Information for the Study Sample: SAT Math by Sex

<table>
<thead>
<tr>
<th>Semester</th>
<th>Mean (SD)</th>
<th>Cohen’s d for sex difference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Female</td>
<td>Male</td>
</tr>
<tr>
<td>Fall 2008 (F=731, M=665)</td>
<td>570(75)</td>
<td>603(71)</td>
</tr>
<tr>
<td>Spring 2009 (F=409 M=281)</td>
<td>537(66)</td>
<td>573(66)</td>
</tr>
<tr>
<td>Fall 2009 (F=746, M=658)</td>
<td>568(78)</td>
<td>607(76)</td>
</tr>
<tr>
<td>Spring 2010 (F=411, M=267)</td>
<td>528(67)</td>
<td>561(68)</td>
</tr>
</tbody>
</table>

In order to give a general idea of how the demographic distribution varies across semesters, Table 3.1 presents the number and SAT Math scores for both females and males of each semester. Sexes distribute similarly across semesters, with more females than males. Males tended to score slightly higher than females on SAT Math, with effect sizes quantified as medium according to Cohen’s d effect size guidelines (Cohen, 1988). Because this difference in
math ability may confound the spatial ability comparison, the SAT math score will be included in the analysis.

Statistical Data Analysis

Descriptive statistics. Descriptive analysis was performed including mean, standard deviation, skewness and kurtosis for the ROT total score. A reliability estimate was calculated by Cronbach’s alpha to reflect how items function consistently to assess the intended construct. Cronbach’s alpha of greater than 0.7 is considered as above the satisfactory level for research purposes (Murphy & Davidshofer, 2005).

Test-retest relationship for temporal stability. The 309 students who took the ROT more than once were used to track score change and correlation. Note that a typical way to measure test-retest reliability is to give a test to a representative sample within an appropriate time interval such as two weeks (Meyer, 2010). The test-retest relationship here is a convenient but not an ideal way to investigate temporal stability of ROT instrument scores. Because the student sample for this analysis is those who retook the course, they are not representative of the ordinary population enrolled in a general chemistry course. In addition, the test interval varied depending on how long they waited to retake the course in this natural setting, rather than within a fixed interval. Due to the nature of convenient data, the actual temporal stability of ROT can be underestimated due to other factors under this situation. For example, the assumption of the same student ability at both administrations may be violated because students may have been motivated to improve their spatial ability during the long interval. Having this knowledge in mind, we are cautious about the appropriate interpretation of this temporal stability data for ROT. Regardless of these limitations, the results can give a rough sense of the magnitude of change when students take the ROT test at different times under the current circumstances.
**Confirmatory factor analysis.** Since the 20 ROT items are designed to load on a general scale, a confirmatory factor analysis was performed on a 1-factor model (items 1 through 20 are loaded on one factor) in Mplus 5.2. All measured variables were categorical and a Means and Variance Adjusted Weighted Least Squares (WLSMV) method was employed to estimate parameters and goodness of model fit using the tetrachoric correlation matrix of the 20 items. The model was identified by fixing the first item on the factor at 1. In general, models based on a large number of scores are likely to have an inflated chi-square value and a significant lack of model fit, so additional fit statistics were examined. The criteria of Comparative Fit Index (CFI) value greater than .95 and Root Mean Square Error of Approximation (RMSEA) value less than .05, and Weighted Root Mean Square Residual (WRMR) less than 1 were used to indicate a good model fit (Bentler, 1990; Hu & Bentler, 1999; Yu, 2002). Chi-square per degrees of freedom ($\chi^2/\text{df}$) closer to 1 indicates good model fit, and less than 2 is recommended as a “rough rule of thumb” (Tabachnick & Fidell, 2007). Alternate models were proposed based on 1-factor model results and test item examination. Once a final model was chosen, it was then used for the follow-up measurement invariance analysis across sexes.

**Measurement invariance between the sexes.** One source of bias emerges when items do not load on a factor in a similar way across groups. Multiple group CFA was performed between the sexes to examine whether the factor loading pattern of ROT holds across sexes (Millsap & Everson, 1993). The measurement invariance can be established in four forms: configural, weak, strong, and strict invariance. Configural or equal form invariance is a basic form of measurement invariance, which requires only that the same item must indicate the same latent factor in each group, while the factor loadings can differ across groups. Weak or metric invariance is established when the fit of the model constraining the factor loadings to be equal
across groups does not fit substantially worse than the configural model. Strong or scalar invariance requires reasonable good model fit with the additional constraints of equal intercepts across groups. This requirement means that the same slope and intercept scheme can be applied across groups to calculate factor scores. Strict invariance adds the constraint of equal residual variance across groups. In order to be in a position to study similarities and differences on the same constructs in all subpopulations, at least strong invariance (factor loading and mean intercept invariance) is necessary, i.e., the same number of factors is tapped into by the same indicators across all groups (Raykov, Marcoulides, & Li, 2012).

However, this traditional procedure sequentially testing measurement invariance from configural to strict invariance has limitations. For example, it cannot test the invariance for anchor items (e.g., the first item) because the factor loading of the first item for each factor is fixed at 1 across groups for model identification purposes. Raykov et al. (2012) proposed a more straightforward approach to test strong invariance of all factor loadings and intercepts including those of the first item. This approach will be used for this study as follows. First, the best-fit model is run for females and males separately within each term. Once the best-fit model holds for both females and males with good fit, then the model is run for strong invariance, which constrains all intercepts and factor loadings to be equal across groups, but not at 1. Note that in this study the best-fit model of ROT was established, as described in the previous section, before measurement invariance testing.

**Effect size and ANCOVA to examine sex difference in spatial ability.** Sex difference in spatial ability was quantified from multiple-group CFA with strong invariance constraints, and results were evaluated with the traditional Cohen’s *d* approach using the total test scores (Cohen, 1988). Cohen’s *d* reveals how many standard deviation units two group means are apart. The guidelines
(d > 0.2, small; d > 0.5, medium, d > 0.8, large) are often used to interpret the meaning of this difference. Because the SAT Math is different across sexes for this sample consistently higher for males, it is important to examine whether the observed difference on ROT reflects the math ability gap or a real spatial ability gap. Thus, ANCOVA was performed in SAS 9.3 using SAT Math score as a covariate to rule out its influence (Heredia, Xu, & Lewis, 2012; Stevens, 2007). We further examined the sex difference in mental rotation ability in terms of item types (the degree of rotation and the number of rotation axes) to explore a possible source of the performance gap.

**Results and Discussion**

**Descriptive Statistics**

Table 3.2 presents means, standard deviations and Cronbach’s alpha for the ROT scores for each semester. Although not presented in Table 3.2, the skewness and kurtosis are consistently less than 1, indicating normal distributions. The means of the ROT scores range from 11.9 (out of 20) in Spring 2010 to 13.1 in Fall 2008, with the largest mean difference across terms at 1.2 points, which is in the small effect size category (d = 0.3). Cronbach’s alpha is greater than 0.7, which is beyond the satisfactory level for each semester and consistent with literature (Battista et al., 1982; Sorby & Baartmans, 1996; Yoon, 2011). Please see the Appendix for item difficulties, item-total correlations, and index of discrimination for all items each term. These item statistics are generally in acceptable ranges. For example, the range of item difficulty is .25 to .90, with no more than three items at .85 or above and only one item at .3 or below in any given semester.

It is notable from Table 3.2 that males scored significantly higher than females on ROT each semester, with an absolute mean difference ranging from 2.2 (Fall 2009) to 3.0 (Fall 2008).
Overall, the effect size (Cohen’s $d$) of the difference varied from 0.58 to 0.80, which translates to a medium to large sex difference in every term from direct interpretation of the observed scores.

**Table 3.2**

*Cronbach’s Alpha and Descriptive Statistics for ROT by Semester*

<table>
<thead>
<tr>
<th>Semester</th>
<th>Overall Mean ($SD$)</th>
<th>Male</th>
<th>Female</th>
<th>Sex difference</th>
<th>Cronbach’s $\alpha$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2008 ($n = 1396$)</td>
<td>13.1 (4.0)</td>
<td>14.7(3.7)</td>
<td>11.7(3.8)</td>
<td>3.0</td>
<td>.80</td>
</tr>
<tr>
<td>Spring 2009 ($n = 690$)</td>
<td>12.3 (4.1)</td>
<td>13.7(4.1)</td>
<td>11.3(3.8)</td>
<td>2.4</td>
<td>.79</td>
</tr>
<tr>
<td>Fall 2009 ($n = 1404$)</td>
<td>12.7 (4.0)</td>
<td>13.9(3.7)</td>
<td>11.7(3.9)</td>
<td>2.2</td>
<td>.78</td>
</tr>
<tr>
<td>Spring 2010 ($n = 678$)</td>
<td>11.9 (4.1)</td>
<td>13.6(3.9)</td>
<td>10.7(3.9)</td>
<td>2.9</td>
<td>.78</td>
</tr>
</tbody>
</table>

Note: The total score is 20 points, one point for each item on the ROT test.

**Test-retest Relationship for Temporal Stability**

The scores for the 309 ROT retakers are shown in Table 3.3 in order to address the first research question. The mean score is 11.9 at Time 1 and 12.6 at Time 2, which yields 0.7 for a simple gain (time2 – time1) out of 20. A dependent means $t$-test indicates the gain score is statistically different from 0 ($t (df = 308) = 3.44, p = .001$), but the effect size is 0.2, which is small. In other words, on average, students’ spatial ability scores slightly improved during the test-retest interval. Overall, the results are in alignment with previous findings that spatial ability is a temporally stable trait (Contreras, Rubio, Pena, & Santacreu, 2010; Newcombe & Dubas, 1992; Plomin, Pedersen, Lichtenstein, & McClearn, 1994; Schaie, Willis, Jay, & Chipuer, 1989).

Because there was no direct intervention with spatial ability training during this time, and the difference between Time 1 and Time 2 in the ROT scores was deemed small, the test-retest correlation ($r = .63$) can be considered as an estimate for the temporal stability for ROT. This statistically significant correlation is moderately large, but not large enough to reach a satisfactory level of reliability (Litwin, 1995). The length of the test-retest interval is likely to have affected these results. In the study setting, the academic year begins in late August, with the second semester beginning in early January. The majority of the retakers (70.6%) took the
ROT test twice in the same academic year, first in August and again in January, for a test-retest interval of 5 months. A small proportion (9.1%) retook the test in the same calendar year, but not the same academic year, for an interval of 8 months. The rest retook the ROT after an interval of 12 months or longer as shown in Table 3.4. Given the length of even the shortest interval, change in the true ability may lead to underestimation of the test-retest reliability (F. M. Lord, Novick, & Birnbaum, 1968; Meyer, 2010). With this limitation, the result therefore establishes some extent of stability for ROT scores in this natural setting, while a study design for the specific purpose of establishing the appropriate test-retest interval will warrant a more accurate estimate of temporal stability for ROT (Litwin, 1995).

Table 3.3

Descriptive Statistics of ROT Scores for Retakers

<table>
<thead>
<tr>
<th>Administration</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Effect size (Cohen’s $d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ROT Time 1</td>
<td>11.9</td>
<td>4.3</td>
<td>-0.5</td>
<td>-0.2</td>
<td></td>
</tr>
<tr>
<td>ROT Time 2</td>
<td>12.6</td>
<td>4.1</td>
<td>-0.4</td>
<td>-0.4</td>
<td></td>
</tr>
<tr>
<td>Difference in ROT</td>
<td>0.7</td>
<td>3.6</td>
<td>0.5</td>
<td>2.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

Note: Cohen’s $d = 0.2$, $0.5$, and $0.8$ as small, medium and large effect sizes, Cohen, p. 40

Table 3.4

Time Interval for Retaking the ROT Test

<table>
<thead>
<tr>
<th>Test-retest Interval</th>
<th>Number of Students</th>
<th>%</th>
<th>Cumulative %</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 months</td>
<td>218</td>
<td>70.6</td>
<td>70.6</td>
</tr>
<tr>
<td>8 months</td>
<td>28</td>
<td>9.1</td>
<td>79.6</td>
</tr>
<tr>
<td>12 months or more</td>
<td>63</td>
<td>20.4</td>
<td>100</td>
</tr>
</tbody>
</table>

Confirmatory Factor Analysis

For the second research question, we need to examine the construct validity evidence to determine whether the ROT test performed the same for both sexes as intended. The first step is to determine a reasonable factor structure. From CFA, the fit estimates of a one-factor model are
similar for all four semesters (see Tables 3.5 and 3.6) and less than ideal. For example, fit results for Spring 2010 are $\chi^2(n = 678, df = 120, p < .001) = 266$, CFI = .92, RMSEA = .04, WRMR = 1.16. Since CFI is less than .95 and WRMR is larger than 1, we cannot conclude that the model fits the data very well. Residual analysis did not indicate serious local misfit, with the largest residual value at 0.244. Based on the modification indices, the addition of error correlation between items 1 and 2, 6 and 7, 11 and 12, or 13 and 14 could improve the model fit by decreasing the value of $\chi^2$ by more than 10.

Table 3.5

*Fit Statistics of Three Models for ROT in Spring 2010*

<table>
<thead>
<tr>
<th>Fit Statistic</th>
<th>1-Factor</th>
<th>2nd-order, 4-Factor</th>
<th>Bi-factor model</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>266</td>
<td>160</td>
<td>125</td>
</tr>
<tr>
<td>df</td>
<td>120</td>
<td>118</td>
<td>109</td>
</tr>
<tr>
<td>p</td>
<td>&lt; .001</td>
<td>.01</td>
<td>.14</td>
</tr>
<tr>
<td>$\chi^2/df$</td>
<td>2.22</td>
<td>1.36</td>
<td>1.15</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.04</td>
<td>.02</td>
<td>.02</td>
</tr>
<tr>
<td>CFI</td>
<td>.92</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td>WRMR</td>
<td>1.2</td>
<td>0.90</td>
<td>0.77</td>
</tr>
</tbody>
</table>

Note: $\chi^2$ = Chi-square, $df$ = degree of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error Of Approximation, WRMR = Weighted Root Mean Square Residual, $n=678$

Rather than modifying the model directly based on empirical results, we investigated items to identify possible meaningful sources of misfit. A specific pattern was observed behind every five items. For Items 1 through 5, the object is rotated 90 degrees around one axis. For example, rotation for Item 2 is 90 degrees anti-clockwise around the axis vertical to the paper plane. For Items 6 through 10, the object is rotated 180 degrees around one axis. Figure 4.2 illustrate this rotation for item 9. From Items 11 through 15, the object is rotated 90 degrees around one axis, and then rotated 90 degrees around another axis. From Items 16 through 20, the object is rotated 90 degrees around one axis, and then rotated 180 degrees around another axis.
The four item sets are in the order of increased difficulty, with the overall average score decreasing from 0.83 for the first item set to 0.64, 0.63, and 0.40, respectively, for the next three sets. Research has shown that difficulty is a linear function of the degrees to be rotated from 0° to 180° (Shepard & Metzler, 1971) or the number of surfaces carried through rotation (Cooper & Shepard, 1973). The first two item sets (items 1-10) rotate on one axis at 90° and 180° respectively, which is consistent with the previous results. The last two item sets (items 11-20) rotate on two different axes, which is deemed to be more difficult. Our results, with the average score for each item set decreasing as the difficulty increases, are reasonable. The score decrease, however, may have an alternate explanation, such as an item order effect. It could be time pressure that influences students to score lower on the last five items. Further research could be done to rule out that possibility, for example, giving alternate test forms with scrambled item sets.

Based on the observed structure, we decided to test a second-order four-factor model. For the first order, items 1 through 5 were loaded on factor 1, items 6 through 10 on factor 2, items 11 through 15 on factor 3, and items 16 through 20 on factor 4. All four first-order factors were loaded on the general second-order factor. This model was run using the same estimation...
approach described above. Results indicate that overall the model fits the data well, with all fit indices improved, as shown in Table 3.5 for Spring 2010 and in Table 3.6 for the other semesters.

A bi-factor model presents another way to conceptualize and model multidimensional data with an underlying construct and an item-set writing pattern (Gibbons et al., 2007). As in Figure 3.3C, the probability to answer each item correctly is influenced by not only the general construct of mental rotation ability, but also the item writing pattern. A bi-factor model has an advantage over a second-order model, in supporting the use of test scores to reflect a single construct even when the data are multidimensional (Chen, West, & Sousa, 2006; Reise, Moore, & Haviland, 2010). In this study, the 20 items load on a single factor, mental rotation ability, while each set of five items simultaneously loads on the respective rotation factor that represents the effect of the item-set writing pattern. Overall the bi-factor model fits the data well, with improvements over the 2nd-order model indicated by most fit indices, as shown in Table 3.5 for Spring 2010 data. Figure 3.3 presents the parameter estimates for Spring 2010 data with the three different models.

CFA results for all three models from the other semesters are in Table 3.6. As presented in Tables 3.5 and 3.6, results support that the bi-factor model fits the data better than the other two models for each semester. Across semesters the bi-factor models meet the criteria of good fit, such as \( \text{CFI} > .95, \text{WRMR} < 1.0, \) and \( \chi^2/df < 2.0 \) (Hu & Bentler, 1995, 1999; Tabachnick & Fidell, 2007; Yu, 2002), showing considerably improved model fit as compared to the 1-factor and the second-order 4-factor models. Therefore, a bi-factor model is theoretically and empirically more tenable for the ROT test and was retained in the following tests for measurement invariance across sexes.
A. One-factor CFA model

B. Second-order CFA model
Figure 3.3 Standardized parameter estimates for one-factor (3A), second-order (3B) and bi-factor (3C) alternate models \((n = 678)\). The ovals designate latent variables, and the rectangles indicate observed variables. Items were set to load on their assigned factors only. All factor loadings on the general construct of mental rotations ability are significantly different from 0 \((p < 0.001)\).

Table 3.6

*Fit Statistics of ROT for Three Models for Three Semesters*

<table>
<thead>
<tr>
<th>Model</th>
<th>1-Factor</th>
<th>2\textsuperscript{nd}, 4-Factor</th>
<th>Bi-factor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fall 2008</td>
<td>Spring 2009</td>
<td>Fall 2009</td>
</tr>
<tr>
<td>(\chi^2)</td>
<td>716</td>
<td>325</td>
<td>530</td>
</tr>
<tr>
<td>(df)</td>
<td>128</td>
<td>115</td>
<td>133</td>
</tr>
<tr>
<td>(p)</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>(\chi^2/df)</td>
<td>5.59</td>
<td>2.83</td>
<td>3.98</td>
</tr>
<tr>
<td>CFI</td>
<td>.85</td>
<td>.88</td>
<td>.90</td>
</tr>
<tr>
<td>RMSEA</td>
<td>.06</td>
<td>.05</td>
<td>.05</td>
</tr>
<tr>
<td>WRMR</td>
<td>1.8</td>
<td>1.3</td>
<td>1.6</td>
</tr>
</tbody>
</table>

*Measurement Invariance between the Sexes*

Strong invariance for the bi-factor structure was examined first for each term and then by sex. Table 3.7 summarizes the fit index results. All CFI values are greater than .95, and RMSEA is consistently less than .05, indicating that strong invariance is tenable. Therefore, ROT test items
function similarly to measure the construct of mental rotations for females and males for each semester. We did not find evidence of test bias from this measurement invariance analysis.

**Table 3.7**

*Measurement Invariance Result across Sexes for CFA Bi-factor Model*

<table>
<thead>
<tr>
<th>Model</th>
<th>n</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
<th>$\chi^2$/df</th>
<th>CFI</th>
<th>RMSEA</th>
<th>WRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFA for F08</td>
<td>1396</td>
<td>207.8</td>
<td>108</td>
<td>.001</td>
<td>1.92</td>
<td>.98</td>
<td>.024</td>
<td>0.95</td>
</tr>
<tr>
<td>CFA for S09</td>
<td>690</td>
<td>153.9</td>
<td>104</td>
<td>.001</td>
<td>1.48</td>
<td>.97</td>
<td>.026</td>
<td>0.88</td>
</tr>
<tr>
<td>CFA for F09</td>
<td>1404</td>
<td>195.2</td>
<td>120</td>
<td>&lt;.001</td>
<td>1.63</td>
<td>.98</td>
<td>.021</td>
<td>0.93</td>
</tr>
<tr>
<td>CFA for S10</td>
<td>678</td>
<td>125</td>
<td>109</td>
<td>.14</td>
<td>1.15</td>
<td>.99</td>
<td>.015</td>
<td>0.77</td>
</tr>
<tr>
<td>CFA for females</td>
<td>2297</td>
<td>215.6</td>
<td>129</td>
<td>&lt;.001</td>
<td>1.67</td>
<td>.98</td>
<td>.017</td>
<td>0.95</td>
</tr>
<tr>
<td>CFA for males</td>
<td>1871</td>
<td>260.4</td>
<td>120</td>
<td>&lt;.001</td>
<td>2.17</td>
<td>.97</td>
<td>.025</td>
<td>1.05</td>
</tr>
<tr>
<td>Strong invariance across term</td>
<td>4168</td>
<td>755.5</td>
<td>457</td>
<td>&lt;.001</td>
<td>1.65</td>
<td>.97</td>
<td>.025</td>
<td>2.16</td>
</tr>
<tr>
<td>Strong invariance across sex F08</td>
<td>1396</td>
<td>292.8</td>
<td>196</td>
<td>&lt;.001</td>
<td>1.49</td>
<td>.96</td>
<td>.027</td>
<td>1.38</td>
</tr>
<tr>
<td>Strong invariance across sex S09</td>
<td>690</td>
<td>211.9</td>
<td>167</td>
<td>.01</td>
<td>1.26</td>
<td>.97</td>
<td>.028</td>
<td>1.28</td>
</tr>
<tr>
<td>Strong invariance across sex F09</td>
<td>1404</td>
<td>298.2</td>
<td>219</td>
<td>&lt;.001</td>
<td>1.36</td>
<td>.98</td>
<td>.023</td>
<td>1.34</td>
</tr>
<tr>
<td>Strong invariance across sex S10</td>
<td>678</td>
<td>187</td>
<td>178</td>
<td>.31</td>
<td>1.05</td>
<td>.99</td>
<td>.012</td>
<td>1.14</td>
</tr>
</tbody>
</table>

Note: $\chi^2$ = Chi-square, df = degree of freedom, CFI = Comparative Fit Index, RMSEA = Root Mean Square Error Of Approximation, WRMR = Weighted Root Mean Square Residual

**Sex Difference in Spatial Ability**

Table 3.8 presents the effect size from multiple group CFA under the constraints of strong invariance (Hancock, 2001; Thompson & Green, 2006) as well as Cohen’s $d$ based on the raw score, i.e., assuming no measurement error. In terms of spatial ability, the mean of males was higher than that of females across semesters. We see larger effect size values for each semester when the measurement error is accounted for by the CFA.

**Table 3.8**

*Effect Size from MCFA Bi-factor Strong Invariance Model and Cohen’s d*

<table>
<thead>
<tr>
<th>Semester</th>
<th>From MCFA</th>
<th>From observed score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fall 2008</td>
<td>0.87</td>
<td>0.80</td>
</tr>
<tr>
<td>Spring 2009</td>
<td>0.66</td>
<td>0.61</td>
</tr>
<tr>
<td>Fall 2009</td>
<td>0.70</td>
<td>0.58</td>
</tr>
<tr>
<td>Spring 2010</td>
<td>0.81</td>
<td>0.74</td>
</tr>
</tbody>
</table>
Because males outperformed females on SAT Math as shown in Table 3.1, ANCOVA was run for Spring 2010 data to rule out the possibility that the ROT difference arises from the math difference. Five hundred forty three students had the complete data of ROT and SAT Math scores and were used for this analysis. First, a model with an interaction between sex and SAT Math indicated no evidence of statistically significant interaction effect \( F(1, n = 543) = 2.68, p = 0.10 \), which means the assumption of no interaction effect in ANCOVA is tenable. From the final model without the interaction term, the results suggest a statistically significant effect for sex \( F(1, n = 543) = 42.1, p < .001 \) even after controlling for SAT Math. The adjusted means on the ROT are 13.0 and 10.9 for male and female students, respectively, with the average SAT Math score. Other semesters exhibited similar results. Therefore, the sex gap for ROT scores persists in our sample when SAT Math is controlled.

Bors and Vigneau (2011) observed an overall sex difference in mental rotation ability favoring males for a similar instrument and also investigated role of item type, reporting no sex effect with respect to what was termed “angular disparity” in their study. Angular disparity refers to the degree of mental rotation required. For this study, the sex difference in the observed score was also examined for each set of ROT item types to see whether results would be similar to those in that prior work. For each item set, males consistently outperformed females. For item type 1 (i.e., item sets 1 and 2), the 3-D objects rotate on one axis, with the degree of angular disparity 90° (set 1) and 180° (set 2). The sex difference increased from 0.56 to 0.67 with the increase of angular disparity (see Table 3.9) within this item type for data from Spring 2010. For item type 2 (i.e., item sets 3 and 4), the 3D objects have an additional 90° rotation on one axis but otherwise parallel items 1 and 2 with respect to the change in angular disparity associated with the other axis. For Spring 2010 data, the sex difference decreased from 0.52 to 0.37 with
the increase of angular disparity (see Table 3.9) within this item type. Therefore, there is an interaction effect between sex and item type (absence or presence of an additional 90° rotation) on the performance gap decrease due to angular disparity. For Spring 2010 data, the interaction effect was $F(1,676)= 8.89, p = .003$. Data from other semesters also support the significance of this interaction effect. The effect of sex on the performance decrease due to angular disparity depends on the item type.

Table 3.9

Sex Difference by Degree of Angular Disparity for Two Item Types

<table>
<thead>
<tr>
<th>Item type</th>
<th>Angular disparity</th>
<th>Female Mean</th>
<th>Female SD</th>
<th>Male Mean</th>
<th>Male SD</th>
<th>Effect size for sex difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>One rotation</td>
<td>Set 1 (90°)</td>
<td>3.61</td>
<td>1.35</td>
<td>4.32</td>
<td>1.13</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Set 2 (180°)</td>
<td>2.62</td>
<td>1.35</td>
<td>3.51</td>
<td>1.30</td>
<td>0.67</td>
</tr>
<tr>
<td>with additional 90°</td>
<td>Set 3 (90°)</td>
<td>2.59</td>
<td>1.58</td>
<td>3.39</td>
<td>1.46</td>
<td>0.52</td>
</tr>
<tr>
<td>rotation</td>
<td>Set 4 (180°)</td>
<td>1.90</td>
<td>1.32</td>
<td>2.39</td>
<td>1.34</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Note: Data are from Spring 2010. Sample size is 411 for females and 267 for males.

Conclusion

This study has successfully generated answers for the research questions. For the first two research questions regarding psychometric evaluation, results did not indicate concerns for test bias. Results support the reliability of ROT scores regarding internal consistency and temporal stability, although temporal stability is worthy of further study. When students retook the ROT in a later semester, the change in the ROT score is 0.7 out of 20, which is a small effect size.

Despite the fact that the test-retest interval was lengthy, the test-retest scores are correlated at .63. It would be of value to conduct a study of temporal stability with varying test-retest intervals to determine the appropriate interval over which to expect temporal stability. Temporal stability is highly relevant for course-based interventions, for which it is desirable to have an estimate of expected score variability across a relatively lengthy pre/post time interval such as a semester.
Furthermore, this study is the first to propose a bi-factor structure for the 20 ROT items, which exhibits a better fit than the previously tested one-factor model (Yoon, 2011). In the bi-factor model, a single construct of mental rotation ability and four item-writing factors (Rotation 1 – Rotation 4 in Figure 3.3C) were constructed to represent different rotation patterns corresponding to four item sets.

As a side product, this study has also found that the four item sets of the ROT test are presented in the order of increasing difficulty. Although the association between item difficulty and rotation pattern is in agreement with previous findings (Towns, 2010), this study does not rule out the possibility of increased difficulty simply due to item order in the current ROT test. Re-ordering the item sets within ROT would enable the exploration of concerns about whether difficulty is associated with time pressure or with the required rotation patterns. For example, switching item sets 1 and 2 may provide evidence for empirical fingerprinting (Harle & Towns, 2011). From the results of testing different versions of ROT, test writers will have a better idea about how the pattern of rotation can affect item difficulty.

Measurement invariance testing in this study showed that the bi-factor model functions similarly between the sexes to support a fair group mean comparison. For research question three, a sex gap with a medium to large effect size was observed in mental rotation ability as measured by ROT, favoring male students. This observed gap is in alignment with previous findings that spatial ability gaps persist, especially in mental rotation skills (Kinsey, Towle, O'Brien, & Bauer, 2008). Furthermore, this study contributes to previous literature by ruling out a potential confounding factor, math ability. This study also investigated the role of item type with respect to the performance gap. There was an interaction effect between sex and item type on the performance decrease associated with increased angular disparity (i.e., larger drop for females.
for item type 1; larger drop for males for item type 2). It may therefore be the case that Bors and Vigneau (2011) did not observe an association between the sex differences and the degree of angular disparity in their study with a similar instrument because of this potential interaction with item type.

This study has limited generalizability due to the use of a convenient sample at one site. It would be of value to replicate this study in a setting with a different student population. Furthermore, this study only gathered quantitative evidence based on internal structure and reliability. Other qualitative evidence, such as the actual strategies test takers utilize to solve the test items, and content validity from experts in spatial intelligence, needs to be collected in the future. Additional quantitative evidence, such as the relationship of ROT scores to other measures of interest, would also be useful to build a nomological network (Cronbach & Meehl, 1955).

All things considered, ROT is a sound mental rotation ability test for college classrooms supported by the current psychometric evidence. As the educational community continues to collect psychometric evidence at multiple sites for multiple populations, we can become more certain of how instruments function to measure the intended constructs. At present, researchers and instructors can use ROT to examine student mental rotation ability and the sex gap with more confidence than before, while continuing to accumulate validity evidence. In addition, for a spatial ability training program focused on mental rotation, instructors may consider using ROT as a pre and post test to determine the effectiveness of the program, i.e., whether the intended goal is achieved.
References


Arendasy, M. E., & Sommer, M. (2010). Evaluating the contribution of different item features to the effect size of the gender difference in three-dimensional mental rotation using automatic item generation. *Intelligence, 38*, 574-581. doi: 10.1016/j.intell.2010.06.004


IV. The role of spatial ability in students’ progression through organic chemistry

Note to Reader

This chapter includes a manuscript, *The role of spatial ability in students’ progression through organic chemistry*, prepared for the *Journal of Chemical Education*.

Why Spatial Ability?

Spatial ability has been proposed as important for success in the Science, Technology, Engineering, and Mathematics (STEM) fields (Uttal & Cohen, 2012; Wai, Lubinski, & Benbow, 2009). In chemistry course, students need to make associations between molecular geometries and properties of materials. In general chemistry, students need to think about the shape of the molecules, such as H₂S, BF₃, CCl₄, to determine polarity (Wang & Barrow, 2011). In organic chemistry, spatial ability is explicitly applied in course materials to interpret molecular structure and to explain reaction mechanisms. For example, spatial ability can be used to identify R or S configuration for 2-butanol as shown in Figure 4.1 using the Cahn-Ingold-Prelog approach (Cahn, Ingold, & Prelog, 1956). The first step is to assign priorities to the four substituents attached to the chiral carbon and this does not require spatial ability. Students then need to *mentally rotate* the molecule to place the group with the lowest priority (here an H atom) pointed away from them and *visualize from that perspective*. For this step, students who can’t successfully visualize become incapable of following the procedure and accomplishing the task. We reviewed eight organic chemistry textbooks and found that this visualization procedure is represented in seven of them (Brown, 2013; Brown, Foote, & Iverson, 2005; Clayden, 2001; Loudon, 2002; Solomons
& Fryhle, 2009; Vollhardt & Schore, 2011; Wade, 2010). One textbook mentioned that students may not have to mentally visualize, and instead they can use paper or a molecular model set; however, this textbook did not offer a clear and elaborated explanation of alternate strategies (Vollhardt & Schore, 2011).

![Figure 4.1 Assign the Configuration of 2-Butanol Using Visualization](image)

Previous literature has shown that the relationship between spatial ability and chemistry performance varies depending on whether exam items require high order cognitive skills (Pribyl & Bodner, 1987; Turner & Lindsay, 2003). Pribyl and Bodner (1987) used two spatial ability tests for students in organic chemistry: the first is the Purdue Visualization of Rotations (ROT) test to measure the mental rotations component, and the second is the Find-A-Shape-Puzzle (FASP), which reflects the spatial perception component of spatial ability. They found that correlation coefficients between spatial ability and organic chemistry term exam scores varied from -0.13 to 0.39, while spatial ability scores did not correlate with the performance on questions answered by rote memory. Similar results were observed using ROT and FASP tests in general chemistry courses by Carter et al. and Merchant et al. (Bodner & McMillen, 1986; Carter, LaRussa, & Bodner, 1987; Merchant et al., 2012). The above visualization studies focused on the correlation between spatial ability and chemistry exam scores and subscores during the same semester, and did not take into account important factors such as math ability, regardless of the fact that math ability measured by national exams, such as SAT math, is widely used to predict college science performance (Heredia, Xu, & Lewis, 2012; Lewis & Lewis, 2007; Pursell, 2007; Spencer, 1996; Tien, Roth, & Kampmeier, 2002; Turner & Lindsay, 2003). Therefore, it is an
open question for researchers whether spatial ability has a unique effect for student chemistry performance when math is controlled.

Recently, there have been efforts to train college students in visual skills to improve performance in the STEM fields. In a one-year study for 77 college students, Miller and Halpern (2013) found that training in spatial ability improved students’ introductory physics scores. Sorby & Baartmans (1996) developed a ten-week course to train spatial thinking skills in an early college engineering curriculum. In a meta-analysis, spatial ability was found to be trainable for both sexes (Baenninger & Newcombe, 1989). There is an assumption in chemistry that spatial ability is necessary to succeed, so we must visualize to learn. Researchers need to continue the effort to better understand the relationship of spatial ability and long-term students’ academic performance.

In addition to spatial skills, students and experts use alternate strategies to solve structure-related chemistry problems as found by Stieff and colleagues (Stieff, 2007; Stieff & Raje, 2010; Stull, Hegarty, Dixon, & Stieff, 2012). To find the identical pairs of molecule diagrams for symmetric stimuli, students tend to employ analytic heuristics more often than mental rotation after analytical strategies training, which is closer to the experts’ approach (Stieff, 2007). Experts use visualization for the translation of representations, but mostly employ algorithms to spatial tasks, such as analyzing molecular feature, predicting reaction products, drawing a mechanism and devising a synthesis (Stieff & Raje, 2010). Additionally, sex differences were observed in spatial skills (Voyer, Voyer & Bryden, 1995). Specifically, females were found to use analytic strategies more often than males after instruction (Stieff, Ryu, Dixon, & Hegarty, 2012). Because the use of alternate strategies and molecular models can obviate the impact of spatial ability (Stull et al., 2012), and tasks requiring spatial skill may only represent a small part of chemistry
performance, student spatial ability is not necessarily associated with exam score or course grade. Devetak and Glaža found that visualization is only weakly correlated with achievements of secondary students in reading and drawing submicrorepresentations (Devetak & Glažar, 2009). Stieff found that an instruction using an analytical algorithm is more effective than an intervention of mental rotations for 40 college students of both sexes (Stieff, 2013), and effect of training on achievement depends on strategy choice (Stieff, Dixon, Ryu, Kumi, & Hegarty, 2014). Accordingly, it is undetermined whether students have to score high on spatial ability tests to be able to progress through organic chemistry courses.

**Research Questions**

Because there is a concern that the lack of spatial ability can be a barrier for students’ progression through organic chemistry, we intended to track students of varying spatial ability from general chemistry to advance through organic chemistry. Specifically, two research questions guide this study:

1. Were the students who could mentally visualize, as measured by the Purdue Visualization of Rotations Test (ROT), more likely to progress through organic chemistry than those who could not mentally visualize?
2. For students who progressed through organic chemistry, did their mental rotations ability make a difference in their organic chemistry course grades when factors (e.g., math ability) are controlled?

**Research Methods**

**Instruments**

In this study, ROT (Bodner & Guay, 1997) was used to measure the mental rotations component of spatial ability. ROT is chosen because it has accumulated significant psychometric
evidence according to the standard (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999; Arjoon, Xu, & Lewis, 2013). ROT has evidence of temporal reliability and construct validity without showing test bias concern against females (submitted to the journal Learning and Individual Differences). ROT is chosen also because it is frequently used in STEM areas (Bodner & Guay, 1997; Brownlow, McPheron, & Acks, 2003; Carter et al., 1987; Morgil, Yavuz, Oskay, & Arda, 2005) to capture a mental rotation component which is important to solve chemistry problem. Both chemistry experts and students utilize this strategy to solve chemistry problems, such as translating molecular representations (Stieff, 2011, 2013; Stieff & Raje, 2010). The test takers were required to finish all 20 items in 10 minutes and they were not allowed to make any mark on the booklet; this minimizes the possibility of using analytical reasoning strategies.

**Participants and data collection**

ROT was given to the students enrolled in the first semester of general chemistry course (General Chemistry I) at a large Southeastern public research university. Data were collected during the second week of four separate semesters (Fall 2008, Spring 2009, Fall 2009, and Spring 2010). More details about data collection and participants’ information are in the submitted manuscript (submitted). Overall there were 3843 students taking the ROT test during the four semesters. We tracked all these students to see if they entered further organic chemistry by Spring 2013. This choice of sufficient time frame allows students to enter the later courses because students (around 90%) take organic chemistry sequence within three years of taking General Chemistry I. Overall, 1062 students progressed through both Organic Chemistry I and II courses, while the other 2781 did not.
All students in this study needed to take final exam in order to get a grade in general chemistry. The First Term General Chemistry Special Exam developed by the American Chemical Society Division of Chemical Education (ACS exam thereafter), which has 40 multiple-choice items (19 for algorithmic and 21 items for conceptual questions) and was used as part of the final exam (Examinations Institute, 1997). Because the ACS score has been supported by psychometric evidence at multiple sites and more generalizable across institutions, it is used to indicate the latent variable of general chemistry achievement (Lewis & Lewis, 2007). For organic course, we don’t have access to their exam so we used the course grade as a proxy of course performance. Although course grade may not be a good measure of student learning and has a lot of things in it, it is the course outcome recorded on official academic transcript and carries a weight for students, advisors and future employer to make suggestions or decisions. In this study, grades when students first time completed the course, if multiple attempts, were used to minimize the extra exposure that could potentially bias the true relationship. Course grade was converted from letter grade to grade point (A = 4, A− = 3.67, B+= 3.33, B = 3, C = 2, and so on) according to university grading policies. Overall, 876 students had complete data of SAT Math, ROT, ACS, and grades in later organic chemistry courses.

Statistical data analysis

Comparison of students who can or cannot visualize

If the assumption holds that the students need to have higher spatial ability in order to progress through organic chemistry, we expect the spatial ability for the students who progressed through be much higher than the other students who did not. To check if this statement is true and to answer research question 1, we need to examine the association of student progression through organic chemistry (yes or no) with mental rotation ability (high or low). The upper 27%
(ROT total score \(\geq 16\)) and lower 27\% (ROT \(\leq 9\)) were used to separate students of different mental rotation ability level. For students groups who progressed through organic chemistry and those who did not, performance differences regarding SAT math, ROT and ACS scores were calculated. Cohen’s \(d\) cutoff was used to quantify the magnitude of difference across sexes (0.2 small, 0.5 medium, and 0.8 for large effect size) where comparing group means. Simple Pearson correlation was performed for these 876 students. Coefficients \(r\) at 0.1, 0.3, and 0.5 were used to indicate the correlation strength as weak, medium and strong effect sizes (Cohen, 1988).

**Structural equation modeling for effect of mental rotations on organic chemistry course grades**

SEM is a multivariate data analysis approach that can answer research questions such as how well a theoretical model fits empirical data, and how multiple variables interact with one another (Raykov & Marcoulides, 2006). By taking the measurement error into account, SEM has the advantage of more accurate estimate of the relationships among latent and measured variables (Kline, 2010, p.71). SEM in this study was performed using Mplus 5.2 (Muthén & Muthén, 2007) to examine the direct and indirect effect of math ability, sex and mental rotations to general chemistry achievement and Organic Chemistry I and II course grades. The measurement part for ROT items was based on the confirmatory factor analysis (submitted). Path analysis parameters, such as model specification, model fit, standardized parameter estimates and effect sizes for path coefficients were reported in compliance with the previous guidelines (Kline, 2010; X. Xu, Villafane, & Lewis, 2013). Similarly as in the CFA analysis, WLSMV was used, the model was identified by fixing the first item on the factor at 1, and the same criteria (CFI > 0.95 and WRMR < 1) was used to judge a good model fit to the data. Cohen’s \(f^2\) was calculated.
according to path coefficient, $f^2$ at 0.02, 0.15, and 0.35 is considered as small, medium and large effect sizes (Cohen, p. 413).

**Results**

*Descriptive statistics*

Descriptive analysis was performed for each variable. Results indicate similar normal distribution for the overall sample and the cohort with complete data for all the interested variables, which allows for following statistical inferences. Table 4.1 presents here are descriptive for the 876 students, who are used for later SEM analysis, including mean, stand deviation for each variable by sex. Overall, males tend to outperform females at varying magnitude. According to Cohen’s $d$ cutoff, the magnitude of difference across sex is quantified as medium in spatial ability at 0.61, small in math at 0.27, small in general chemistry ACS exam at 0.23, and trivial in organic chemistry course grades.

**Table 4.1**

*Descriptive of SAT Math, ROT, ACS, and Organic Chemistry Grade by Sex*

<table>
<thead>
<tr>
<th></th>
<th>Mean (SD)</th>
<th>Cohen’s $d$ in sex difference</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>M</strong></td>
<td><strong>F</strong></td>
<td></td>
</tr>
<tr>
<td>SAT Math</td>
<td>611.6 (79.1)</td>
<td>593.5 (74.5)</td>
</tr>
<tr>
<td>ROT</td>
<td>14.3 (3.6)</td>
<td>12.1 (3.8)</td>
</tr>
<tr>
<td>ACS in general chemistry</td>
<td>29.1 (5.8)</td>
<td>27.8 (5.8)</td>
</tr>
<tr>
<td>Organic I grade</td>
<td>2.9 (0.7)</td>
<td>2.8 (0.7)</td>
</tr>
<tr>
<td>Organic II grade</td>
<td>2.7 (0.9)</td>
<td>2.7 (0.9)</td>
</tr>
</tbody>
</table>

Note: sample size is 393 for male, and 483 for female

Pearson correlations among variables are shown in Table 4.2. ROT correlates with SAT math at 0.43 (a medium effect size), because both mental rotations ability and SAT math measure some aspects in cognitive ability. ROT correlates significantly with General Chemistry I
achievement during the same semester as measured by the ACS scores, with the coefficient at .30 (a medium effect size), which is consistent with previous literature (Bodner & Guay, 1997; Carter et al., 1987). ROT correlates with later Organic I and organic II course grades at 0.15 and 0.08 respectively (small effect sizes), which are at the low range of literature results when spatial tests data and course grade were obtained during the same semester (Pribyl & Bodner, 1987). All correlations are statistically significant with $p < 0.05$. The magnitudes of sex difference and correlation strength from these 876 students are very similar to results of overall 3843 students.

Table 4.2

Pearson Correlations among SATM, ROT, and ACS Scores, Course Grade for Organic Chemistry I and II

<table>
<thead>
<tr>
<th></th>
<th>ROT</th>
<th>ACS</th>
<th>Grd Organic I</th>
<th>Grd Organic II</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATM</td>
<td>0.43</td>
<td>0.54</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>ROT</td>
<td>0.30</td>
<td>0.15</td>
<td>0.08</td>
<td></td>
</tr>
<tr>
<td>ACS</td>
<td></td>
<td>0.33</td>
<td>0.29</td>
<td></td>
</tr>
<tr>
<td>Grd Organic I</td>
<td></td>
<td></td>
<td>0.53</td>
<td></td>
</tr>
</tbody>
</table>

Note: * $p < 0.05$; all the other correlations, $p < 0.001$, $n=876$

Simple correlation coefficient $r$: 0.1, 0.3, 0.5 as weak, medium and strong effect size

Comparison of students who can or cannot mentally visualize

As shown in the frequency table 4.3 below, 849 students scored at the lower 27% of ROT score, indicating they probably cannot mentally visualize. Among them, 208 (24.5%) students progressed through organic chemistry course sequence. For the 1066 students with high mental rotation ability scores, 29.8% progressed through organic chemistry. Because both variables are treated as categorical, Chi-square test was performed. The results from Chi-square test ($\chi^2(df = 1, n = 1915) = 6.7, p = .009$) indicated the two variables are not independent. Effect size $w$ was
0.12, which is quantified as small according to Cohen’s w guideline (small =0.1, medium=0.3, large =0.5).

**Table 4.3**

*Mental Rotation vs Progression through Organic Chemistry Course Sequence*

<table>
<thead>
<tr>
<th>Progress thru org. chem.</th>
<th>Mental rotation ability level by ROT score</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low (n=849)</td>
<td>High (n=1066)</td>
</tr>
<tr>
<td>No (freq, %)</td>
<td>641 (75.5%)</td>
<td>748(70.2%)</td>
</tr>
<tr>
<td>Yes (freq, %)</td>
<td>208 (24.5%)</td>
<td>318 (29.8%)</td>
</tr>
</tbody>
</table>

Note the math ability is correlated with mental rotation and could confound the relationship. SAT math score for both visual ability groups were examined. The upper 27% group on ROT has a SAT math at 627.6 (SD = 73.4), which is significantly higher than the lower group at 541.0 (SD = 71.7), we need to look at the possible effect of math ability.

![Figure 4.2 Comparing Students Who Progressed through Organic Chemistry (OC) or Not, error bars represent the standard deviation](image)

From Figure 4.2, the 1062 students who progressed through organic chemistry course sequence did not show a sizable higher score on ROT than others who did not, with a small
effect size at 0.1 SD based on Cohen’s standards (Cohen, 1988). On the contrary, students progressed through outperformed on SAT math (medium effect size at 0.5) and ACS (around large effect size at 0.8) than those who did not. Considering SEM results that math and general chemistry performance are more important than spatial ability to influence students’ later course grade, similarly, math and prior chemistry are more likely to influence students’ decision to take organic chemistry courses. Therefore, there is little concern that only students with high ROT scores can enter organic chemistry.

*Structural equation modeling for the effect of mental rotations on organic chemistry course grade*

SEM was performed to answer research question 2 by taking into account of measurement error and investigating multiple variables simultaneously. Among these 876 students who had complete data, 90.1% took the Organic Chemistry I with two year of taking general chemistry I, and 92.3% took the Organic Chemistry II within one year of taking Organic Chemistry I. The full SEM model was specified in Figure 4.2. The measurement part for the 20 ROT items was based on the previous established bi-factor structure (submitted). For the regression part, sex and mental rotations are set to have an effect on general chemistry achievement and follow up organic chemistry course grades. These three exogenous variables/predictors are set to correlate with each other. If the model fit data well, we focused on regression part. For example, H1, H2, and H3 are testing the direct effect of sex on chemistry performance, while H4, H5 and H6 are for the direct effect of mental rotations on chemistry performance. Alternate models were created with less covariates to examine the effect of mental rotations from multiple models (please find the model in the supplement B).
Figure 4.3 presents the SEM model coefficients. The model fitness to the data of 876 students is: $\chi^2 (n = 876, df = 167, p = 0.004) = 219$, CFI = 0.98, WRMR = 0.89. Using the criteria of CFI >0.95 and WRMR < 1, results supports the tenability of the proposed SEM model. Then we can use the parameter estimate in Figure 4.4 to interpret the effect of math ability, sex, and mental rotations on chemistry performance.

Figure 4.3 SEM Model with Unstandardized Parameter Estimate; Female was coded 1 and male 0. The numbers on the line are unstandardized path or correlation coefficients.

Relationship between three predictors. The correlation coefficient between sex and mental rotations is -0.45, which is consistent with the sex difference in mental rotations which favors male students. The correlation between sex and math is -0.15, which means the males tend to score higher than females in this population. The correlation between ROT and SAT
math is 0.52, which means the students with higher math ability tend to have higher mental rotations ability.

Figure 4.4 SEM Model with Standardized Parameter Estimate; Female was coded 1 and male 0. The numbers on the line are standardized path or correlation coefficients.

*Relationship between math ability and chemistry performance.* As shown in Table 4.3, the direct effect of math ability on general chemistry achievement is 0.58, which is quantified as large and consistent with literature results (Lewis & Lewis, 2007). As for Organic Chemistry I grade, because prior general chemistry achievement was added into prediction, the direct effect of math is downplayed, while math only has a sizable indirect effect through the General Chemistry course. As for Organic Chemistry II grade, math has a direct effect with a standardized coefficient at 0.19, which is quantified as small. Only the Organic Chemistry I grade has a direct effect with a large effect size on Organic II grade.
### Table 4.4

*Standardized Path Coefficient and Effect Size for Chemistry Performance*

<table>
<thead>
<tr>
<th>Parameter estimate</th>
<th>Gen. Chem. achievement</th>
<th>Organic I course grade</th>
<th>Organic II course grade</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Path Coefficient</td>
<td>Effect size</td>
<td>Path Coefficient</td>
</tr>
<tr>
<td>Math ability</td>
<td>0.58</td>
<td>0.51</td>
<td>-0.03</td>
</tr>
<tr>
<td>Sex</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.04</td>
</tr>
<tr>
<td>Mental rotations</td>
<td>0.11</td>
<td>0.01</td>
<td>-0.05</td>
</tr>
<tr>
<td>Chem. achievement</td>
<td>-</td>
<td>-</td>
<td>0.38</td>
</tr>
<tr>
<td>Organic I grade</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: Cohen’s $f^2 = 0.02$, 0.15, and 0.35 as small, medium and large effect sizes, Cohen, p. 413

**Relationship between sex and chemistry performance.** From Table 4.3, the direct effect of sex is -0.01, -0.04 and 0.04 for general chemistry achievement, Organic I and II course grades, which is not statistically significant and the effect size is trivial. This means, although sex is related with math and mental rotations, sex has no direct effect on chemistry performance when other variables are controlled. In other words, students with the same math and mental rotations ability are able to perform equally well in chemistry regardless of sex.

**Relationship between mental rotations and chemistry performance.** Mental rotations has a significant direct effect on general chemistry achievement when sex and Math ability are controlled.

Regarding Organic Chemistry I course grade, the direct effect of mental rotations is found to be not significantly different from zero when math and the general chemistry achievement effect is controlled. For Organic Chemistry II course grade, the effect of mental rotations is not significantly different from zero while math ability and Organic Chemistry I grade were controlled. The finding from this full model is consistent with that from alternate model when the predictor of general chemistry performance was removed (in supplement).
In sum, mental rotations skill has a direct effect on chemistry achievement during the same semester when math ability is controlled. In the later Organic I and II, the course grade is directly affected by the previous chemistry measure, while the math and rotations ability have no or small direct effect.

Discussion

This study has several advantages over previous reports. It tracks a large number of students, analyzes relationships among multiple variables simultaneously by SEM, and uses a spatial ability instrument widely used in STEM areas with established psychometric evidence (Xu & Lewis, Submitted; Yoon, 2011), to answer research questions. Question one asks whether students who score high on ROT are more likely to progress through Organic Chemistry. This study reveals a statistically significant but weak association between ROT score and advancing through Organic Chemistry when other variables are not included. Students who progressed through tend to score slightly higher on ROT than those who did not however, SAT math and ACS exam scores in General chemistry between these groups are more significantly different, indicating possible confounding factors that could affect whether or not students progress through organic chemistry courses. It has been reported that general reasoning skills and prior content knowledge, rather than visualization skills, are more relevant regarding students’ future academic performance (Baker & Talley, 1972; Keig & Rubba, 1993; Wu & Shah, 2004). On the contrary, literature has also reported that spatial ability could be more important than math ability regarding student retention in CEPS from a relative small sample of 49 students (Kinsey et al., 2008). Accordingly, we should not rush to the claim that students have to score high on ROT to progress through Organic Chemistry.
Regarding question two which deals specifically students who have progressed through Organic Chemistry, it is worth investigation how spatial ability, math ability and other abilities correlate with one another and influence students’ performance in Organic Chemistry courses. This study included 876 students from General Chemistry I who progressed into Organic Chemistry I and II courses to allow for more accurate estimates of relationships among the interested variables. From an SEM model that can best explain data, the mental rotations ability did not show a significant effect on organic chemistry course grades. The modeled results suggest sex did not show a significant effect on performance in Organic Chemistry when other variables were controlled. In other words, regardless that sex gap as measured by ROT persistent in literature (Maeda & Yoon, 2013), being a female or doing poorly on a spatial ability test like ROT does not constitute a big concern for getting good grades in the later course sequence of Organic Chemistry as long as students do well on SAT math and General Chemistry exams.

This study has several limitations. First, there are other possible factors that can affect student performance that are not captured here. The role of spatial ability in learning chemistry can be overshadowed by many factors. For example, spatial components other than mental rotations and student approaches to solve chemistry problems are not available, and the actual learning strategies are not measured. A recent study has revealed that both spatial ability and analytic strategies are important for achievement while math ability is not included in the SEM model (Stieff, Dixon, Ryu, Kumi, & Hegarty, 2014). Second, course grade is used as a proxy for student performance, but is far from an ideal measurement. Because course grade policy varies regarding the existence of curve and grade points can consist of many elements such as homework, attendance points or instructor-created exams, a final grade does not have the same meaning across instructors and semesters and does not necessarily reflect students’ meaningful
learning. For students who get a good grade in organic chemistry courses, we don’t know whether they were able to solve stereochemistry problems or what strategies they used. It is still undetermined whether spatial ability components play an important role in specific chemistry problems, and this is worth further examination (Stieff, 2011). Finally, ROT was administered for students at one institution at the beginning of General Chemistry to avoid the stereotype that students need visualization in order to succeed in organic chemistry. However, it should be noted that the effect of spatial ability could be downplayed here due to the interval students have between each course. Students could have already improved their mental rotations when they started to take organic chemistry courses. So administering ROT during the same semester of organic chemistry may reveal a significant relationship and are worth further examination at more institutions.

This study raises many interesting questions for future research. For example, what is the relationship between a specific type of stereochemistry problem and a relevant spatial ability component not limited to mental rotations? What is the relationship between a specific type of stereochemistry problem and a relevant spatial ability component not limited to mental rotations? Future studies can also investigate the role of spatial ability with more nuanced discriminations in each component and corresponding training tasks. Educators may be interested in the effective approaches to teach spatially related problems.

This study has practical implications in chemistry instruction. One implication of this study is special caution when considering the claim that females in general lack spatial ability; therefore, females cannot learn organic chemistry well until they receive extra spatial ability training. This study did not support that concern in the studied setting. Based on this study, females students who score low on a widely-used spatial ability test like ROT, should not be
scared away from entering more advanced courses like organic chemistry in which lack of sufficient visualization seems to be a barrier.

Furthermore, it is traditionally emphasized by many organic chemists that students must be able to visualize molecular structures in their minds in order to master the content knowledge (Coleman & Gotch, 1998; Harle & Towns, 2011). Taking the previously discussed task in Figure 4.1 as an example, most textbooks require visualization skills to solve chiral configurations following the original approach Cahn et al. developed. If instructors assume students should be able to visualize as an important course objective, a positive relationship is expected between visualization skills and course grades (Wai et al., 2009). However, this study did not find such a relationship, indicating that instructors may not have aligned this objective with exams or grading policies. In order to support this assumption, students may be required to score high on ROT to enter Organic Chemistry course, course instructions can be adjusted to target visualization, and grading policies need to change to be able to fail students who cannot visualize to solve space-related problems.

Instead of assuming everyone needs to be able to visualize in order to succeed in organic chemistry, some educators believe that students can achieve meaningful learning through multiple approaches which are not limited to visualization (Stieff et al., 2014). For these educators, the findings from this study are not a surprise, because course grades are typically composed of many things other than mental rotations ability. The students without strong mental rotations ability, whether male or female, may resort to alternate spatial skills, use analytic strategies, build molecular models, or learn chemistry in their own comfortable ways. To better help students who don’t visualize, alternative approaches, such as analytic strategies and molecular models, should be provided. As shown by studies of Stieff et al, people who use
diagrammatic analytic strategies to solve space-related chemistry problems are equally successful (Stieff, 2007; Stieff & Raje, 2010), and the group that received training in analytical strategies performed even better than those who received spatial skills training (Stieff, 2013). Active instructional approaches like the Process Oriented Guided Inquiry Learning can also support students at different levels of spatial ability to learn abstract concepts such as chemical applications of group theory (Southam & Lewis, 2013). Since both sexes can benefit from molecular visualization tools and training (Baenninger & Newcombe, 1989; Miller & Halpern, 2013), all students should have access to valid visualization tools and activities. Textbooks and instructions should clearly elaborate on the procedures and reduce the high working memory load required for structure-related chemistry problems. One out of eight organic chemistry textbooks we examined provide a great example to guide students to use alternate strategies and minimize visualization skills (Bruice, 2012). For students who cannot visualize and do not have access to Bruice’s textbook, there are some popular videos introducing alternate strategies on Youtube, such as Organic chemistry: R and S naming (http://www.youtube.com/watch?v=GxN8k7OniA4), and Converting Line Diagrams to Fischer Projections (http://www.youtube.com/watch?v=VGRARhqr3tI). More work needs to done for educators to provide guidance to students regarding the validity of using these online approaches. With more resources on alternate strategies and more entering students, we have a better chance of encouraging them to stay and succeed in chemistry and STEM areas.

References


73


V. Attitude toward the Subject of Chemistry in Australia: An ALIUS and POGIL Collaboration to Promote Cross-national Comparisons

Note to Reader


Introduction

A collaborative network of like-minded tertiary science educators was formed as part of Active Learning in University Science (ALIUS) (Bedgood, 2008). The nascent project “Developing leaders of change in the teaching of large university chemistry classes”, funded by the Australian Learning and Teaching Council, focused on capacity building to drive student-centred curriculum design and pedagogy in the discipline (Bedgood, Yates, et al., 2010). The aim was to develop local leadership capacity, by forging partnerships between local and international educators, that would guide sustainable and long-term embedded practice (Bedgood, Bridgeman, et al., 2010). One such student-centred strategy introduced was Process Oriented Guided Inquiry Learning (POGIL), where students work through carefully crafted activities in small groups (POGIL, 2012). The use of POGIL in the classroom has demonstrated benefits to
students in terms of content mastery, higher-order thinking skills, metacognition, and teamwork (Moog et al., 2009).

POGIL has been implemented in our large first year chemistry classes at Curtin University by replacing some or all of the didactic lecture presentation with small group work, clicker questions and online mini-lectures in a blended learning environment (McIntyre, 2010). Dissatisfaction with the current student-teacher paradigm motivated our change. We felt a need to improve the attitude of our students toward the discipline by enhancing engagement, and this potential was identified within POGIL. As we seek to encourage others to adopt similar strategies for similar purposes the question arises: are these aims met at Curtin and what tools are available to assess the impact, in these realms, of innovations in the classroom?

Cognition and affect are two domains central to the aims of implementing POGIL at Curtin. To focus on attitude within the affective domain, a short-form eight-item semantic differential instrument, Attitude toward the Subject of Chemistry Inventory (ASCIv2), was chosen as key assessment tool (Xu & Lewis, 2011). ASCIv2 captures responses on a seven-point semantic differential scale to eight items regarding “intellectual accessibility” and “emotional satisfaction”, reflecting the cognitive and affective components of a tripartite attitude framework.

Here we will present preliminary findings from data collection in two POGIL classes at Curtin. Before commencing broader studies we must ascertain whether ASCIv2 will perform, in the local context, within psychometric standards for reliability and validity. Secondly, we can compare the preliminary findings to existing research using the same instrument. Finally, we can demonstrate how networks can demonstrate efficacy through the use of such an instrument and build knowledge internationally about education of science at the tertiary level.
**Instrument**

Attitude toward the Subject of Chemistry Inventory (ASCI) was originally designed by Bauer to measure students’ attitude toward chemistry (Bauer, 2008). ASCIv2 is a refined version that contains only eight items in two subscales, “intellectual accessibility” (e.g., how difficult is chemistry?) and “emotional satisfaction” (e.g., how comfortable is chemistry?), and performs well from a psychometric perspective (Xu & Lewis, 2011). In this study, ASCIv2 was used to collect data from Curtin students during October 2011. Although the instrument has been used successfully at two different US universities, Curtin students may have culturally-influenced interpretations of the instrument items that are different from those of US students, so it is necessary first to examine reliability and validity evidence for the data obtained at Curtin.

**Sample**

Curtin students from two different classes participated in the data collection during October 2011. Class I is a second semester general chemistry course undertaken by Science (34%), Engineering (59%) and double degree (6%) students. The students from Science undertake this course in the first year, second semester of their studies, while those from Engineering and double degrees primarily undertake it in their second year. This class is predominantly male (80%) and from diverse cultural and linguistic backgrounds from Asia, Africa and the Middle East, in line with the general student population at Curtin. Class II is a second semester, first year organic and biological chemistry course undertaken by health sciences students studying science majoring in biomedical science (33%), nutrition (52%) and food science (3%). This class is predominantly female (83%) and comprises mostly domestic students (88%).
From the data collected, there are 61 students from one class (Class I, 61/251 = 24%) and 53 from the other (Class II, 53/107 = 50%). Based on descriptive analysis for each class separately, during which no major differences emerged, the two classes were combined for this report. Overall, 108 students had complete data (answered all eight items) and are included in the following analysis. The ASCIv2 results collected from students at a large US public research university, near the end of the semester are presented for comparison. Similar findings have been reported at another medium-sized public research university in the mid-western United States (Brandriet, Xu, Bretz, & Lewis, 2012). The data at the US university was collected from students who were enrolled in the first general chemistry class in the university curriculum. This class is taught with a tested and proven POGIL approach (Lewis & Lewis, 2005, 2008).

**Data Analysis Methods**

First, basic descriptive statistics for the eight items of ASCIv2 were obtained using SAS software version 9.2. The results are presented below in Table 5.1A along with the results from the US research university in Table 5.1B.

Next, evidence for reliability and validity was examined. For these analyses, the two data sets from Curtin were combined. For reliability, Cronbach’s alpha estimates were calculated using SAS 9.2. Cronbach’s alpha indicates the degree of internal consistency of the items within a particular scale. A high Cronbach’s alpha suggests that the items scores are positively correlated with each other and with the total scale score as well. The larger the Cronbach’s alphas for a set of items, the greater the assurance the items measure the same construct. The upper bound for Cronbach’s alpha is 1.0, and values of at least 0.7 are generally desirable for research purposes (Murphy & Davidshofer, 2005).
For validity, evidence based on internal structure was obtained to support construct validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). Confirmatory factor analysis (CFA) was performed in Mplus 5.2 to estimate how well the designed two-factor correlated structure for the instrument fit the responses obtained with the sample (Muthen & Muthen, 2007). In addition to recommended cutoffs for model fit statistics, factor correlation and item loadings were used to evaluate the model fit (Bentler, 1990; Hu & Bentler, 1999).

Finally, once validity and reliability evidence supported score interpretation, students’ attitude scores from Curtin were compared with scores from US students, using a standard effect size method (Cohen, 1988).

**Results**

**Descriptive Statistics**

The four negatively stated items (#1, 4, 5, and 7) were reversely coded for ease of interpretation. This means, the highest score of 7 on the original data was transformed to 1, and the lowest score of 1 was changed to 7, and so on. After recoding, the score interpretation for all eight items is in the same direction. A score higher than middle point 4 means positive attitude, i.e., students feel chemistry is intellectually accessible or emotionally satisfying. Descriptive statistics are shown in Table 5.1A for each item for both classes. The average scores range from 2.54 to 4.59, and standard deviations range from 1.16 to 1.54. No item was found to have skewness or kurtosis greater than .90, which suggests good normality of the item scores. The mean scores for most items are around the middle point of 4. The items with the highest mean scores are # 4, 5, 8, which means students in general feel chemistry is comfortable, satisfying and
organized. Item 6 has the lowest mean score, which means students in both classes feel chemistry is challenging.

**Table 5.1A**

*Descriptive Statistics for Curtin after Recoding Items 1, 4, 5 and 7*

<table>
<thead>
<tr>
<th>Item</th>
<th>Class I (n=58)</th>
<th>Class II (n=50)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
</tr>
<tr>
<td>1*Hard/easy</td>
<td>3.78</td>
<td>1.27</td>
</tr>
<tr>
<td>2 Complicated/simple</td>
<td>3.67</td>
<td>1.28</td>
</tr>
<tr>
<td>3 Confusing/clear</td>
<td>3.79</td>
<td>1.41</td>
</tr>
<tr>
<td>4* Uncomfortable/comfortable</td>
<td><strong>4.59</strong></td>
<td>1.34</td>
</tr>
<tr>
<td>5*Frustrating/satisfying</td>
<td>4.34</td>
<td>1.55</td>
</tr>
<tr>
<td>6 Challenging/unchallenging</td>
<td><strong>3.29</strong></td>
<td>1.17</td>
</tr>
<tr>
<td>7* Unpleasant/pleasant</td>
<td>4.29</td>
<td>1.31</td>
</tr>
<tr>
<td>8 Chaotic/organized</td>
<td>4.57</td>
<td>1.43</td>
</tr>
</tbody>
</table>

From comparison to US data in Table 5.1B, there is a similar trend for the 8 items. From the items with extreme scores (#6 & 8), both groups of students feel chemistry is organized, and is challenging. Because of the measurement error for the observed item scores, we will not compare the group difference based on item scores.

**Table 5.1B**

*Descriptive Statistics for the US University for Comparison*

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*Hard/easy</td>
<td>2.81</td>
<td>1.28</td>
</tr>
<tr>
<td>2 Complicated/simple</td>
<td>2.95</td>
<td>1.43</td>
</tr>
<tr>
<td>3 Confusing/clear</td>
<td>3.36</td>
<td>1.40</td>
</tr>
<tr>
<td>4* Uncomfortable/comfortable</td>
<td>3.64</td>
<td>1.36</td>
</tr>
<tr>
<td>5*Frustrating/satisfying</td>
<td>3.24</td>
<td>1.70</td>
</tr>
<tr>
<td>6 Challenging/unchallenging</td>
<td><strong>2.50</strong></td>
<td>1.50</td>
</tr>
<tr>
<td>7* Unpleasant/pleasant</td>
<td>3.38</td>
<td>1.41</td>
</tr>
<tr>
<td>8 Chaotic/organized</td>
<td><strong>4.26</strong></td>
<td>1.66</td>
</tr>
</tbody>
</table>

*Note: Each item score ranges from 1 to 7, while 4 is the middle point. Bolded values indicate the range (largest and smallest values in a column.) A high value means students feel chemistry is intellectually accessible, emotionally satisfying. Adapted with permission from X. Xu and J. E. Lewis, *Refinement of a Chemistry Attitude Measure for College Students*, J. Chem. Educ. 2011, 88, 561–568, Table 2. Copyright 2011 by American Chemical Society and Division of Chemical Education, Inc. Item score is reversed before averaging; items are shown reversed for ease of interpretation.
Validity Results from CFA

CFA was performed to estimate goodness of fit for the designed 2-factor model. Items 1, 2, 3 and 6 were set to load on factor “intellectual accessibility” only; items 4, 5, 7 and 8 were set to load on factor “emotional satisfaction” only, and the two factors were allowed to correlate. The model was identified by fixing the first item on each factor at 1.

The estimation, illustrated in Table 5.2, of the 2-factor model fit for the Curtin data is: $\chi^2 (N = 108, df = 19, p < .001) = 53$, CFI = .94, SRMR = .06. Since CFI was close to .95, and SRMR is less than .08, we can consider that the overall model does fit the data reasonably well. Factor correlation is .77. From Table 5.3, all eight items are loaded well on their intended scales. A factor loading represents the correlation between that item and the factor as a whole.

Table 5.2

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>53</td>
<td>77</td>
</tr>
<tr>
<td>p-value</td>
<td>.00</td>
<td>.00</td>
</tr>
<tr>
<td>df</td>
<td>19</td>
<td>19</td>
</tr>
<tr>
<td>CFI</td>
<td>.94</td>
<td>.95</td>
</tr>
<tr>
<td>SRMR</td>
<td>.06</td>
<td>.04</td>
</tr>
</tbody>
</table>

2-Factor model fit for Curtin is similar with the results from a large US research university results. As shown in Table 5.2, the estimation of the 2-factor model fit with data obtained from a US university is: $\chi^2 (N = 354, df = 19, p < .001) = 77$, CFI = .95, SRMR = .04. Results from US students are very comparable to the fit obtained with the Curtin data, which supports the tenability of the 2-factor structure for the ASCIv2 instrument.
Table 5.3

*CFA Item Loadings for the 2-Factor Solution (N=108 for Curtin 2011)*

<table>
<thead>
<tr>
<th>Intended Factor</th>
<th>Item #</th>
<th>Item Loading</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Accessibility</td>
<td>Item 1</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Item 2</td>
<td>0.89</td>
</tr>
<tr>
<td></td>
<td>Item 3</td>
<td>0.83</td>
</tr>
<tr>
<td></td>
<td>Item 6</td>
<td>0.55</td>
</tr>
<tr>
<td>Emotional Satisfaction</td>
<td>Item 4</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Item 5</td>
<td>0.87</td>
</tr>
<tr>
<td></td>
<td>Item 7</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>Item 8</td>
<td>0.51</td>
</tr>
</tbody>
</table>

* All loadings are significantly different from 0 at the \( p=.05 \) level.

**Reliability**

The internal consistencies were estimated for each subscale. Cronbach’s alpha for Curtin students is .84 for the accessibility scale, which is above the satisfactory level of .7 and comparable to that obtained from a large US research university. Cronbach’s alpha is .84 for the satisfaction scale, which is again above .7, and also comparable to a large US research university as shown in Table 5.4.

Table 5.4

*Internal Consistency Reliability by Cronbach’s Alpha for ASCI-V2*

<table>
<thead>
<tr>
<th>Subscales of ASCI</th>
<th>Curtin 2011(n=108)</th>
<th>US(n=354)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Accessibility (1,2,3,6)</td>
<td>.84</td>
<td>.82</td>
</tr>
<tr>
<td>Emotional Satisfaction (4,5,7,8)</td>
<td>.84</td>
<td>.79</td>
</tr>
</tbody>
</table>

**Factor Score Comparison**

Based on the evidence of internal consistency reliability and factorial validity, it is safe to combine the items within each factor to calculate a factor score. Factor scores (for intellectual accessibility and for emotional satisfaction) and factor reliability by Cronbach’s alpha are
recalculated and shown in Table 5.5. The intellectual accessibility factor is the average score of three items 1, 2, 3, and 6. The emotional satisfaction factor is the average of four items of 4, 5, 7, and 8. All negative items are reverse coded before summing, and the possible factor score may range from 1 to 7, with the 4 at the middle point. Higher values mean greater intellectual accessibility and greater emotional satisfaction.

Although students are from two countries, we cannot make generalizations for comparing attitude levels between the two countries because the sample is not representative of the population of each country. Likewise, there are also differences in the data collection, where US students are in the same course and combined, but not so in Australia. However, we can have an impression of the overall attitude status for students in each chemistry program in this study. As a result, effect size is more meaningful than a t-test when it is comparing samples with different sample sizes in different contexts.

Cohen's $d$ effect size can be used to quantify the difference in the attitude between US students and Curtin students and are provided in Table 5.5 (Cohen, 1988). Values for Cohen's $d$ of 0.2 to 0.3 are generally considered a "small" effect, around 0.5 a "medium" effect and 0.8 or above, a "large" effect. A medium effect size reflects a difference that would be noticeable to a careful observer. For the two classes at Curtin, Class I students feel chemistry is more accessible (by .46 standard deviation units, a medium effect size) and slightly more emotionally satisfying (by .36 standard deviation units, a small effect size) than Class II students do. The more dramatic differences emerge when the Curtin students from Class I are compared to the students from the US. In this case, all of the effect sizes are above .5, signaling observable differences between attitudes in the US and at Curtin. The t-Test in Table 5.3 indicates that the group differences is statistically significant between Curtin class I and the US students.
Table 5.5

Factor Scores for Each Subscale for US and Curtin Students

<table>
<thead>
<tr>
<th>School</th>
<th>Variable</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>p value for t-Test</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Curtin I</td>
<td>Accessibility</td>
<td>3.63</td>
<td>0.98</td>
<td>-0.49</td>
<td>0.79</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>4.45</td>
<td>1.10</td>
<td>-0.21</td>
<td>-0.17</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Curtin II</td>
<td>Accessibility</td>
<td>3.14</td>
<td>1.16</td>
<td>-0.24</td>
<td>-0.55</td>
<td><em>p = .02</em></td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td>(N=50)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>4.02</td>
<td>1.31</td>
<td>-0.51</td>
<td>-0.10</td>
<td><em>p = .07</em></td>
<td>.36</td>
</tr>
<tr>
<td>US</td>
<td>Accessibility</td>
<td>2.91</td>
<td>1.13</td>
<td>0.32</td>
<td>-0.25</td>
<td><em>p &lt; .001</em></td>
<td>.65</td>
</tr>
<tr>
<td>(N=354)</td>
<td>Satisfaction</td>
<td>3.63</td>
<td>1.2</td>
<td>-0.03</td>
<td>-0.08</td>
<td><em>P &lt; .001</em></td>
<td>.69</td>
</tr>
</tbody>
</table>

1 All effect size comparisons are made with respect to Curtin I.
2 Statistically significant at the adjusted alpha level of .05/4=.013

Conclusions

The first important issue, whether the instrument “works” in Australia from a basic psychometric perspective, seems to have been resolved positively. From the data analysis, ASCIv2 functions in a similar way for Curtin University students as with US students. In other words, the 2-factor model produced a decent fit with the combined Curtin data, with each of the eight items loaded on its appropriate subscale. The internal consistency reliability as measured by Cronbach’s alpha was above .8 for both scales, which indicates the instrument meets acceptable standards for research (Murphy & Davidshofer, 2005).

The second question of interest, whether the student attitudes measured at Curtin are more or less positive than those currently in the literature, also had a definitive answer. Compared with US students from a large southeastern public research university at the 11th week of a semester of the first college chemistry class in the curriculum, Curtin students from October 2011 in a first year general (Class I) and organic and biological chemistry (Class II) unit of study think that chemistry is more accessible and more emotionally satisfying (by .51 standard deviation units), both medium effect sizes (Cohen, 1988). A medium effect size reflects a difference that would
be noticeable to a careful observer. For the two classes at Curtin, class I students feel chemistry is more accessible (by .47 standard deviation units, medium effect size) and more emotionally satisfying (by .36 standard deviation units) than class I students do.

Our desire to determine whether ASCIv2 could be useful in the Australian context, where increased student engagement is a key aspect of curriculum reform efforts, seems to have borne fruit. The instrument scores obtained from trialing the instrument with two classes of Curtin University students met some commonly accepted psychometric standards, and we believe it is likely that other institutions in Australia would also find the instrument easy to use and to interpret. What remains is to explain how the nature of the two Curtin classes may give rise to these observable differences to those in the US.

Acknowledgements

The Australian Government, through the Australian Learning and Teaching Government, funded “Developing leaders of change in the teaching of large university chemistry classes” (LE8-818). The support of the Australian Council of the Deans of Science, the Royal Australian Chemical Institute, and Curtin University is acknowledged. The authors thank the students of Curtin University for participating in this study by voluntarily completing surveys. Collection of data described in this experiment was authorised by the Human Research Ethics Committee at Curtin University, project number SMEC 107-11.

References


Bedgood, D. R. (2008). Why are we still teaching the way we were taught in the 1980s? *Chemistry in Australia, 75*(11), 22-23.


VI. Gathering Psychometric Evidence for ASCIv2 to Support Cross-Cultural Attitudinal Studies for College Chemistry Programs

Note to Reader

This chapter presents work of a book chapter of Affective Dimensions in Chemistry Education in press, coauthored with Drs. Alhooshani, Southam, and Lewis. This chapter has been reproduced with kind permission of Springer Science+Business Media and coauthors.

Introduction

The Need for Attitude

Student attitude toward science is a concern in many countries. For various reasons, such as public and disciplinary concerns about content knowledge and uncertain reliability of measurement for the attitude construct, education accountability policy usually does not consider a student’s attitude (The General Assembly of the State of South Carolina, 1998). Similarly, many countries are faced with the pressure to make educational policies focus on content knowledge in order to catch up with top-achieving countries as measured by the international assessments like TIMSS (Martin & Mullis, 2008). This has created a concern that some educators might not perceive the affective domain as important and may pay more attention to students’ content knowledge than to their affective growth, which may impact student learning in the long term. Because some also may perceive an attitude score to be subjective and unreliable, effort should be taken to measure and interpret data in a scientifically valid way.
When an attitude instrument is designed for international use, there is a validity concern that the students of various cultural backgrounds may interpret the items differently. The different interpretations of survey items can potentially bias the survey results if users do not take this validity concern into account. In their review, Osborne cited the studies by Taylor et al. (Taylor, 2003; Modood, 2003) and posited that “Asian students have a clear preference to study for degrees in medicine-related studies, engineering or mathematics … Afro-Caribbean students seem to shy away from science preferring to pursue degrees in the social sciences,” while “Japanese-Americans were most positively inclined towards scientific careers” (Osborne, Simon, & Collins, 2003). While these sorts of conclusions are intriguing, without examining the instrument function across a number of populations, we cannot be sure whether the observed difference is due to test bias or represents a real difference in the group means (Jiang, Xu, Garcia, & Lewis, 2010; Schroeder, Murphy, & Holme, 2012). Therefore, we need to gather sufficient evidence as to whether or not the instrument functions similarly to measure the same construct across different groups, so we can make a fair comparison using the instrument scores.

In this pilot study an attitude instrument was given to a student sample in Saudi Arabia, and compared to previous results from Australia and the USA (Xu & Lewis, 2011; Xu, Southam, & Lewis, 2012), to examine how an attitude instrument functions in context and how college students’ attitude status toward chemistry might differ between each group. Saudi Arabian students are of particular interest in the area of student attitude due to the unique features of the college chemistry programs as compared to other western educational systems such as Australia or the USA. For example, in Saudi Arabia, courses are mostly offered in classes segregated by sex, even in co-educational universities. The major choice is also limited to the programs that
these specialist universities can offer, compared to the comprehensive and mixed sex classes in both Australia and the USA.

Research on student attitude has been undertaken in Saudi Arabia with targets as diverse as computers, smoking, eating, and so on (Al-Khaldi & AlJabri, 1998). There is only one study concerning attitude performed in a school chemistry education context from Saudi Arabia (Harty & Alfaleh, 1983). The authors reported very promising results that high school students in a small-group laboratory setting performed better on a chemistry achievement test and exhibited more positive attitudes toward science than those in the traditional setting. Unfortunately, there are no accessible follow up publications in English on Saudi Arabian students’ attitudes in the area of chemistry, though there are some in Arabic (Albaz, 2007; Albusylee, Sadieg, & Abdukader, 1990; Balfagheh, 2001; Fatallah, 2009; Hijazi, 2008). Therefore, this field is in its infancy in Saudi Arabia and needs to be further explored. This study is intended to contribute to understanding how attitude instruments function for college students in a Saudi Arabian context to provide a preliminary attitude comparison with student samples in Australia and the US. The findings from this study can provide an example to establish evidence for a robust instrument; otherwise it is uncertain whether research findings are just an artifact of the instrumentation. With more knowledge of an instrument’s function, the chemistry education community can use the tool with more confidence to support further research investigations, such as how attitude relates to student achievement and the school curriculum.

*Measurement for Attitude*

With increased international collaboration in educational reform, there is a need to develop measurement tools with reported psychometric evidence for cross-cultural comparison (Abell, Springer, & Kamata, 2009; van de Vijver & Tanzer, 2004). Valid instrument scores are
critical to evaluate the effectiveness of educational programs, to inform instruction, and to help make data-driven educational policies. Psychometric evidence of an instrument’s function in the specific context is necessary, to interpret an instrument score, to support claims based on instrument scores, and to meet the current standards for educational and psychological testing (AERA, APA, & NCME, 1999). Accordingly, a conceptual framework for evidence sources was proposed in the area of chemical education (Arjoon, Xu, & Lewis, 2013). Validity refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (AERA et al., 1999). Reliability is necessary for validity and refers to the consistency of a measure and scoring procedures. Reliability evidence can be gathered based on temporal stability, which requires the same respondents to take the test more than once, and internal consistency, which involves examining the degree to which responses to related items correlate with one another. Validity evidence can be collected based on temporal stability, internal consistency, test content, response processes, internal structure, and relations to other variables (Arjoon et al., 2013).

Test content evidence is typically established by asking a panel of domain experts to judge whether the items appropriately sample the domain of interest. Cognitive interviews are often used for gathering response process evidence, providing insight into whether thought processes invoked by test items are those intended by the test developer. Constructed-response items can also be useful tools for examining this sort of validity evidence. Respondents need to first understand the nuances of the item and then mentally retrieve relevant information in order to make a decision about how to respond to the item; response process evidence demonstrates respondents’ understanding of an item by illuminating their thinking about that item. Relational validity evidence is typically inferred from statistical analysis, such as confirmatory factor
analysis and correlation analysis. The internal structure of an instrument, or how the items in the instrument relate to each other, is important because usually an instrument prescribes the intended construct as unidimensional or multidimensional, with specific item sets measuring different aspects of the construct in the latter case. Evidence based on internal structure establishes the degree to which the item scores for the instrument confirm to the hypothetical construct. Evidence based on relations to other variables concerns hypothesized relationships between the construct measured by the instrument and other variables within a specific theoretical framework. Accumulating this evidence requires information about the other variables of interest, gathered via additional tests or surveys of the respondents.

All the sources of evidence mentioned above provide support for instrument function from multiple perspectives. Gathering evidence, even for an existing instrument, is a long and iterative process, and should never be viewed as complete. Instead of developing a new instrument from scratch, it is desirable to use and evaluate an established instrument for respondents in different contexts. While many instruments relating to attitude toward chemistry are available, five have been specifically evaluated with respect to published validity evidence in the college chemistry context (Arjoon et al., 2013): the Cognitive Expectations for Learning Chemistry Survey (CHEMX) (Grove & Bretz, 2007), Colorado Learning Attitudes about Science Survey (CLASS) (Barbera, Adams, Wieman, & Perkins, 2008), Chemistry Self-concept Inventory (CSI) (Bauer, 2005), Attitude toward the Subject of Chemistry Inventory (ASCI) (Bauer, 2008) and its shortened version ASCIv2. Among these instruments, ASCIv2 has the advantage of clear connection with the attitude definition and framework in psychology (Rosenberg & Hovland, 1960), which is supported by empirical data from student samples at multiple sites (Xu & Lewis, 2011; Xu, Southam, & Lewis, 2012). The ASCIv2 retains eight
items from ASCI in two subscales, “intellectual accessibility” (items 1, 4, 5, and 10 from ASCI) about the difficulty of chemistry, and “emotional satisfaction” (items 7, 11, 14, and 17 from ASCI) about how satisfied students feel about chemistry in general, which are congruent with two components (cognitive and affective respectively) of attitude theory. This study focuses on quantitative analysis of internal structure, internal consistency and some qualitative evidence based on response processes for ASCIv2 in a new context, Saudi Arabia. These analyses were used to investigate the evidence that the ASCIv2 items are internally consistent and measuring the intended two attitudinal subscales for three student cohorts, thus it can provide a potential valid tool to support the cross-cultural attitudinal studies.

**Research Questions**

The goal of this study is to examine an attitude instrument, ASCIv2, which can be used to support cross-cultural studies for college chemistry programs. First, we are interested to see how ASCIv2 behaves for students with different backgrounds. We are also interested in examining the attitude profile across student groups.

The three specific research questions that guide the study are:

1. How did the attitude instrument function at a Saudi Arabian institution to measure the attitude construct, as compared to institutions in Australia and the US regarding the construct validity and internal consistency reliability?

2. If there are any items performing differently at the Saudi Arabian institution, how did students interpret the problematic items?

3. What is the attitude status for students enrolled in general chemistry courses for a Saudi Arabian institution as compared to an Australia and a US institution?
**Settings**

A university in the Kingdom of Saudi Arabia, or KU, participated in this study. KU data was compared with those from SE, a southeastern university in the USA, and WU in Western Australia. Note that the attitude scores are for that specific university and we cannot generalize from one university to the whole country. The detailed comparison of cultural, religious, and educational features in these three countries is beyond this report. Many factors can influence student attitude toward science including school, classroom and family levels (Papanastasiou, 2002; Papanastasiou & Papanastasiou, 2004). Here we provide some general information about these three universities that could potentially affect how students respond to an attitude instrument, from differences in attitudinal status. Four similarities connect these universities: they are large in size, public, research-oriented (with high research activity), and the language of instruction is English. However, great distinctions exist in the school organization and the role of the chemistry course. KU ranks among the top four universities in Saudi Arabia according to the Academic Ranking of World Universities (ARWU) (Shanghai Jiaotong University, 2012). With an acceptance rate of only 6.5% based on internal institutional data during the recent years, KU accepts among the best 2% of high school graduates especially those who intend to study engineering disciplines. WU, on the other hand, has a different profile. It ranks among the top 20 out of 39 tertiary institutions in Australia (Australian Education Network, 2013) and is the largest university in its region. It is an internationally-focused institution with 30-50% international students. In the USA, the Carnegie Foundation for the Advancement of Teaching classifies institution characteristics, including students’ prior academic preparation and selectivity of undergraduate admissions. SE is categorized as selective (Carnegie Foundation for the Advancement of Teaching, 2010). This category locates at roughly the middle two-fifths of baccalaureate institutions in the USA. As compared to the other two institutions, due to the
relatively high ranking and selectivity of KU within Saudi Arabia, its students likely have good academic records, which may be associated with more positive attitudes (e.g., Osborne, et al. 2003; Brandriet et al., 2012, Xu et al., 2013).

Additional distinctions between the institutions also are worth consideration. KU offers courses in the college of engineering, sciences, and industry management. The first major choice for most students is engineering. It is possible for students to change majors; however, few students do so. Social science or liberal arts programs are not available at the university, so there is not a diverse set of options. Since Saudi Arabian high school students are taught in different tracks (e.g., science/engineering vs. social science) starting in 11th grade, the KU students who enter the engineering major are mostly from the science/engineering track. Based on their high school experiences and institutional choice, entering KU students tend to have a good idea of chemistry and the importance of chemistry for their major when they take their first college chemistry course.

At WU, students are accepted for admission at university with a major; however, it is not unusual for students to consider changing their enrollment major or course of study. WU offers over 850 undergraduate and postgraduate courses in business, engineering, health sciences, humanities, science, mining and agriculture. Students can change a major or add a second major at the beginning of semester at any point during their enrollment. Portability of courses between majors is encouraged through an established credit transfer system and the commonality of chemistry as a first year course in many science-related majors means it is often taken by students in preparation for transfer to more competitive majors, such as Engineering or Pharmacy.

At SE, all students must officially declare a major or pre-major before they register for more than 36 credits, usually by the end of the second year. SE offers a balanced arts &
sciences/professions undergraduate instructional program. During the first two years, students usually take general college courses that are applicable to all majors as they develop their knowledge of and perceptions about various major programs. A college general chemistry course, for example, can serve as a general course for both science and non-science majors, meaning that students who may have been considering a science major can change their minds without penalty after taking chemistry. After the first two years, students enroll in the specific courses that apply toward completion of degree requirements for the chosen major. As compared to KU and WU, SE students may have a relatively vague idea about their major when taking their college general chemistry course and have more choices regarding major throughout the undergraduate period.

According to the university characteristics above, KU is the most selective among these universities, and KU students may understand best the importance of passing the college chemistry course for their degree, while SE is the least selective. We hypothesize that attitude follows the selectivity trend between KU and SE, with WU sitting in between. The participating universities also differ a great deal in the timing for students to declare a major and the options for changing majors, with KU students entering with significant experience in STEM-related courses and a relatively firm major choice, and SE students entering without much experience and with the option to delay major choice until after taking chemistry. Again, we hypothesize that attitude would follow trend regarding major, with KU most positive, SE least, and WU in between.

**Research Methods**

**Instrument**

The English version of ASCIv2, was used at the three universities (Bauer, 2008; Xu & Lewis, 2011). The instrument is intended to measure students’ attitude toward chemistry in
general in a 7-point semantic differential format; e.g., chemistry is easy vs. hard for item 1, comfortable vs. uncomfortable for item 4. It includes eight items which can be grouped in two subscales: intellectual accessibility (4 items) and emotional satisfaction (4 items). The entire instrument and instructions can fit on half a page, and it takes at most 5 min to administer. For a copy of the instrument, see the supplementary material (Xu & Lewis, 2011) or contact the corresponding author directly.

Participants and Data Collection

Detailed information about participants and data collection processes in Australia and the USA was provided in previous literature (Brandriet, Xu, Bretz, & Lewis, 2012; Xu & Lewis, 2011; Xu, Southam, & Lewis, 2012). The participants at KU are all male students because KU is a single-sex public university in Saudi Arabia. All KU participants were majoring in Engineering, in keeping with the specialist nature of this institution.

Data were collected from freshman students enrolled in a section of the first semester General Chemistry course at KU in Saudi Arabia in Feb 2011. This section is taught in a traditional lecture-based way. There are a total of 190 students in the class. One hundred and seventy students returned the survey with complete data for all eight items and are included in the data analysis. We did not see any suspicious patterns in the missing data that could potentially bias the research findings.

Data Analysis

First, basic descriptive statistics including mean, standard deviation, skewness and kurtosis for the eight items of ASCIv2 for KU students were obtained using SAS software version 9.1. In addition, a box-and-whisker plot was graphed for each item score, overlapped with the test items. A box-and-whisker plot is a convenient way to graphically indicate the
degree of spread for each item through the seven values: the lowest score, the lower quartile (the score at the low 25% rank in the sample), median (the score at the 50% rank), upper quartile (the score at the high 25% rank), and the highest score.

Next, evidence for reliability and validity of internal structure was examined. For reliability, Cronbach’s alpha estimates were calculated using SAS 9.1. A high Cronbach’s alpha suggests that the item scores are positively correlated with each other and with the total scale score as well. Cronbach’s alpha of at least 0.7 is generally desirable for research purposes (Murphy & Davidshofer, 2005).

For validity evidence based on internal structure, confirmatory factor analysis was performed in Mplus 5.2 to estimate how well the designed two-factor correlated structure for the instrument fit the responses obtained with the sample (Muthén & Muthén, 2007). Fit indices such as chi-square ($\chi^2$), Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR) were examined to assess the fitness of the model to the data, and item loadings were also evaluated. The criteria of CFI value greater than 0.95 and SRMR value less than 0.08 were used to indicate a good model fit, and CFI > 0.90 as acceptable fit (Bentler, 1990; Hu & Bentler, 1995).

When results raised concerns about items performing differently across the countries, we followed up by investigating how students interpreted the items. To achieve this, ASCIv2 was given to a group of students at KU who had previously completed the survey, to collect their written feedback about how they understood the problematic item. The feedback was independently coded to examine whether the item was interpreted as expected to indicate the intended subscale or not. The inter-rater reliability was calculated according to Cohen’s Kappa (Cohen, 1960). The widely used interpretation for the value of Kappa is applied to examine the
strength of agreement: ≤ 0.00 as poor, 0.00 - 0.20 slight, 0.21 - 0.40 as fair, 0.41 - 0.60 moderate, 0.61 - 0.80 as substantial, 0.81 - 1.00 almost perfect (Landis & Koch, 1977). Note that, since students could not be asked to provide identification information, we could not match their initial answer on ASCIv2 with their feedback.

For the student feedback that did not capture the aspect of attitude intended by the subscale, three coders performed another round of open-ended coding to examine if any pattern existed. A code is any tag or label that assigns a sense-making attribute for a portion of qualitative data (Miles & Huberman, 1994). Patterns in the codes were reported after the three coders reached consensus via discussion.

Finally, once validity and reliability evidence supported score interpretation, students’ attitude scores from KU were compared with scores from WU and SE students, using a standard effect size method (Cohen, 1988). Cohen's $d$ effect size can be used to quantify the difference in the attitude score. Values for Cohen's $d$ of 0.2 to 0.3 are generally considered a "small" effect, of around 0.5 a "medium" effect and of 0.8 or above, a "large" effect. A medium effect size reflects a difference that would be noticeable to a careful observer.

**Results**

*Descriptive Statistics for Item Scores for KU*

Descriptive statistics are shown in Table 6.1 for each item (four reverse ordered items were reverse coded for interpretation). High scores mean that students feel chemistry is intellectually accessible and emotionally satisfying. The average scores range from 3.37 to 4.54, and standard deviations range from 1.39 to 1.92. No item was found to have skewness or kurtosis greater than 1.15, which suggests good normality of the item scores. From the items (6 & 8) with extreme scores in Table 6.1, students feel chemistry is organized, and is challenging. This pattern
is very similar with that observed for data from WU and SE (Brandriet et al., 2012; Xu & Lewis, 2011; Xu et al., 2012).

Table 6.1

Descriptive Statistics of Item Scores for KU

<table>
<thead>
<tr>
<th>Item</th>
<th>Mean</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>1* hard</td>
<td>4.24</td>
<td>1.41</td>
<td>-0.24</td>
<td>-0.49</td>
</tr>
<tr>
<td>2 complicated</td>
<td>3.89</td>
<td>1.39</td>
<td>-0.13</td>
<td>-0.65</td>
</tr>
<tr>
<td>3 confusing</td>
<td>3.96</td>
<td>1.61</td>
<td>0.16</td>
<td>-1.05</td>
</tr>
<tr>
<td>4* uncomfortable</td>
<td>4.09</td>
<td>1.54</td>
<td>-0.11</td>
<td>-0.53</td>
</tr>
<tr>
<td>5* frustrating</td>
<td>4.41</td>
<td>1.57</td>
<td>-0.37</td>
<td>-0.56</td>
</tr>
<tr>
<td>6 challenging</td>
<td>4.37</td>
<td>1.80</td>
<td>0.44</td>
<td>-0.81</td>
</tr>
<tr>
<td>7* unpleasant</td>
<td>4.08</td>
<td>1.92</td>
<td>-0.09</td>
<td>-1.15</td>
</tr>
<tr>
<td>8 chaotic</td>
<td>4.54</td>
<td>1.67</td>
<td>-0.35</td>
<td>-0.81</td>
</tr>
</tbody>
</table>

Note: Items with * are reverse coded but are shown with the word pairs also reversed for clarity of interpretation. The extreme values are in bold.

The box-and-whisker plot for each survey item for KU students is presented in Figure 6.1. The star near the middle of box represents the mean score for each item, ranging from 3.37 for item 6 to 4.54 for item 8. The line near the middle of the box represents the median score for each item. Five items have a median at the middle point of 4, while items 5 and 8 have the median toward the positive side, with item 6 toward the negative side. The left and right of the box represent the lower and upper quartiles of the item score. For most items, students tend to pick between 3 and 5. The ends of the whiskers represent the minimum and maximum of all responses for each item. Note that the students used the full range; the whiskers extend from 1 to 7. From Figure 6.1, most students tend to pick up the neutral attitude around 4, and there is a lot of overlap for the item scores.

Because the instrument has the internal structure of two subscales, we proceeded to examine the evidence based on the intended test design and interpret on the subscale level rather than on the item level.
Figure 6.1 Box-and-whisker plot for KU students overlapped with ASCIv2 items. Note that the item pairs have been reversed for items 1, 4, 5, and 7 for ease of interpretation.

Two-Factor CFA Model Fit

CFA was performed to estimate an appropriate fit for the 2-factor model. Items 1, 2, 3 and 6 were set to load on the factor “intellectual accessibility” only; items 4, 5, 7 and 8 were set to load on the factor “emotional satisfaction” only, and the two factors were allowed to correlate. Models were identified by fixing the first item on its factor at 1.

The estimation of the 2-factor model fit for KU is: \( \chi^2 (n = 170, df = 19, p < .001) = 75, \) CFI = .88, SRMR = .07. Since CFI was less than .90, we can consider that the overall model does not fit the data well for the KU students. By comparison, the 2-factor model fit the data from WU and SE relatively well (Xu et al., 2012).

Item loading was examined to identify the source of the misfit for the KU data. As shown on Table 6.2, item 6 (regarding whether chemistry is challenging/not challenging) is negatively related to other items in that scale of intellectual accessibility and the correlation coefficient is not significantly different from zero. This means that, KU students did not consider item 6 as an indicator of this proposed factor (intellectual accessibility), and the measurement of this scale...
should be reconstructed. The other scale of emotional satisfaction works as intended. For WU and SE data, all items loaded on the intended subscale well. The item-total correlation was examined with item 6 (challenging vs not challenging) found to have an extremely low value of -0.1. Again the observation suggests that KU responses to item 6 are not correlated to three other items (1, 2, and 3) for the “intellectual accessibility” subscale. One possible reason is that KU students might have a different understanding of this word pair of challenging/not challenging from students at WU and SE.

Table 6.2

*CFA Item Loadings for the 2-Factor Solution and Item-Total Correlation within Each Intended Subscale*

<table>
<thead>
<tr>
<th>Factor</th>
<th>Item #</th>
<th>Item Loading</th>
<th>Item-total Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Accessibility</td>
<td>Item 1</td>
<td>0.76</td>
<td>.49</td>
</tr>
<tr>
<td></td>
<td>Item 2</td>
<td>0.80</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>Item 3</td>
<td>0.66</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>Item 6</td>
<td>-0.11 (n.s.)</td>
<td>-.04</td>
</tr>
<tr>
<td>Emotional Satisfaction</td>
<td>Item 4</td>
<td>0.75</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>Item 5</td>
<td>0.71</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td>Item 7</td>
<td>0.74</td>
<td>.61</td>
</tr>
<tr>
<td></td>
<td>Item 8</td>
<td>0.52</td>
<td>.45</td>
</tr>
</tbody>
</table>

Note: All loadings are significantly different from 0 at the p=.05 level except those labeled (n.s.)

*Internal Consistency Reliability for ASCIv2*

The internal consistencies were calculated for each subscale. Cronbach's alpha for the KU data is only 0.56 for the intellectual accessibility scale, which is quite a bit lower than the satisfactory level of 0.7. The Cronbach's alpha increases to 0.78 if item 6 is deleted from this scale. Cronbach's alpha for both subscales for WU and SE are above the rule-of-thumb satisfactory level of 0.7.
Students’ Interpretation of Item 6

Based on the CFA and Cronbach's alpha, KU students responded to item 6, challenging/not challenging, differently from their counterparts and were not consistent with the intended internal structure. It is critical to examine the cognitive process of how the respondents respond to the survey item, and the first task is to explore how respondents interpret the survey item (Schwarz, 1999). Specifically, we need to know whether students understand and respond to item 6 as an indicator of chemistry attitude on the intellectual accessibility subscale as expected.

Fifty KU students who took the ASCIv2 were willing to provide anonymous written responses to the second author regarding their interpretation of the meaning of item 6. The first author and a chemistry graduate student with qualitative coding experience independently coded all the feedback. If the student interpreted item 6 as expected for the intellectual accessibility scale, the feedback was coded as 1. If not, the code was 0. From Table 6.3, both raters agreed on 46 out of the 50 total codes, with percent agreement at 92%. Inter-rater reliability calculated by Cohen’s Kappa is .70, which is considered substantial strength of agreement (Landis & Koch, 1977).

Table 6.3

<table>
<thead>
<tr>
<th>Rater 1</th>
<th>A good indicator or not</th>
<th>Yes</th>
<th>No</th>
<th>Total for rater2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rater 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yes</td>
<td>40</td>
<td>1</td>
<td>41</td>
</tr>
<tr>
<td></td>
<td>No</td>
<td>3</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Total for rater1</td>
<td></td>
<td>43</td>
<td>7</td>
<td>50</td>
</tr>
</tbody>
</table>
Both raters believed that a large portion of students (40/50) interpreted item 6 to involve considerations of whether chemistry is difficult or not, which does fit the intended scale of intellectual accessibility. Both raters also agreed that six students did not interpret the item as a good indicator of the intended subscale, but they disagreed on the other four students’ feedback. Overall, then, 10 students explicitly interpreted this item in terms that seemed outside intellectual accessibility to at least one rater.

Three chemistry graduate students engaged in a consensus coding process to identify the common characteristics in the 10 student interpretations that were not in good alignment with the intended subscale. Out of the 10 interpretations, one student interpreted the meaning of challenging as including strong negative emotional arousal. He wrote, “Challenging for me it means challenging (frustrating) in negative side of it, and usually frustrating does not lead to a positive result as far as what sort of despair and melancholy in the soul.” This means he treats this item as more on the emotional satisfaction rather than on the intellectual accessibility subscale. Five student responses were in alignment with neither intended subscale. Three of these students interpreted item 6 as chemistry relating to daily life, for example, “Chemistry helps us to understand how the things around us work.” The other two students viewed the item as indicating the role of chemistry in discovery, e.g., “Challenge in chemistry appears to me on how new discovery can help, improve and solve of today life.” These interpretations are both off target for intellectual accessibility.

The other four students mentioned multiple elements in their responses, such as the effort needed to learn chemistry, the lab work, competition for a high score, the learning strategy, the grading system and so on. For example, one student wrote, “It let you try to do your best to get the best result [e]specially in the lab you try to get the perfect result however we can’t but we try
our best.” These interpretations are loosely connected to intellectual accessibility, but they also involve other ideas not intended by the subscale so are too broad to really be on target.

Recently, we also have begun to collect cognitive interview data regarding how students at WU interpret and respond to ASCIv2 items. As for item 6, WU students have reported thinking about whether chemistry-related topics are difficult or easy, which is consistent with the intended construct of intellectual accessibility. While a few of the WU students also allude to emotions, in every case the emotion associated with “challenging” is positive, unlike the KU student’s negative emotional connotation. For example, one student explicitly said: “I do not mean challenging in terms of overly difficult. I mean challenging in terms of it’s something I enjoy getting involved in.” These sorts of insights into how the items can be interpreted differently highlight the need for gathering evidence based on response processes, either via interviews or constructed response prompts.

**Attitudinal Profile**

Due to the problem of the unintended interpretation of item 6 for KU students, a slightly revised test structure of two factors measured by seven items without item 6 was considered for the purpose of interpretation. CFA was performed and Cronbach’s alpha determined for each data set without item 6 to support this new score interpretation approach. As shown in Table 6.4, the model fits for the three data sets are all tenable without item 6. The model fit for KU improved when item 6 was removed. For the WU and the SE data, the model fit without item 6 remains good based on accepted fit criteria. For KU, the new Cronbach’s alpha is .78 for the scale of intellectual accessibility, and .77 for the scale of emotional satisfaction. The final Cronbach’s alphas are .86 and .84 for the WU data, and .80 and .79 for the SE data. All values are above the satisfactory level of .7.
Table 6.4

*CFA Model Fit for the 2-Factor Solution without Item 6*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>KU</th>
<th>WU</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$</td>
<td>56.7</td>
<td>21</td>
<td>56</td>
</tr>
<tr>
<td>$p$-value</td>
<td>&lt; .001</td>
<td>.06</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>$df$</td>
<td>13</td>
<td>13</td>
<td>13</td>
</tr>
<tr>
<td>CFI</td>
<td>.90</td>
<td>.98</td>
<td>.96</td>
</tr>
<tr>
<td>SRMR</td>
<td>.05</td>
<td>.03</td>
<td>.04</td>
</tr>
</tbody>
</table>

Table 6.5

*Factor Scores for Each Subscale for KU, WU and SE*

<table>
<thead>
<tr>
<th>School</th>
<th>Subscale</th>
<th>$\alpha$</th>
<th>Mean</th>
<th>SD</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>KU (n=170)</td>
<td>Accessibility</td>
<td>.78</td>
<td>4.03</td>
<td>1.22</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>.77</td>
<td>4.28</td>
<td>1.29</td>
<td>-</td>
</tr>
<tr>
<td>WU (n=108)</td>
<td>Accessibility</td>
<td>.86</td>
<td>3.56</td>
<td>1.20</td>
<td>0.39 (medium)</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>.84</td>
<td>4.25</td>
<td>1.22</td>
<td>0.02 (no/trivial)</td>
</tr>
<tr>
<td>SE (n=354)</td>
<td>Accessibility</td>
<td>.80</td>
<td>3.04</td>
<td>1.16</td>
<td>0.84 (large)</td>
</tr>
<tr>
<td></td>
<td>Satisfaction</td>
<td>.79</td>
<td>3.63</td>
<td>1.20</td>
<td>0.53 (medium)</td>
</tr>
</tbody>
</table>

Note: All effect size comparisons are made with respect to KU.

We proceeded to calculate and compare the composite scores without item 6 for the two subscales for KU, WU, and SE. The intellectual accessibility composite score was obtained by averaging the three items 1, 2, and 3. The intellectual satisfaction composite score was the average of four items: 4, 5, 7, and 8. Cohen's $d$ effect size is presented to quantify the difference in attitude between SE students and KU students (Table 6.5). Compared with the SE students, KU students think that chemistry is more intellectually accessible with the difference of .84 SD (a large effect size) and more emotionally satisfying with .53 SD (a medium effect size). The KU
students feel chemistry is more intellectually accessible than WU students with a medium effect size, and feel similarly on the scale of emotional satisfaction as WU.

**Discussion and Conclusions**

This study reported results using the ASCIv2 attitude instrument for college students from three universities in Saudi Arabia, Australia, and the USA, and successfully answered the research questions. First, based on the evidence analysis using Cronbach’s alpha, confirmatory factor analysis, and student feedback about their response processes, the instrument of ASCIv2 measures attitude toward chemistry in a similar way across student groups with the exception of item 6. Care should be taken with using the same instrument in multiple countries, particularly when considering cultural interpretation of items that can influence validity. This study presents a case that a group of Saudi Arabian students, with a different background than two other groups of students, from Australia and the USA, respond to the survey items in a different pattern, even when the instrument language is the same.

Secondly, from the written feedback for KU students, we found 10 students interpreted item 6 (challenging vs. not challenging) not as intended. Rather than considering the item as measuring intellectual accessibility, students aligned it with other unintended constructs. It is not clear what cued Saudi Arabian students to relate this item to daily life, discovery, strong negative emotions, or learning behaviors. Interestingly, four students from Australia respond to this item with positive emotion. Cognitive interviews or constructed responses from a larger student sample can help better understand the way student groups interpreted item 6 as related to their cultural background. Since the data in this study is all from KU males, further work with female students from Saudi Arabia is warranted. In addition, since the students in this study cannot be
identified to track them, we do not know whether the 10 students with unintended interpretations was a large enough sample to influence the validity of score interpretation.

Last but not least, based on the composite score without item 6 for the intellectual accessibility subscale, we found that students enrolled in a general chemistry course at a US university exhibited a more negative attitude toward chemistry than their counterparts in Saudi Arabia and Australia. Effect sizes were large enough that the difference can be considered meaningful. This order in attitude is reasonable for these three institutions. SE ranks in the middle two-fifths of baccalaureate institutions in the US and is not highly selective. In addition, SE students are not required to decide their major until after they take general chemistry and may not be prepared for nor committed to a STEM major. KU is a top engineering university in Saudi Arabia, very selective, and the students have already committed to STEM degrees before entering the university. Although it would not be appropriate to draw conclusions about national differences based on only one institution from each country, a similar order in attitude has been observed for middle school students: Saudi Arabian middle school students in general exhibit more positive attitudes toward science as measured by the index of Positive Affect Toward Science (PATS) on TIMSS as compared to Australian and US students (Martin & Mullis, 2008).

The need to foster positive attitudes toward chemistry, which is one central goal of science literacy as postulated by American Association for the Advancement of Science (1989), and to increase students’ interest in the pursuit of STEM-related careers, may inspire instructional interventions at SE and WA that may not be necessary for KU.

The use of a convenience sample from one university in each country limits generalizability for this study; however, the study represents an important initial step. As discussed, the Saudi Arabian university in this study is highly competitive and selective, and we
may find that students at a less selective university are more similar to WU and SE students with respect to attitude, or that KU students would be more similar to students at more selective universities in Australia and the US. This study does show that ASCIv2 can be used to obtain interpretable attitude scores from students in multiple countries, and that it can discriminate attitudinal differences.

As ASCIv2 becomes established as appropriate for cross-cultural use in studies such as these, we hope that this initial study has drawn attention to issues of item interpretation, and we believe that the principles and procedures described here can be useful to others engaged in similar work. For example, further psychometric evidence can be gathered for the international use of ASCIv2 in additional cultural contexts. If the problem for any instrument, such as the one with item 6, persists in multiple contexts, the problem can potentially be addressed by deleting the item from the instrument or by exploring alternative word pairs, with careful attention to the impact on validity of using the item to measure the intended construct.

Moreover, this study signifies the importance of examining psychometric evidence in context when an attitude instrument is used in cross-cultural scenarios, to ensure that responses in a new context still match the trait underlying the instrument and the comparison based on instrument scores is fair and sound. The administration of ASCIv2 in three countries makes us aware that, for validity in cross-cultural studies based on an instrument, the population may not share a similar background that cues participants to interpret items similarly, even when the language is the same. When instruments are translated and adapted to different languages, language factors alone can impact the test function and should be carefully examined (Allalouf, 2003; Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2012). With such efforts and
evidence, we can have more confidence in conclusions of attitudinal differences by sex and/or ethnicity as posited by Osborne.

Once a robust instrument is chosen which works equally well for multiple groups with different backgrounds in the study setting, the next step is to use pre-post designs to track student attitude change in order to make fair cross-cultural attitudinal comparisons for college chemistry programs. Even with a robust instrument, we still must avoid the danger of determining program quality across countries based on one administration, because classes can be quite different. Other important variables, such as students’ academic performance, class organizations, school environment, school curriculum, and institutional selectivity, need to be tracked and considered as potential confounds. Over time, with the accumulation of psychometric evidence for the cross-cultural use of attitude instruments, we can have more confidence in the understanding of attitude status and its relationship to other variables such as academic performance and curriculum. Accordingly, more effective curriculum innovations can be chosen based on evidence, and then implemented to foster students’ positive attitude while improving content learning.

References


VII. College Students’ Attitudes toward Chemistry, Conceptual Knowledge and Achievement: Structural Equation Model Analysis

Note to Reader

This chapter is a published paper coauthored with Villafane and Lewis in the Chemistry Education Research and Practice, 14(2): 188-200, doi: 10.1039/C3RP20170H, and have been reproduced with permission from the Royal Society of Chemistry and coauthors. The paper can be accessed via http://pubs.rsc.org/en/content/articlelanding/2013/RP/c3rp20170h#!divAbstract.

Introduction

Introductory college-level chemistry courses are required for students to continue to advanced chemistry courses and to progress toward science-related careers. Therefore, student performance in introductory college chemistry courses remains a concern for educators. In order to guide decisions about which interventions could be effective in improving student achievement, the first step is to understand the factors that influence achievement. Many factors, such as student background and school environment, have been reported to influence students’ performance in science (e.g. (Bunce & Hutchinson, 1993; Cakici, Aricak, & Ilgaz, 2011; Seery, 2009)). In chemical education, research has been done to predict students’ achievement in college chemistry courses. Mathematical ability has been important for students’ placement into specific college chemistry courses as well as for prediction of their chemistry achievement (Cooper & Pearson, 2012; Pienta, 2003; Scott, 2012). Prior conceptual knowledge has also been
found to have an effect on student subsequent achievement (Seery, 2009; Wagner, Sasser, & DiBiase, 2002). Besides those cognitive factors, non-cognitive factors, such as attitude, are extensively studied in the chemistry education area as a potential influence on students’ achievement (Bauer, 2008; Cukrowska, Staskun, & Schoeman, 1999; House, 1995; Xu & Lewis, 2011).

In this study, we are interested in examining the relationships among three key factors, attitude toward chemistry, prior conceptual knowledge, and math ability, and their relationship with chemistry performance for college students enrolled in the first semester of introductory college chemistry. These three factors have been shown to be good predictors of achievement in previous studies of similar contexts (e.g., (Gooding, Swift, Schell, Swift, & McCroskery, 1990; House, 1995; Pienta, 2003; Xu & Lewis, 2011)). In addition, these factors are of practical importance for educational practice. Effective instructional strategies can be implemented in science classrooms to influence these factors. For example, Freedman (1997) used a hands-on laboratory program to foster students’ positive attitudes toward science, while Abdullah, Mohamed, and Ismail (2009) used individualized microscale experiments in chemistry to improve students’ attitudes toward chemistry practical work. Ebenezer, Chacko, Kaya, Koya, and Ebenezer (2010) implemented interventions to promote positive changes in students’ conceptual knowledge; in the same way, Regan, Childs, and Hayes (2011) used students’ misconceptions in chemistry to develop a course of tutorials to target those difficult areas. Finally, Pienta (2003) developed a mathematical assessment and tutorial for general chemistry students to help them with math knowledge and skills. Since different interventions have been used to improve each of these factors separately, it is reasonable to wonder whether attention to all of these factors would be likely to improve student chemistry achievement to a greater degree,
and whether there are relationships between the factors that hint at the potential for synergy. This study examines these issues for a group of students enrolled in a typical first semester general chemistry course in the United States via a structural equation modeling (SEM) approach. The first step was choosing how to measure each of the relevant factors.

**Student achievement in college chemistry**

Course grades or final exam scores are often used to measure student achievement in the first semester general chemistry course. In the United States, course grade can be a relevant criterion for investigating student progress toward eventual careers in chemistry, because students often need a grade of C or better in the introductory course to be able to register in more advanced courses. Alternatively, more general measurement tools, especially national exams such as the chemistry exams prepared by the Examinations Institute of the American Chemical Society’s (ACS) Division of Chemical Education, can be used as a measure of student achievement (Examinations Institute, 1997). Some ACS exams have explicitly blended both traditional (algorithmic) and conceptual questions to measure students’ performance in concepts covered during college chemistry courses (Examinations Institute, 1997; Lewis & Lewis, 2008). The main advantage of using an ACS exam score over course grade, as a measure of performance or achievement in chemistry, is that the results are more readily generalizable across institutions (Examinations Institute, 1997; Lewis & Lewis, 2007).

**Attitude toward chemistry and relationship with achievement**

Contemporary theories of attitude define it as “a psychological tendency that is expressed by evaluating a particular entity with some degree of favor or disfavor” (Eagly & Chaiken, 1993, p.1). The tripartite theoretical model consisting of three attitudinal dimensions, affect, behavior, and cognition was proposed by Rosenberg and Hovland (1960, p. 3). The affective component
relates to an emotional response that expresses an individual's level of preference with respect to a particular entity, such as the subject of chemistry. The cognitive component is an evaluation that comprises an individual's beliefs and knowledge about the entity. The behavioral component corresponds to a verbal intention or behavioral tendency of an individual to act towards that entity. Attitude is not resistant to change: an individual gains most of his or her attitudes from either direct experience or observational learning from the environment (Maio & Haddock, 2010).

There are many different methods for measuring attitude. Standard attitude measures include scales such as Likert, Guttman, Thurston, and semantic differentials. In chemistry education, several extant attitude instruments are the Chemistry Expectations Survey (CHEMX) (Grove & Bretz, 2007), Colorado Learning Attitudes about Science Survey (CLASS) (Barbera, Adams, Wieman, & Perkins, 2008; Heredia & Lewis, 2012), Attitude toward the Subject of Chemistry Inventory (ASCI) (Bauer, 2008; Brandriet, Xu, Bretz, & Lewis, 2011; Xu & Lewis, 2011), Chemistry Attitudes and Experiences Questionnaire (CAEQ) (Coll, Dalgety, & Salter, 2002), and a questionnaire to measure attitude toward chemistry learning (Berg, 2005). In its shortened version, the Attitude toward the Subject of Chemistry Inventory version 2, ASCIv2, has the most reported evidence for psychometric quality and reflects two components in the tripartite framework, cognition and affect (Brandriet et al., 2011; Xu & Lewis, 2011).

There have been many attitudinal studies where the relationship between attitude toward science and science achievement has been investigated (Cukrowska et al., 1999; House, 1995; Weinburgh, 1995). With high school students, students’ achievement has been found to be significantly related to their attitudes (Gooding et al., 1990; Tuan, Chin, & Shieh, 2005). In college chemistry, positive relationships have been found for first-year chemistry using
correlation analysis (Cukrowska et al., 1999) and logistic regression analysis (House, 1995). However, these positive relationships have been found to depend on the instrument used to measure students’ attitude, making instrument choice an important issue (Xu & Lewis, 2011). Not only does ASCIv2’s psychometric information make it a reasonable choice, Xu and Lewis (2011), using multiple regression, found that one of the components of attitude measured by ASCIv2 plays a significant and unique role in predicting final achievement in a first semester general chemistry course.

**Prior conceptual knowledge and relationship with achievement**

Students’ prior conceptual knowledge is another factor that has been considered critical for learning chemistry in a meaningful way. Students’ incorrect ideas or alternate conceptions about important topics could hinder deeper learning of those and other relevant topics in chemistry, thereby affecting students’ final achievement in the course (Gabel, 1999; Nakhleh, 1992; Nicoll, 2001). For that reason, investigating the relationship between students’ prior knowledge and achievement is important. A recent study reported that prior knowledge is the most significant factor in predicting achievement for undergraduate chemistry (Seery, 2009). However, as Dochy (1999) pointed out, the effect of prior knowledge depends on the assessment method used.

Multiple instruments have been developed with a focus on chemistry conceptual knowledge rather than rote memorization, including the Chemistry Concept Understanding Test (CCUT) (Danipog & Ferido, 2011), Student Conceptual Understanding Test (SCUT) (Lin, Lee, & Treagust, 2005), Chemical Concept Inventory (CCI) (Mulford & Robinson, 2002), the Particulate Nature of Matter and Chemical Bonding Diagnostic Instrument (Heredia, Xu, & Lewis, 2012; Othman, Treagust, & Chandrasegaran, 2008), and the Targeted Misconception
Inventory (TMI) (Heredia, 2012). All of these instruments have been designed and used to probe students’ prior knowledge in chemistry and to identify students’ incorrect ideas, but none appear to have been used to predict achievement. Additionally, publicly available psychometric evidence is limited for many of these instruments. Because the TMI has been designed with a factor structure and some psychometric evidence can be obtained, it is a good candidate for measuring students’ prior conceptual knowledge and for investigating the relationship between prior knowledge and students’ achievement.

**Math ability and relationship with achievement**

Math ability has long been considered important for achievement in chemistry (Spencer, 1996). The SAT is the most widely used college admission test in the U.S., and it has been successfully used as a predictor of science performance in college (Ewing, Huff, Andrews, & King, 2005). The SAT as a whole probes students’ aptitude in different subject areas such as reading, writing, and math, with student math ability measured by the quantitative portion of the SAT, SATM. The relationship between SATM and students’ performance in college chemistry has been well documented in the research literature (Bunce & Hutchinson, 1993; Lewis & Lewis, 2007, 2008; Spencer, 1996; Wagner et al., 2002). In one study, Spencer (1996) found that SATM scores were an important factor in determining students’ grades in college-level general chemistry. In another study focused on college-level general chemistry, Lewis and Lewis (2007) found a medium-size correlation between SATM scores and ACS final exam score. While other measures of math ability exist (Cooper & Pearson, 2012; Pienta, 2003; Wagner et al., 2002), SATM has been found to be a suitable measure of students’ math ability and a good predictor of student achievement in introductory college chemistry.
**Conceptual relationship among the four factors in chemistry**

As the foregoing discussion indicates, these three factors and their relationship to achievement in chemistry have been examined previously, but typically in separate studies. The proposed simultaneous relationship among the different factors for the present study is shown in Figure 7.1. Students’ performance in chemistry, as measured by students’ achievement in the course, is directly influenced by three predictors: students’ attitude toward chemistry, their prior conceptual knowledge and their math ability. Relationships among these variables are expected; therefore they are set to correlate with one another in this model as shown by the double arrows in Figure 7.1.

![Figure 7.1 Conceptual Model of Chemistry Achievement and Three Predictors (Model 1)](image)

**Structural Equation Modeling (SEM)**

Structural equation modeling (SEM) is a multivariate data analysis approach used to study complex relationships among variables. The typical research questions that can be answered using SEM are twofold: 1) How do multiple variables interact with one another?, and 2) Which of several hypothetical models best fit empirical data?

SEM has been widely used in multiple disciplines such as economics (Goerzen & Beamish, 2005; Kaynak, 2003) and psychology (MacCallum & Austin, 2000; Wei, Heppner, & Mallinckrodt, 2003). In science education, it has been employed to investigate the learning process and to improve instruction (Abbott & Berninger, 1993; Cakici et al., 2011; Hemmings, 2011).
Grootenboer, & Kay, 2011; Schreiber, Nora, Stage, Barlow, & King, 2006). Recently in chemistry education, Stamovlasis, Tsitsipis, and Papageorgiou (2012) reported the relationships among ninth-grade junior high school Greek students' understanding of the change of state of matter and three cognitive variables: logical thinking, field-dependence/field-independence and convergence/divergence, with the final model accounting for 85% of the variance after structural understanding had been included among the predictors. In another study, Merchant et al. (2012) investigated the relationships among variables such as spatial ability, self-efficacy, and chemistry learning achievement for college chemistry students, with the model accounting for 45% of variance in performance as measured by a chemistry test. Kim and Song (2010) examined the effects of Korean high school students’ dichotomous (intrinsic and extrinsic) attitudes towards science, interest, and conceptual understanding on school achievement with a model that explained 33% of the variance in achievement. Turning the focus to conceptual understanding, Nieswandt (2007) explored the relationships between affective variables, including attitudes toward chemistry and chemistry self-concept, and student understanding of chemistry concepts for ninth grade students in Germany and found that students’ self-concept has a positive and significant effect on conceptual understanding.

Schreiber (2006) and Schreiber et al. (2008) has recommended a set of guidelines for the use, report, and evaluation of SEM studies in social and educational areas. Standard nontechnical issues for research, such as the fact that research questions should warrant the use of the chosen analysis technique, whatever it may be, also apply to SEM studies, but using SEM adds some technical requirements as well. For example, discussion of sample size, missing data, fit indices, parameter estimates, effect sizes, explained variance, rationale for model modification, and the use of nested models are important for SEM studies (Kline, 2010; Schreiber, 2008;
Schreiber et al., 2006). The recent SEM publications in chemistry education cited above abide well by these guidelines with respect to reporting fit indices and parameter estimates; however, there are some limitations in other areas. First, one study does not report explained variance (Nieswandt, 2007). Second, two studies were performed with very small sample size (Kim & Song, 2010; Nieswandt, 2007), which suggests that replication is warranted. For example, Kim and Song (2010) found no relationship between topic attitude and school achievement for 10th-graders (n = 109) and a negative relationship for 11th-graders (n = 110), which is very interesting, but needs to be verified in a larger population. Third, although all of these studies examined multiple models, they neither report effect sizes for the reported relationships, nor investigate nested models, both of which would enable more direct comparisons between models. In general, although these four studies represent a promising start, there remains a scarcity of SEM analysis in chemistry education research.

**Present Study**

The goal of the present study is to examine the relationship of potentially relevant predictor variables such as students’ attitude toward chemistry, students’ prior conceptual knowledge of chemistry, and students’ math ability to students’ achievement in chemistry, while accounting for relationships among the predictor variables. Specifically, the goal is to use Structural Equation Modeling (SEM) to assess the hypothetical relationships among variables as shown in Figure 7.1.

The four specific hypotheses for this study are:

1) The full SEM model with all three predictors will fit the data better than a model without the inclusion of prior conceptual knowledge as a predictor.
2) The proportion of variance in student achievement explained by the model will significantly decrease if measurement error is neglected or if prior conceptual knowledge is deleted as a predictor.

3) Attitude and prior conceptual knowledge have direct effects on chemistry achievement when math ability is controlled.

4) There is a significant correlation among attitude, math ability, and prior conceptual knowledge.

Method

Instruments

Attitude. Students’ attitude toward chemistry was measured by the Attitude toward the Subject of Chemistry Inventory version 2 (ASCIv2 hereafter) (Xu & Lewis, 2011). The instrument comprises two scales: “Intellectual Accessibility”, depicting the cognitive dimension of attitude and “Emotional Satisfaction”, depicting the affective dimension. The instrument consists of eight items, four items in each scale as shown in Table 7.1. The scale of the instrument is a semantic differential ranging from 1-7.

Table 7.1

Variables in ASCIv2 Instrument

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Variable</th>
<th>Item #</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intellectual Accessibility</td>
<td>ASCI1</td>
<td>1</td>
<td>easy</td>
</tr>
<tr>
<td>(cognitive dimension)</td>
<td>ASCI2</td>
<td>2</td>
<td>complicated</td>
</tr>
<tr>
<td></td>
<td>ASCI3</td>
<td>3</td>
<td>confusing</td>
</tr>
<tr>
<td></td>
<td>ASCI6</td>
<td>6</td>
<td>challenging</td>
</tr>
<tr>
<td>Emotional Satisfaction</td>
<td>ASCI4</td>
<td>4</td>
<td>comfortable</td>
</tr>
<tr>
<td>(affective dimension)</td>
<td>ASCI5</td>
<td>5</td>
<td>satisfying</td>
</tr>
<tr>
<td></td>
<td>ASCI7</td>
<td>7</td>
<td>pleasant</td>
</tr>
<tr>
<td></td>
<td>ASCI8</td>
<td>8</td>
<td>chaotic</td>
</tr>
</tbody>
</table>

Note: Items are shown in the orientation used in the instrument, Items 1, 4, 5, 7 are reverse scored before analysis.
**Prior conceptual knowledge.** Students’ prior conceptual knowledge in chemistry was assessed by the Targeted Misconception Inventory (TMI) (Heredia, 2012). The TMI is a two-tier multiple-choice instrument that targets three specific concepts from general chemistry: bond energy, phase changes, and ionic bonding. These basic concepts should be covered in previous chemistry courses such as high school chemistry or preparatory college chemistry and are further built upon in college general chemistry. TMI is designed to identify students’ difficulty or incorrect ideas with these basic concepts at the beginning of the first semester of college general chemistry. The test consists of 18 multiple-choice questions, which corresponds to 9 items, with each two-tiered question counted as one item. Students need to answer both tiers correctly to get 1 point for an item. Each concept in the TMI is represented by three items following a parallel structure as shown in Table 7.2. The possible score ranges from 0 to 9 points. Although TMI covers three basic concepts in general chemistry, it is limited in scope.

**Table 7.2**

*Items in TMI Instrument*

<table>
<thead>
<tr>
<th>Concept</th>
<th>Variable</th>
<th>Item #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phase Changes</td>
<td>TMI1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>TMI5</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>TMI7</td>
<td>7</td>
</tr>
<tr>
<td>Bond Energy</td>
<td>TMI2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>TMI4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>TMI9</td>
<td>9</td>
</tr>
<tr>
<td>Ionic Bonding</td>
<td>TMI3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>TMI6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>TMI8</td>
<td>8</td>
</tr>
</tbody>
</table>

**Math ability.** Students’ math ability was measured using the quantitative part of the SAT (SATM hereafter). The SAT is typically administered in the last year of high school, and it is used as a college entrance exam in the U.S. The maximum score for the SATM during this time period
was 800. SATM scores have been found to have internal consistency coefficients (Cronbach’s alpha) of .92, which is satisfactory for a high-stakes test (Ewing et al., 2005).

**Chemistry achievement.** Students’ achievement in chemistry was measured by students’ scores on the First-Term General Chemistry Blended Examination from the Examinations Institute of the American Chemical Society Division of Chemical Education (ACS exam hereafter). It consists of 40 questions divided evenly among traditional (algorithmic) and conceptual questions (Examinations Institute, 1997).

**Data collection and participants**

Participants are students enrolled in the first term of a general chemistry course during Fall 2010 at a large southeastern public research university. All students took the same exams at the same time. ASCIv2 and TMI were given to students during the second week of the semester as a paper and pencil test within regular exam hours, with ASCIv2 proceeding TMI. The ACS exam was given to students as the final course exam at the end of the semester.

A total of 1656 students were enrolled in the course. Among them, 1319 students had SATM scores, 1399 had ASCIv2 scores, 1427 had TMI scores, and 1259 had ACS final exam scores. Overall, 963 students had the complete set of data for SATM, ASCIv2, TMI, and the ACS exam, and were included in the SEM analysis.

Table 7.3 presents the demographic information for the participants in the study. There are 504 females (52.3%) and 459 males (47.7%). The decomposition for this sample is very similar to the enrolled population of 1656 students (52.1% Females and 47.9% males). Regarding race/ethnicity, 53.2% are White, 18.7% Hispanic 14.4% Asian, and 10.3% Black. 54.5% students are in the second year, 24.6% are in the third year, and 12.1% in the first year. Examination of the demographic information for the missing data did not indicate suspicious
patterns with respect to race/ethnicity or year of study that could bias the results based on this sample.

Table 7.3

Demographics: Number and Percentage of Students by Sex and Race/Ethnicity \( (n = 963) \)

<table>
<thead>
<tr>
<th>Sex</th>
<th>No. of students</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>504</td>
<td>52.3</td>
</tr>
<tr>
<td>Male</td>
<td>459</td>
<td>47.7</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>White (Not of Hispanic Origin)</td>
<td>512</td>
<td>53.2</td>
</tr>
<tr>
<td>Hispanic or Latina</td>
<td>180</td>
<td>18.7</td>
</tr>
<tr>
<td>Asian</td>
<td>139</td>
<td>14.4</td>
</tr>
<tr>
<td>Black (Not of Hispanic Origin)</td>
<td>99</td>
<td>10.3</td>
</tr>
<tr>
<td>Non-Resident Alien</td>
<td>15</td>
<td>1.6</td>
</tr>
<tr>
<td>American Indian or Alaskan Native</td>
<td>5</td>
<td>0.5</td>
</tr>
<tr>
<td>Ethnicity Unspecified</td>
<td>9</td>
<td>0.9</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islanders</td>
<td>4</td>
<td>0.4</td>
</tr>
</tbody>
</table>

Model Specification for SEM Analysis

Three SEM models were specified and tested in this study. The first model is represented in Figure 7.1. This model assumes no measurement error in each variable; therefore, it includes only measured variables. This model is a simple path analysis model. Achievement as measured directly by the ACS final exam is predicted by math ability as measured by the SATM, attitude as measured by the total score on ASCIv2, and prior conceptual knowledge as measured by the total score on the TMI.

Figures 7.2 and 7.3 present two SEM models included in the analysis. Both of these models include two parts, the path or structural model and the measurement model, in the analysis. Figure 7.2 shows the same relationship among variables as in the path model (Figure 7.1) for the student variables, achievement, math ability, prior conceptual knowledge, and
attitude toward chemistry. Note that, in this more complex model, three variables are included as latent variables, which are measured by indicators such as SATM, ASCI1, and TMI1. All three models as illustrated in Figures 7.1 through 7.3 posited correlations among the predictors.

The TMI, which is a measure of prior conceptual knowledge in this study, is only measuring three basic concepts from all of chemistry; therefore, it might be argued that the TMI is not a relevant predictor of achievement as measured by ACS exam scores. In order to address this concern, TMI was removed from the prediction for Model 3 as represented in Figure 7.3. Accordingly, Model 3 examines the prediction of achievement with one less predictor than does Model 2. Except for the difference of one path coefficient, both of these models include the same measured variables, latent variables, and relationships among variables, thus a comparison of chi-square change between the two nested models will provide information about which model better fits the data. This comparison will also allow us to determine whether prior conceptual knowledge as measured by the TMI is a relevant predictor of achievement for this data.

Both SEM models have a total of nine latent variables (represented with ovals) and 19 measured variables (represented as rectangles), and 23 error variances (reflected in circles). Each SEM model has 190 pieces of information in the covariance matrix. Model 2 estimates 47 parameters, with degrees of freedom at 143. Model 3 estimates 66 parameters, with degrees of freedom at 144.

For model identification purposes, the loading for the first item on its corresponding latent variable was set to 1. The error variances for two measured variables, ACS and SATM, were calculated from the Cronbach’s alpha coefficient for each one according to the procedure recommended by Valiente and co-workers, because they are proprietary exams and no further information was available (Valiente, Lemery-Chalfant, & Reiser, 2007). The ACS final exam
had a Cronbach’s alpha of .85 for the 40 items for this sample; therefore the value of the error variance was set to 15% of the total variance obtained (47.1), resulting in an error value set to 7.09. For the SATM, previous reports indicated the quantitative SAT had an alpha of .92; therefore the error variance was set to 482, which is 8% of the total variance for the variable (6026).

**Figure 7.2** SEM Model 2 with Three Predictors

**Figure 7.3** SEM Model 3 with Two Predictors
Data Analysis

Descriptive statistics and assumption checking for SEM

Descriptive statistics for each measured variable were obtained using SAS 9.3. Univariate and multivariate normality and outliers were examined.

Structural Equation Model (SEM)

SEM was performed using the CALIS procedure in SAS 9.3 to investigate both model fitness and parameter estimates. Fit indices such as Chi-Square ($\chi^2$), Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR) were examined to assess the fitness of the model to the data. The criteria of CFI value at least .95 and SRMR value less than 0.08 are used to indicate a good model fit (Hu & Bentler, 1999). SEM analysis parameters such as model specification, standardized parameter estimates and effect sizes, and predicted variance indicated by $R^2$, were reported in compliance with the guidelines proposed by Kline (2010).

Results

Descriptive statistics and assumption checking for SEM

The analysis was performed using 963 complete sets of data for the variables included in the model. Descriptive statistics for the sample were examined to observe the general differences in the variables and to examine the univariate normality of each variable. As shown in Table 7.4, the mean for the ASCIv2 variables ranged from 2.75 to 4.56. Four of the variables (ASCI1, ASCI4, ASCI5, ASCI7) were reverse coded for interpretation; therefore, higher scores mean students feel chemistry is intellectually accessible and emotionally satisfying. For the TMI, items 2, 3, and 9 are apparently quite difficult, because less than 20% of the students got these items correct.
Univariate normality was first assessed via examination of the skewness and kurtosis for each variable. According to Table 7.4, all the variables for ASCIv2 (ASCI1-ASCI8), SATM, and ACS Exam have absolute skewness and kurtosis values less than one; however, for the TMI, items 2, 3 and 9 have kurtosis values beyond two, which may be cause for concern (Curran, West & Finch, 1996). These items were difficult for students in our sample; therefore, these items may need to be revised for future administrations. A SEM model using only the composite (sum) score for the nine TMI items was run, since the skewness and kurtosis for the composite TMI score are both less than one (Table 7.4). The model fit and parameter estimates for the model did not change, suggesting that the model is robust, and indicating relative univariate normality. Therefore, the SEM model with the individual TMI was used for the analysis and interpretation.

**Table 7.4**

*Descriptive Statistics for Each Measured Variable* (n=963)

<table>
<thead>
<tr>
<th>Measured variables</th>
<th>M</th>
<th>SD</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASCI1</td>
<td>3.45</td>
<td>1.42</td>
<td>0.31</td>
<td>-0.33</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI2</td>
<td>3.33</td>
<td>1.42</td>
<td>0.37</td>
<td>-0.40</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI3</td>
<td>3.81</td>
<td>1.47</td>
<td>0.05</td>
<td>-0.56</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI6</td>
<td>2.75</td>
<td>1.38</td>
<td>0.80</td>
<td>0.39</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI4</td>
<td>4.10</td>
<td>1.48</td>
<td>-0.06</td>
<td>-0.45</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI5</td>
<td>4.04</td>
<td>1.68</td>
<td>0.04</td>
<td>-0.76</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI7</td>
<td>3.95</td>
<td>1.46</td>
<td>0.04</td>
<td>-0.18</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>ASCI8</td>
<td>4.56</td>
<td>1.54</td>
<td>-0.35</td>
<td>-0.51</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>TMI1</td>
<td>0.34</td>
<td>0.47</td>
<td>0.67</td>
<td>-1.55</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI5</td>
<td>0.30</td>
<td>0.46</td>
<td>0.88</td>
<td>-1.22</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI7</td>
<td>0.58</td>
<td>0.49</td>
<td>-0.34</td>
<td>-1.89</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI2</td>
<td>0.12</td>
<td>0.33</td>
<td>2.31</td>
<td>3.32</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI4</td>
<td>0.20</td>
<td>0.40</td>
<td>1.51</td>
<td>0.27</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI9</td>
<td>0.09</td>
<td>0.29</td>
<td>2.82</td>
<td>5.96</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI3</td>
<td>0.11</td>
<td>0.32</td>
<td>2.46</td>
<td>4.07</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI6</td>
<td>0.28</td>
<td>0.45</td>
<td>0.96</td>
<td>-1.09</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>TMI8</td>
<td>0.22</td>
<td>0.41</td>
<td>1.37</td>
<td>-0.13</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>SATM (α=.92)</td>
<td>596.00</td>
<td>77.63</td>
<td>-0.11</td>
<td>0.02</td>
<td>370</td>
<td>800</td>
</tr>
<tr>
<td>ASCIv2 (α=.88)</td>
<td>3.75</td>
<td>1.09</td>
<td>-0.10</td>
<td>0.21</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>TMI (α=.45)</td>
<td>2.25</td>
<td>1.61</td>
<td>0.90</td>
<td>0.99</td>
<td>0</td>
<td>9</td>
</tr>
<tr>
<td>ACS Exam (α=.85)</td>
<td>26.20</td>
<td>6.86</td>
<td>-0.20</td>
<td>-0.51</td>
<td>7</td>
<td>40</td>
</tr>
</tbody>
</table>

Note: The relatively large kurtosis values for the three difficult TMI items are indicated in red.
Outlier assessment testing revealed a single outlier with the largest Mahalanobis distance of 59.4 away from the sample mean. The SEM analysis was re-run after removing the outlier to examine the consequence of the outlier. The values of the fit indices ($\chi^2$, CFI, SRMR) did not change significantly, suggesting that the fitness of the data to the model remains the same. The path coefficients also did not change for the analysis without the outlier. For these reasons, the original SEM with the complete data set was used for the analysis and interpretation. Since relative univariate normality was obtained, the outlier assessment did not reveal any consequential outliers, and the sample size is large, Maximum Likelihood (ML) estimate was used as the estimation method for the proposed SEM models (Finney & DiStefano, 2006).

**Structural equation model analysis**

**Model fit.** For Model 1, model fit is not assessed and fit indices are not reported since it is a path analysis, e.g., its degrees of freedom are zero and there is no measurement error (Raykov & Marcoulides, 2006). The overall fit for SEM Models 2 and 3 is reported as shown in Table 7.5. Although the Chi-square ($\chi^2$) test is statistically significant (p<.001) for both models, suggesting that the models did not fit the data well, the large value of $\chi^2$ arises from the large sample size (Brown, 2006, p. 81). Other fit indices, insensitive to sample size, indicate that both SEM models have good fit (Hu & Bentler, 1999). For Model 2, the CFI value (0.95) meets the suggested criterion of at least .95 and the SRMR value (0.04) meets the suggested criterion of smaller than 0.08. For Model 3, the CFI value (0.94) is close to the suggested criterion, and the SRMR value (0.04) remains less than 0.08. Since Model 3 is a nested model within Model 2, the change in $\chi^2$ for the two model fits can be used to compare the models (West, Taylor & Wu, 2012, p. 211).
The change in $\chi^2$ is 71, with a change in degrees of freedom of 1, which indicates that the data fits Model 2 significantly better than Model 3.

**Table 7.5**

*Fit Indices and Predicted Variance for Models (N = 963)*

<table>
<thead>
<tr>
<th>SEM Model</th>
<th>$\chi^2$ (df)</th>
<th>p-value</th>
<th>SRMR</th>
<th>CFI</th>
<th>$\Delta$ df</th>
<th>$\Delta \chi^2$</th>
<th>Predicted Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>46%</td>
</tr>
<tr>
<td>2</td>
<td>395 (143)</td>
<td>&lt;.001</td>
<td>.04</td>
<td>.95</td>
<td>1</td>
<td>71 (s.s.)</td>
<td>69%</td>
</tr>
<tr>
<td>3</td>
<td>466 (144)</td>
<td>&lt;.001</td>
<td>.04</td>
<td>.94</td>
<td>1</td>
<td>71 (s.s.)</td>
<td>57%</td>
</tr>
</tbody>
</table>

As shown in Figure 7.6, Model 3 has one less predictor for achievement, which allows examination of how removing prior content knowledge from the model might influence the prediction accuracy. The prediction accuracy was assessed using the percentage of variance in achievement that is accounted for the sets of predictors. Model 1 has the lowest prediction accuracy with 46% of the variance explained as shown in Table 7.5. For Model 3, 59% of the variance in the achievement variable is explained by the set of predictors for this model while for Model 2, 69% of the variance is explained.

The following formula, for determining the $F$-value, was used to examine whether adding the variable of prior conceptual knowledge contributed significantly to the prediction in Model 3 (Pedhazur, 1997).

$$F = \frac{(R_{l arg e}^2 - R_{small}^2)/(k_{l arg e} - k_{small})}{(1 - R_{l arg e}^2)/(N - k_{l arg e} - 1)}$$

where $R^2$ is the proportion of variance explained by the model

$k$ is “number of predictors”

$N$ is total sample size

Results from this test, $F(1, 959) = 371.2$, $p < .001$, indicate that the unique contribution of prior conceptual knowledge is significant for the prediction.
Based on the change in $\chi^2$ and the explained variance, Model 2 can better explain the relationships among the variables than Model 3. Therefore, Model 2 is the better model to represent the relationships among the studied variables. Below the relationships for the three models are explained in more detail.

**Decomposition of relationships**

The standardized parameter estimates resulting from the three models are presented in Figures 4 through 6, respectively.

**Measurement model for ASCIv2**

The first set of coefficients estimated in both SEM Models 2 and 3 represent the attitude toward chemistry variable as measured by ASCIv2. This attitude variable has two components: the cognitive (“Intellectual Accessibility”) and the affective (“Emotional Satisfaction”), indicated as “Cognitive” and “Affective” in the diagrams. As can be seen from the high positive (.57-.80) path coefficients for the cognitive component in the path diagram, “Intellectual Accessibility” has a strong influence on all of the measured variables. Similarly, as shown by the positive (.53-.79) values of the path coefficient for the affective component “Emotional Satisfaction” has a strong influence on its measured variables. The second order of this attitude test structure is the relationship between these two latent variables with overall “Attitude toward Chemistry.” The path coefficient from attitude toward chemistry to cognitive is .89 and from attitude toward chemistry to affective is .93, which provides strong validity evidence based on the internal structure of ASCIv2 that items are measuring two components of attitude toward the subject of chemistry. The variance in attitude toward chemistry explained by the cognitive component is 78%, and by the attitude component is 86%. These results are the same for both SEM models and consistent with literature (Brandriet et al., 2011; Xu & Lewis, 2011).
Figure 7.4 Standardized Parameters for the Path Analysis Model 1

Figure 7.5 Path Diagram with the Standardized Parameter Coefficients for Model 2

Figure 7.6 Path Diagram with the Standardized Parameter Coefficients for Model 3
Measurement part for TMI

TMI addresses three basic concepts, bond energy, phase changes, and ionic/covalent bonds. Each concept is measured using three items. As shown in Figures 5 and 6, the test items loaded well on their respective scales. However, the scale of bond energy does not load well under the general construct of chemistry conceptual knowledge, since its loadings are -.06 and -.04 for Model 2 and 3, respectively. For the other two concepts, the loadings are higher, .85 and .84 for phase changes in Model 2 and 3, respectively and for ionic/covalent bond the loadings are .71 and .72 for Model 2 and Model 3. This situation indicates that the bond energy concept, as measured by the TMI, does not really function well with the other two concepts as an indicator of overall understanding.

Relationships among the three predictors

Table 7.6 presents the relationship among the three predictors as measured by correlation coefficients for the three models. For Model 2, the correlation between math ability and prior conceptual knowledge is .61, between math ability and attitude toward chemistry is .29, and between attitude toward chemistry and conceptual knowledge .48. These correlation values are considered moderate, and these results are consistent with literature related to math ability and conceptual understanding (Holme & Murphy, 2011; Nicoll, 2001), math ability and attitude toward chemistry (Brandriet et al., 2011; Xu & Lewis, 2011), and attitude toward chemistry and conceptual understanding (Kim & Song, 2010; Nieswandt, 2007).

The correlation coefficients from Model 2 are comparable with those of Model 3. However, since Model 1 did not account for measurement error, the correlation for Model 1 is underestimated. For example, math ability is moderately correlated to prior conceptual
knowledge with the $r$ at .37 for Model 1, but a much higher correlation is observed for Models 2 and 3 with the estimates for $r$ at .61 and .64, respectively.

**Table 7.6**

*Standardized Correlation Coefficients*

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Math ability - Prior conceptual knowledge</td>
<td>.37</td>
<td>.61</td>
<td>.64</td>
</tr>
<tr>
<td>Prior conceptual knowledge - Attitude</td>
<td>.28</td>
<td>.48</td>
<td>.52</td>
</tr>
<tr>
<td>Math ability - Attitude</td>
<td>.25</td>
<td>.29</td>
<td>.29</td>
</tr>
</tbody>
</table>

Note: $r = .1, .3, .5$ for weak, medium and strong effect size (Cohen, 1988, p.78)

**Relationships with student achievement in college chemistry**

The next part of the model concerns the relationships between achievement and the predictors. Standardized path coefficients and effect sizes from the three models are shown in Table 7.7. All path coefficients achieved significance. For Model 1, the path coefficient from attitude to achievement is .19, which indicates a positive, though small ($f^2 = .04$), effect on achievement. The path between math ability and achievement is .49, which indicates a strong positive effect ($f^2 = .32$). The path coefficient between the third predictor, prior conceptual knowledge and achievement is .20, which again indicates a small ($f^2 = .04$) positive effect of the conceptual knowledge on achievement.

Model 2 assessed the relationship of achievement with three predictors. First, the path coefficient between math ability and achievement is .50, which indicates a strong ($f^2 = .33$) positive effect. The path coefficient from attitude to achievement is .16, which indicates a small ($f^2 = .03$) positive effect on achievement. The path coefficient between prior conceptual knowledge and achievement is .36, which indicates a moderate ($f^2 = .15$) positive effect of the conceptual knowledge on achievement. For Model 3, the path coefficient from math ability to achievement increases to .62, a much larger ($f^2 = .62$) effect, and from attitude to .28, a slightly
larger but still small ($f^2 = .09$) effect. In other words, when the prior conceptual knowledge is not included in the model, an overestimation of the path coefficients for math ability and attitude is observed. This overestimation is most obvious for math ability.

**Table 7.7**

*Standardized Path Coefficient and Effect Size for the Regression on Achievement*

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Model 1</th>
<th></th>
<th>Model 2</th>
<th></th>
<th>Model 3</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Path Coefficient</td>
<td>Effect size $f^2$</td>
<td>Path Coefficient</td>
<td>Effect size $f^2$</td>
<td>Path Coefficient</td>
<td>Effect size $f^2$</td>
</tr>
<tr>
<td>Math ability</td>
<td>.49</td>
<td>0.32</td>
<td>.36</td>
<td>0.15</td>
<td>.62</td>
<td>0.62</td>
</tr>
<tr>
<td>Attitude</td>
<td>.19</td>
<td>0.04</td>
<td>.16</td>
<td>0.03</td>
<td>.28</td>
<td>0.09</td>
</tr>
<tr>
<td>Prior concept</td>
<td>.20</td>
<td>0.04</td>
<td>.50</td>
<td>0.33</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Note: $f^2 = .02$, .15, and .35 for small, medium and large effect sizes, Cohen (1988), p. 413

**Discussion and Implications**

This study examined three alternate models to predict student achievement in introductory college chemistry. All three proposed models have reasonable fit to account for variance/covariance in the observed variables. We found that Model 2 can best represent the relationships among these achievement-relevant variables, based on the model fit, chi-square change when removing one predictor, and the proportion of variance accounted for student achievement in chemistry.

Each hypothesis was addressed in this study. First, the model results suggest a significant effect for each predictor on the performance of college chemistry. Based on Model 2, each predictor contributed a unique portion for prediction. Students’ achievement as measured by an ACS exam at the end of the semester can be mostly predicted by the combination of predictors at as early as the second week of the semester. This means, besides student math ability, both prior conceptual knowledge and attitude toward chemistry exert a significant unique effect on student achievement. This finding has a very important implication for instructors, since they can
improve students’ achievement in chemistry not only by focusing on helping students to build conceptual knowledge, but also by fostering their positive attitude toward chemistry.

Second, the model suggests significant relationships among the predictors of attitude, math ability, and prior chemistry conceptual knowledge as shown in Figures 5 and 6. These results imply that instructors may emphasize the interaction of multiple aspects in the classroom as a way to help students’ performance in the course.

Third, the comparison of the nested models suggests that Model 2 fits the data better than Model 3. Although math ability is often used as a proxy for cognitive factors to predict chemistry achievement in the literature, the importance of student conceptual knowledge should not be diminished. Leaving out prior conceptual knowledge resulted in significant over-estimation of the importance of math ability. This clear result should encourage instructors to re-think simplistic math-focused instructional interventions as insufficient to maximize improvement in students’ potential for achievement in college-level general chemistry.

Fourth, the amount of explained variance in students’ achievement significantly increased with a SEM model (Model 2 and 3) as opposed to a path analysis model (Model 1). Model 1 predicts only 46% of the variance in achievement, since the measurement tools used are assumed to be perfect and the prediction effect tends to be underestimated. Model 2 predicts 69% of the variance in student achievement by taking into account the measurement error associated with each observed variable. Model 3 predicts less than Model 2 due to missing one predictor, prior conceptual knowledge. Therefore, the accuracy of predicting achievement can significantly decrease if measurement error is neglected, and the same is true if important predictors are absent.

Some limitations are present in this study. First, the sample was drawn from a unique
population, a single semester of a general chemistry course at a particular institution; therefore, the model may work for this unique context but not apply to other situations. Second, an appropriate assessment that covers a broader range of concepts that students need to know when they enter college chemistry course needs to be developed. In this study, the prior knowledge exam is based on three important concepts, but there is evidence that it does not represent the broad range of chemistry conceptual knowledge. If the measure is improved, we expect to have better prediction of achievement and model fit. Third, only three predictors were added to the model. Additional predictors were not explored, nor were alternative measurement strategies.

This study has several implications for chemistry education researchers with respect to further studies. In one hand, the obtained model needs to be replicated in another context. On the other hand, there are multiple ways to optimize the model. Due to the scarcity of SEM studies in chemistry education, multiple perspectives on aspects of SEM modeling, such as model specification, predictor choices, and operationalization and measurement of the variables, have yet to be explored.

Considering that there is 31% of variance that is not explained by the model in the study, future studies can explore what other factors could explain the rest of the achievement. For example, adding more appropriate predictors or mediating variables can help better explain the relationships and improve prediction. Other factors such as spatial ability (Carter, Larussa, & Bodner, 1987; Merchant et al., 2012; Pribyl & Bodner, 1987), self-efficacy (Kan & Akbas, 2006; Merchant et al., 2012; Nieswandt, 2007; Uzuntiryaki & Aydin, 2009), or instructional strategy (Merchant et al., 2012; Schroeder, Scott, Tolson, Huang, & Lee, 2007) are reported as factors influencing student achievement in previous studies. More work is needed to capture the broad range of factors that are potentially important to understand chemistry learning. Moreover, the
operationalization and measurement of key variables can be refined to better reflect the learning theory and empirical data. In sum, whether the necessary predictors are included and how well they are measured will affect the model fit and change the magnitude of path coefficients. By trying multiple combinations of models in multiple samples, over time researchers can achieve a better understanding of interactions among factors influencing student achievement in chemistry, help to build discipline-based learning theory, and provide information to support curriculum innovation.

References


VIII. Conclusions and Implications

Summary Conclusion

With the aim of improving STEM education in mind, the objective of this dissertation is to evaluate and recommend appropriate instruments for measuring constructs of spatial ability and attitude respectively that people can trust the use in college chemistry and STEM education, and to offer better insights than the previous literature into the unique roles played by factors in student chemistry performance. A substantial amount of work has been done to examine the factors influencing student performance; however, due to the complex nature of educational research and the constraints of classroom settings, many problems remain unsolved. For example, the instruments used may not meet the quality criteria required by the Standards (AERA, APA, & NCME, 1999; Arjoon, Xu, & Lewis, 2013). Examination of previous literature demonstrates conflicting results on the relationship between attitude, spatial ability and chemistry achievement (Carter, LaRussa, & Bodner, 1987; Xu, & Lewis, 2011). This dissertation attempts to establish two high-quality instruments and rigorous correlation analysis, such as structural equation modeling (SEM), in order to rule out major sources of disagreement from the measurement itself and the analysis technique. Thus, the findings can provide the educational community with trustworthy knowledge of the influence of these two factors.

This dissertation included five studies adopting primarily a quantitative approach. Studies 1, 3 and 4 provided psychometric evidence for using the ROT and ASCIv2 instruments to measure the mental rotation component of spatial ability and attitude toward chemistry.
respectively, in college chemistry classrooms. The findings reveal that ROT and ASCIv2 can serve as appropriate measurement tools for college student samples in the United States. By using ROT and ASCIv2, which have been well-characterized in previous findings, studies 2 and 5 focused on establishing a trustworthy model to explain the contributions of these two important factors to student chemistry performance.

Regarding spatial ability, results from the second study, *The role of spatial ability in students’ progression through organic chemistry*, did not support the statement that low mental rotations ability prevented students from moving on through the organic chemistry course sequence. This is consistent with general view in the literature that student chemistry performance is more related to general reasoning skills and prior knowledge than visualization skills (Baker & Talley, 1972; Keig & Rubba, 1993; Wu & Shah, 2004). In other words, students progressed and obtained good course grades in organic chemistry in the current setting regardless of their spatial ability level. This finding is interesting because literature suggests that females tend to underperform males in the spatial ability domain, especially with respect to mental rotations, which can be an unconscious sex stereotype that prevents entry into areas deemed visually intensive (Hill, Corbett, & St. Rose, 2010; Xu, 2008). For the population of this study, a grade of A in organic chemistry did not necessarily mean a high spatial ability when math ability was controlled, and vice versa. According to the findings of this dissertation, ROT is not recommended as a screening test to suggest that only students of high mental rotation ability should make through the organic chemistry course sequence.

Regarding attitude, findings from the last study, *College students’ attitudes toward chemistry, conceptual knowledge and achievement: SEM analysis*, revealed that using three predictors at the beginning of the semester can explain a very high percentage (69%) of the
variance in chemistry achievement at the end of the course. Both prior conceptual knowledge and attitude toward chemistry contribute a significant unique portion to the prediction of chemistry achievement when controlling for math ability. This is consistent with other studies that have examined affective variables such as attitude for college chemistry achievement (House, 1995). The results suggest that instructors can improve students’ college chemistry achievement not only by helping them develop a conceptual understanding of chemistry and mathematics, but also by nurturing a positive attitude toward chemistry.

Regarding the psychometric quality of the instruments, the first study provides evidence that a 20-item ROT test is an appropriate tool to measure the mental rotation component of spatial ability for college students. A measurement invariance test indicated that ROT items functioned in the same way across sexes; therefore, ROT can be used for comparing sex difference with less concern about test bias. The third and fourth studies offered evidence that ASCIv2 items functioned to measure attitude toward chemistry for students outside of the United States. The score interpretation of ASCIv2 has been supported for students in three institutions in the United States by previous studies (Brandriet, Xu, Bretz, & Lewis, 2012; Chan & Bauer, 2014; Xu & Lewis, 2011). The fourth study identified a problematic item (challenging vs. not challenging) for students at an institution in Saudi Arabia and raised concern about using this item to indicate the construct for students with a background similar to that of Saudi Arabian students. Findings from these studies suggest that other institutions with characteristics and student populations similar to those of US participants may use these instruments and interpret them in a similar fashion.

ROT has been widely used in college classrooms (Maeda & Yoon, 2013). Encouraged by psychometric evidence of ROT scores from this work, instructors can continue the use of
ROT to track student ability to mentally rotate a block shape in their classrooms. For the attitude instrument, ASCIv2, caution should be taken for students like those in Saudi Arabia, i.e., if the item challenging vs. unchallenging is associated with negative emotional arousal, then this item should be excluded. Consequently, not every existing instrument can be assumed to be “good”, students may not respond to the instrument items as expected, and how to appropriately interpret item score should be carefully examined. This work strongly urges educators and researchers to always check whether instruments function as expected to measure the intended variable, especially when there is a large sample size suitable for techniques such as factor analysis. When internal structure evidence is not supported by data that has been collected from students through the use of an existing instrument, the response process of students should be further examined by qualitative approaches, such as interviews, to elucidate how the instrument functions for the specific population and how it can be modified accordingly. This work indicate the need for the chemistry education community to continue collectively gathering evidence on site each time a test is administered, in order to support research findings.

This work (especially studies 1, 3, and 4) highlights the importance of using instruments with sufficient psychometric evidence to yield valid interpretations in order to answer questions about spatial ability and attitude. There are many situations where data collected from instruments are used to support learning theories, to confirm claims, and to inform practice (Creemers & Kyriakides, 2010). Examples of claims could be that only gifted visual-spatial learners should enter an area requiring spatial ability, or a training curriculum is effective and worthy of wide implementation (Wanzel et al., 2002). Only with high-quality data can more convincing conclusions be reached. The findings from this work contribute to the educational community by successfully using the framework of evidence sources to evaluate the reliability
and validity of instrument scores to meet the commonly accepted standards (AERA, APA, & NCME, 1999; Arjoon, Xu, & Lewis, 2013). Specifically, this work emphasized the importance of the alignment of instrument design (measure a construct using multiple indicators), psychometric evidence (e.g., based on internal consistency and internal structure) and the interpretation of the total score of multiple items to support claims.

**Limitations and Implications**

This dissertation has several limitations and implications. First, it only explored limited aspects of two constructs, the mental rotations component of spatial ability of block shapes and two subscales of attitude toward chemistry (i.e., intellectual accessibility and emotional satisfaction). Other aspects of these constructs may also be worthy of further study, e.g., mental transformation of molecular structure and organic reactions (Stieff, 2011), and attitude that chemistry is generally dangerous and bad for the environment (Gilbert, 2006). In addition, many other factors, such as motivation (Ames, 1992), self-efficacy (Bandura, 1994; Villafañe, Garcia, & Lewis, 2014), problem-solving skills (Bodner & McMillen, 1986; Mahalingam, Schaefer, & Morlino, 2008), may also be important factors in students’ decision to enter and stay in STEM areas, and thus are worthy of further examination.

Second, the approaches used to evaluate the instruments in this dissertation are not necessarily ideal to measure the intended variables, because it is hard to establish the “right” measurement for a complex variable such as chemistry performance in a natural educational setting. For example, this dissertation used student scores on a nationwide exam created by the American Chemical Society Exams Institute (Examinations Institute, 1997) as a proxy for general chemistry achievement, and used organic chemistry course grade to loosely quantify student chemistry performance, according to what measures were the most appropriate and
available in this study setting. It is even harder to capture overall school performance and job performance, as it must be generalized across institutions (Kuncel, Hezlett, & Ones, 2004). This work did not investigate multiple variables crossed with multiple methods; that is, multitrait-multimethod design is the ideal methodology to evaluate multiple traits from maximally different ways such as questionnaire and interview (Campbell & Fiske, 1959; Eid, 2000). This dissertation has implications for researchers seeking to identify the most appropriate measures to utilize in the statistical analysis.

And third, it is noteworthy that this work used convenience samples at each participating institution, which is not representative of the whole college student population and limits the generalizability of these findings. Whether instruments function well or the relationships among variables hold warrants further investigation at more sites. This work calls for more studies to fill this research gap and to replicate the findings in other contexts.

Finally, this dissertation used a survey design to explore the relationship among variables, and did not directly evaluate a specific curriculum using an experimental design. Note that measurement alone, without actual classroom instruction, cannot improve the education system. Established measurements need to be used together with solid research, practice, and rigorous evaluation to bring forth sustainable progress. From the findings of this dissertation, instruments such as ROT and ASCIv2 can be used in future experimental design to better evaluate a reform effort regarding spatial ability or attitude for the purpose of wider implementation of effective instructions.

Attitude is important in the prediction of chemistry performance, while the role of spatial ability is still ambiguous and worthy of further examination. Instructors should focus more on fostering students’ positive attitude toward chemistry and helping them learn the concepts of
chemistry. Active learning strategies, as discussed in the Introduction, are encouraged for further implementation and evaluation in order to foster positive attitude and learning of content.

For educators who are interested in spatial ability research and practice, Stieff and colleagues have provided important information regarding how students and faculty use spatial reasoning, especially mental rotations, to solve specific problems (Stieff, 2007, 2011, 2013; Stieff & Raje, 2010). More work can be done to better understand spatially related chemistry problems, and to inform effective instruction in that regard. Because the ROT is widely used for STEM majors (Ceci, Williams, & Barnett, 2009; Sorby, 2009), and the ROT is supported by the current psychometric evidence as a sound mental rotations ability test for college students, this instrument can conveniently be used. Educators can use ROT to identify sex gaps in their samples or the relationship with other key constructs with more confidence than before, while continuing to accumulate evidence appropriate to their setting and research purpose. Additionally, in spatial ability training programs, instructors can use it for pre and post tests to determine the effectiveness.

Active learning approaches such as the Process-Oriented Guided Inquiry Learning (POGIL), can help students at different levels of spatial ability to learn abstract concepts such as chemical applications of group theory (Southam & Lewis, 2013). Since both sexes can benefit from visualization tools and spatial ability training, all students should have access to visualization tools and activities to facilitate mental visualization (Baenninger & Newcombe, 1989). For students who cannot visualize well or prefer using analytic strategies, textbooks and instruction should clearly elaborate on the procedure and reduce the high working memory load required to solve space-related problems. Therefore, in order to help more students learn,
instruction and training should not be limited to mental rotations or spatial ability, but rather should also encourage a combination of strategies (Stieff et al., 2014).

This study cautions against the claim that females generally lack spatial ability and thus cannot learn chemistry well until they receive extra training in spatial thinking to improve their spatial ability. Instead, students of both sexes should not be overly concerned that a lack of spatial ability will hinder their academic success. Instead, they can apply analytic strategies, use molecular models, or learn in whatever ways feel comfortable to them. As shown by Stieff and colleagues, people who use analytic strategies to solve space-related problems are equally successful (Stieff, 2007; Stieff & Raje, 2010), and a group that has received training in analytical strategies can even perform better on certain tasks than a group that has received training in spatial skills (Stieff, 2013). Using concrete and pseudo-concrete types of representations (e.g. molecular model sets, computer-generated models) is a more effective way to help students understand molecular structure than using more abstract types (Ferk & Yrtacnik, 2003). With these available research sources and training programs in mind, educators can encourage and help more students into the advanced courses regardless of their spatial visualization skills.

Overall, this dissertation highlights the importance of gathering more evidence for instrument use in STEM contexts. These instruments can lay a solid foundation for research and practice. By thoroughly researching critical factors, educators can be much clearer on how the factors interact with one another and influence student performance and persistence. Instruments can also facilitate the evaluation of curricular innovations to help students persist in STEM areas. For the proposed classroom intervention, well-established instrument tools such as ROT and ASCiV2 can be used with solid research design to evaluate the effectiveness before wider dissemination is recommended. Eventually, professionals in classroom instruction should use
instruments in collaboration with experts from various areas, such as measurement, research, and evaluation. More efforts can be made to improve student’s attitude and build the confidence needed to overcome any barriers and succeed. In this way, we can take action collectively to improve the education system and to recruit and prepare more people for future careers in STEM areas (National Academy of Sciences, National Academy of Engineering, & Institute of Medicine, 2010).

References


Appendices
Appendix A: Administration instruction for ROT

Fall 2008

ROT Proctoring Instructions

This exam has two portions – the ROT test and an additional questionnaire.

1. Hand out the instruction sheet and scantrons first.
   ➢ Try to make sure that there is at least one seat between each student and that there are no notes, books, etc. out and open during the test. In some rooms students may have to sit closer together. If so you will need to stand near them during the exam.

2. Make your announcements.

3. Hand out the ROT test booklet.
   ➢ Each student receives only one test booklet.

4. Once all students have received the test booklet, tell them to read the directions in the first 2 pages of the test booklet. Tell them "you are NOT allowed to proceed to the 3rd page with the real test questions until I tell you so".
   ➢ Give students 3 minutes to read the directions (in the first 2 pages of the booklet). Do NOT allow any student to proceed to the 3rd page; stop them if they try to!

5. Three (3) minutes later, tell them to proceed to the 3rd page; immediately start the 10-minute "clock" for the ROT test.
   ➢ Students are not allowed to talk during the exam. Even if they finish before the 10 minutes has passed, they must sit quietly.

6. When a student completes the ROT test portion, take only the ROT test booklet and give the student the additional set of questions. The students will use the original scantron to answer the additional questionnaire on. Check the ROT test booklet to make sure it has not been written on – if it has been written on, set it aside.
   ➢ After the 10 minute time period, students may no longer work on the ROT test and you must collect all booklets and distribute the additional questions. Make additional announcement at this time.

7. At the end of the whole exam (by 6:05 pm), collect everything.
   ➢ Students may not take any paper used during the exam out of the room.

Make these announcements to your students:

1. "Make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U** ID number without the U, and your course section number. It is NOT enough to write your name or other information down, you must BUBBLE IN the information. The scantron machine does NOT read handwriting; it reads bubbles."
2. No one can turn in the ROT test until 10 minutes has passed.
3. No one can receive a test after 10 minutes of the test being distributed.
4. No one may leave the exam room before 30 minutes has passed.
5. "Here are some specific announcements pertaining to the ROT test":
   • "You MAY NOT WRITE on the ROT test. The instruction sheet should be used as scratch paper if needed."
   • "The first two pages in the test booklet are directions with two examples. You will have 3 minutes to read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until I tell you so."
   • "The ROT test questions starts on the 3rd page and they are numbered from 1 to 20. You will have 10 minutes to complete the ROT test."
   • "At the end of 10 minutes, the ROT test booklet (but not the scantron) will be collected and the additional questionnaire distributed."
   • "At the end of the exam (by 6:05 pm), you must turn in everything (including the additional questionnaire, the scratch paper AND the scantron)."

Information for Proctors After the Test:

1. Collect the scantrons and check to make sure that the scantron is filled out correctly. This may take more time than just picking up the scantron, but the students can wait. It is very important that you check to make sure everything is completely bubbled in.
2. Collect everything (including scratch paper and additional questionnaire). This is to protect the confidentiality of the exam and prevent potential lawsuits.
3. Return the collected scantrons and test materials to ROOM *** immediately following the exam.

It is your job to make sure that the scantrons are filled out correctly (i.e. everything bubbled in, filled out in pencil, etc.) and to have them in a neat and organized manner.
Instruction Sheet for Students

1) Before the test begins, make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U** ID number without the U, and your course section number.
   ▶️ It is not enough to write your name or other information down, you must **BUBBLE IN** the information with a pencil. The scantron machine does not read handwriting; it reads bubbles.

2) First you will receive the ROT test portion of the exam. You will have 10 minutes to work on this portion.

3) **Do not write on the test booklet.** You may use this instruction sheet as scratch paper if needed.

4) The first two pages in the test booklet are directions with two examples. Read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until told by your exam proctor.

5) The real test questions are numbered 1 through 20. Bubble in your answers to these questions on the scantron using a pencil (not pen).
   ▶️ This test is intentionally timed so that you must rely on visualization instead of reasoning. Just visualize (picture the rotations of 3-D objects in your head). Do NOT try to reason it out; there is NOT enough time.

6) When time is called, stop working. You will turn in the ROT test booklet and receive an additional set of questions.

7) You will have the remainder of the time to answer the additional questionnaire.

8) Additional questions begin at question number 21. Make sure you begin bubbling in your scantron at question number 21 for these questions. There will be **32 additional questions (numbered 21 through 52).**

9) When finished, hand in everything. **You may not take any paper used during the exam out of the exam room.**
**ROT Proctoring Instructions**

This exam has two portions – the ROT test and an additional questionnaire.

1. **Hand out the instruction sheet and scantrons first.**
   - Try to make sure that there is at least one seat between each student and that there are no notes, books, etc. out and open during the test. In some rooms students may have to sit closer together. If so you will need to stand near them during the exam.

2. **Make your announcements.**

3. **Hand out the ROT test booklet.**
   - Each student receives only one test booklet.

4. **Once all students have received the test booklet, tell them to read the directions in the first 2 pages of the test booklet.** Tell them "you are NOT allowed to proceed to the 3rd page with the real test questions until I tell you so".
   - Give students 3 minutes to read the directions (in the first 2 pages of the booklet). Do NOT allow any student to proceed to the 3rd page; stop them if they try to!

5. **Three (3) minutes later, tell them to proceed to the 3rd page; immediately start the 10-minute “clock” for the ROT test.**
   - Students are not allowed to talk during the exam. Even if they finish before the 10 minutes has passed, they must sit quietly.

6. **When a student completes the ROT test portion, take only the ROT test booklet and give the student the additional set of questions.** The students will use the original scantron to answer the additional questionnaire on. Check the ROT test booklet to make sure it has not been written on – if it has been written on, set it aside.
   - After the 10 minute time period, students may no longer work on the ROT test and you must collect all booklets and distribute the additional questions. Make additional announcement at this time.

7. **At the end of the whole exam (by 6:05 pm), collect everything.**
   - **Students may not take any paper used during the exam out of the room.**

**Make these announcements to your students:**

1. "Make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U**ID number without the U, and your course section number. It is NOT enough to write your name or other information down, you must BUBBLE IN the information. The scantron machine does NOT read handwriting; it reads bubbles."
2. No one can turn in the ROT test until 10 minutes has passed.
3. No one can receive a test after 10 minutes of the test being distributed.
4. No one may leave the exam room before 30 minutes has passed.
5. Here are some specific announcements pertaining to the ROT test:
   - "You MAY NOT WRITE on the ROT test. The instruction sheet should be used as scratch paper if needed."
   - "The first two pages in the test booklet are directions with two examples. You will have 3 minutes to read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until I tell you so."
   - "The ROT test questions starts on the 3rd page and they are numbered from 1 to 20. You will have 10 minutes to complete the ROT test."
   - "At the end of 10 minutes, the ROT test booklet (but not the scantron) will be collected and the additional questionnaire distributed."
   - "At the end of the exam (by 6:05 pm), you must turn in everything (including the additional questionnaire, the scratch paper AND the scantron)."
6. Here are some specific announcements pertaining to the additional instrument:
   - The additional questions begin at question number 21. Make sure you begin bubbling in your scantron at question number 21 for these questions. There will be 47 additional questions (numbered 21 through 67).
   - Please answer each question honestly. Your answers are confidential. Your instructor is not going to know your individual results.

Information for Proctors After the Test:

1. Check photo ID while collecting the scantrons. Check to make sure that the scantron is filled out correctly in pencil and includes UID. This may take more time than just picking up the scantron, but the students can wait. It is very important that you check to make sure everything is completely bubbled in.
2. Collect everything (including scratch paper and additional questionnaire). This is to protect the confidentiality of the exam and prevent potential lawsuits.
3. Return the collected scantrons and test materials to ROOM *** immediately following the exam.

It is your job to make sure that the scantrons are filled out correctly (i.e. everything bubbled in, filled out in pencil, etc.) and to have them in a neat and organized manner.
Instruction Sheet for Students

1) Before the test begins, make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U** ID number without the U, and your course section number.
   ➢ It is not enough to write your name or other information down, you must BUBBLE IN the information with a pencil. The scantron machine does not read handwriting; it reads bubbles.

2) First you will receive the ROT test portion of the exam. You will have 10 minutes to work on this portion.

3) Do not write on the test booklet. You may use this instruction sheet as scratch paper if needed.

4) The first two pages in the test booklet are directions with two examples. Read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until told by your exam proctor.

5) The real test questions are numbered 1 through 20. Bubble in your answers to these questions on the scantron using a pencil (not pen).
   ➢ This test is intentionally timed so that you must rely on visualization instead of reasoning. Just visualize (picture the rotations of 3-D objects in your head). Do NOT try to reason it out; there is NOT enough time.

6) When time is called, stop working. You will turn in the ROT test booklet and receive an additional set of questions.

7) You will have the remainder of the time to answer the additional questionnaire.

8) Additional questions begin at question number 21. Make sure you begin bubbling in your scantron at question number 21 for these questions. There will be 47 additional questions (numbered 21 through 67).

9) When finished, hand in everything. You may not take any paper used during the exam out of the exam room.
Fall 2009 ROT

ROT Proctoring Instructions

This exam has two portions – the ROT test and an additional questionnaire.

8. Hand out the instruction sheet and scantrons first.
   ➢ Try to make sure that there is at least one seat between each student and that there are no notes, books, etc. out and open during the test. In some rooms students may have to sit closer together. If so you will need to stand near them during the exam.

9. Make your announcements.

10. Hand out the ROT test booklet.
   ➢ Each student receives only one test booklet.

11. Once all students have received the test booklet, tell them to read the directions in the first 2 pages of the test booklet. Tell them "you are NOT allowed to proceed to the 3rd page with the real test questions until I tell you so".
   ➢ Give students 3 minutes to read the directions (in the first 2 pages of the booklet). Do NOT allow any student to proceed to the 3rd page; stop them if they try to!

12. Three (3) minutes later, tell them to proceed to the 3rd page; immediately start the 10-minute “clock” for the ROT test.
   ➢ Students are not allowed to talk during the exam. Even if they finish before the 10 minutes has passed, they must sit quietly.

13. When a student completes the ROT test portion, take **only the ROT test booklet and give the student the additional set of questions**. The students will use the original scantron to answer the additional questionnaire on. Check the ROT test booklet to make sure it has not been written on – if it has been written on, set it aside.
   ➢ After the 10 minute time period, students may no longer work on the ROT test and you must collect all booklets and distribute the additional questions. Make additional announcement at this time.

14. At the end of the whole exam (by 6:05 pm), collect everything.
   ➢ **Students may not take any paper used during the exam out of the room.**

Make these announcements to your students:

7. "Make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U** ID number without the U, and your course section number. It is NOT enough to write your name or other information down, you must **BUBBLE IN** the information. The scantron machine does NOT read handwriting; it reads bubbles."

8. No one can turn in the ROT test until 10 minutes has passed.

9. No one can receive a test after 10 minutes of the test being distributed.
10. No one may leave the exam room before 30 minutes has passed.

11. **Here are some specific announcements pertaining to the ROT test:**
   - "You MAY NOT WRITE on the ROT test. The instruction sheet should be used as scratch paper if needed."
   - "The first two pages in the test booklet are directions with two examples. You will have 3 minutes to read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until I tell you so."
   - "The ROT test questions starts on the 3rd page and they are numbered from 1 to 20. You will have 10 minutes to complete the ROT test."
   - "At the end of 10 minutes, the ROT test booklet (but not the scantron) will be collected and the additional questionnaire distributed."
   - "At the end of the exam (by 6:05 pm), you must turn in everything (including the additional questionnaire, the scratch paper AND the scantron)."

12. **Here are some specific announcements pertaining to the additional instrument:**
   - The additional questions begin at question number 21. Make sure you begin bubbling in your scantron at question number 21 for these questions. There will be 47 additional questions (numbered 21 through 67).
   - Please answer each question honestly. Your answers are confidential. Your instructor is not going to know your individual results.

**Information for Proctors After the Test:**

4. Check photo ID while collecting the scantrons. Check to make sure that the scantron is filled out correctly in pencil and includes UID. This may take more time than just picking up the scantron, but the students can wait. It is very important that you check to make sure everything is completely bubbled in.

5. Collect everything (including scratch paper and additional questionnaire). This is to protect the confidentiality of the exam and prevent potential lawsuits.

6. Return the collected scantrons and test materials to ROOM *** immediately following the exam.

It is your job to make sure that the scantrons are filled out correctly (i.e. everything bubbled in, filled out in pencil, etc.) and to have them in a neat and organized manner.
Instruction Sheet for Students

10) Before the test begins, make sure you completely fill out your scantron in pencil not pen. This includes bubbling in your name, U** ID number without the U, and your course section number.

- It is not enough to write your name or other information down, you must **BUBBLE IN** the information with a pencil. The scantron machine does not read handwriting; it reads bubbles.

11) First you will receive the ROT test portion of the exam. You will have 10 minutes to work on this portion.

12) **Do not write on the test booklet.** You may use this instruction sheet as scratch paper if needed.

13) The first two pages in the test booklet are directions with two examples. Read them carefully to understand the examples. Do NOT proceed to the 3rd page with the real test questions until told by your exam proctor.

14) The real test questions are numbered 1 through 20. Bubble in your answers to these questions on the scantron using a pencil (not pen).

- This test is intentionally timed so that you must rely on visualization instead of reasoning. Just visualize (picture the rotations of 3-D objects in your head). Do NOT try to reason it out; there is NOT enough time.

15) When time is called, stop working. You will turn in the ROT test booklet and receive an additional set of questions.

16) You will have the remainder of the time to answer the additional questionnaire.

17) Additional questions begin at question number 21. Make sure you begin bubbling in your scantron at question number 21 for these questions. There will be 47 additional questions (numbered 21 through 67).

18) When finished, hand in everything. **You may not take any paper used during the exam out of the exam room.**
Proctoring Instructions  01/20/10
Arrive at the exam room at least 15 minutes early. The exam formally begins at 7:35, so you should be in the room no later than 7:20. You have several tasks to accomplish before you can hand out the exams.

Task One: Figure out how to display a clock
You will need a dry-erase marker. (most likely)
If you can use the computer projection system to display a clock, terrific. If not, you need to write the time on the board every ~10 minutes for a normal exam, or ~5 minutes for an exam that is shorter. Regardless, right after you start the exam (or a portion of the exam) you need to write the ending time on the board so that students can compare it with the clock time.

Task Two: Check U**-IDs, hand out scantrons, and seat students
You will need a list of all of the students who are supposed to be in the room, scantrons, and confidence.
As a student comes in the door, ask to see her/his U**-ID. Find his/her name on the list, and check it off. Hand him/her a scantron and say that it should be filled out with name, U**-ID, and section number in pencil. Direct the student to a seat so that there will be one empty seat between students. If a student has a disability, allow her/him to sit where works best to accommodate their needs, but otherwise, YOU decide where each student sits. DO NOT START THE EXAM UNTIL YOU HAVE CHECKED the IDs of all students present.

Task Three: Announcements
Tell the students the exam will be in three parts, each harder than the next, and they will get credit for trying. (We reserve the right to not give credit for extremely low scores.) Tell them that getting credit also depends on filling in their scantron correctly IN PENCIL with NAME, U**-ID, and SECTION NUMBER. They should also sign and date their scantrons. Ask students to silence and put away all cell phones. Cell phones may not be used as calculators for any exam. For this exam, they are not allowed to have any calculator at all.

PART 1: CLASS-Chemistry
When you hand this first part of the exam out, tell students to read the instructions and not turn the page until instructed by you. Tell them that they should not write on the exam, that they will have 20 minutes to complete it, and that when they have finished, they should sit quietly. Once you have instructed that they turn the page, don’t forget Task One about the clock. At the end of 20 minutes, call time. Collect the exam by having them hand them forward or across. They should KEEP their scantrons.

PART 2: Chemistry Questions
When you hand this second part of the exam out, tell students to read the instructions and not turn the page until instructed by you. Tell them that they should not write on the exam, that they will have 35 minutes to complete it, and that when they have finished, they should sit quietly. Remind them that these questions start with number 18, the first blank line on their scantron. Once you have instructed that they turn the page, don’t forget Task One about the clock. At the
end of 35 minutes, call time. Collect the exam by having them hand them forward or across. After you have the exam, **COLLECT the scantrons** in the same way.

**PART 3: Rotations**

**HAND OUT SCANTRONS.** Instruct students to fill out their name, U**-ID, and section number (AGAIN). Give them time to do this before handing out the third part of the exam.

When you hand the third part of the exam out, tell students to read the instructions and not turn the page until instructed by you. Tell them that **THIS EXAM IS QUITE DIFFICULT FOR MANY PEOPLE, EVEN MANY GOOD CHEMISTS.** Remind them that they should **answer the questions by trying to picture the objects in their mind,** not by reasoning it out. Tell them that they will need to complete 20 questions in 10 minutes, so there is not time to use reasoning.

Once you have instructed that they turn the page, don’t forget Task One about the clock. Give a 5 minute and a 2 minute verbal warning (“You have two minutes remaining”). At the end of 10 minutes, call time. Collect the exam by having them hand them forward or across. After you have the exam, **collect the scantrons** in the same way.

**Task Four: RETURNING EVERYTHING NEATLY**

Separate the exams into “written on” and “not written on” stacks. Place both stacks in the box. Separate the scantrons into neatly organized Parts 1&2 and Part 3 stacks, ready for scanning. Check that name, U**-ID, and section number have been bubbled in for each scantron. If not, place the problem scantrons at the top of the correct stack and notify *** when you drop off the box, Place both stacks in the box.

Bring the box to ROOM *** when the exam is over. Be prepared to check in with *** or ***, showing them what is in the box. You may not leave until you have been checked in.

**Contact 1’s cell phone:** ********

**Contact 2’s cell phone:** ********

**Contact 1’s cell phone:** ********
Appendix B: Administration instruction for ROT

Spring 2009 ASCIv2

Instructions
(3/26/2009)

When students take the survey, make these announcements.

- “This short survey takes at most 5 minutes to complete.”

- “Your answers to this survey are very important to help us understand how students at U** view chemistry. Please respond to each item honestly and about the subject of chemistry itself, not this particular chemistry course or instructor. Your answers will be kept confidential and there are no wrong answers.”

- "Make sure you circle your answer choices for the survey on paper first. When you are finished with all 8 items, then you can record your answers on the scantron. Each item has 7 likert scale, so your answer on the scantron should range from “1” to “7” or from “B” to “H”. When you bubble your 8-digit U** ID on the scantron in the “IDENTIFICATION NUMBER” box, start from column “A” through column “H”. Make sure to bubble in your name and U**-ID on the scantron to get credit.”
A list of opposing words appears below. Rate how well these words describe your feelings about chemistry. Think carefully and **try not to include** your feelings toward chemistry teachers or chemistry courses. For each line, choose a position between the two words that describes **exactly how you feel.** Circle that number on this sheet. The middle position is if you are undecided or have feelings related to the terms on that line. When you are done with all 8 items, also record your response on the scantron. (B=1, C=2, D=3, E=4, F=5, G=6, H=7) **Make sure to bubble in your name and USF ID.**

**CHEMISTRY IS**

<table>
<thead>
<tr>
<th></th>
<th>easy</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>hard</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>chaotic</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>organized</td>
</tr>
<tr>
<td>3</td>
<td>confusing</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>clear</td>
</tr>
<tr>
<td>4</td>
<td>comfortable</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>uncomfortable</td>
</tr>
<tr>
<td>5</td>
<td>satisfying</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>frustrating</td>
</tr>
<tr>
<td>6</td>
<td>challenging</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>not challenging</td>
</tr>
<tr>
<td>7</td>
<td>pleasant</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>unpleasant</td>
</tr>
<tr>
<td>8</td>
<td>complicated</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>simple</td>
</tr>
</tbody>
</table>

171
Arrive at the exam room at least 15 minutes early. The exam formally begins at 7:35pm, so you should be in the room no later than 7:20pm. All the exams (except the final) are scheduled from 7:35pm to 9:00 pm. You have several tasks to accomplish before you can hand out the exams.

Task One: Figure out how to display a clock
You will need a dry-erase marker. (most likely)
If you can use the computer projection system to display a clock, terrific. If not, you need to write the time on the board every ~10 minutes for a normal exam, or ~5 minutes for an exam that is shorter. Regardless, right after you start the exam (or a portion of the exam) you need to write the ending time on the board so that students can compare it with the clock time.

Task Two: Check U**-IDs, hand out scantrons, and seat students
You will need a list of all of the students who are supposed to be in the room, scantrons, and exams. This list will be placed in the box.
As a student comes in the door, ask to see her/his U**-ID. Find his/her name on the list, and check it off. Hand him/her a scantron and say that it should be filled out with name, U**-ID, and section number in pencil. Direct the student to a seat so that there will be one empty seat between students. If a student has a disability, allow her/him to sit where works best to accommodate their needs, but otherwise, YOU decide where each student sits. DO NOT START THE EXAM UNTIL YOU HAVE CHECKED the IDs of all students present. If a student does not have a U** ID, ask him/her to show you any photo ID. If a student does not have any photo ID, ask him/her to write their name on the student list and sign it. Tell them that they will NOT be able to take the exam next time without proper identification. If a student is in the wrong room, and there is enough time for him to go to the correct room, give them the correct room number (provided in the box). If there is no time for a student to go to the correct room, ask the student to write their name on the student’s list and sign it.

Task Three: Announcements (It is very important that you announce this to the students)
- Announce that the exam will be in three parts. There will be two surveys and one exam, and they will get credit for trying. (We reserve the right to not give credit for extremely low scores.).
- Tell the students that they will be filling out two scantrons. The first scantron will be used for Part 1, and the second scantron will be used for Part 2 and Part 3.
- Tell students that their survey information will be kept confidential, and their exam information will be shared with their instructor for course improvement.
- Tell them that getting credit for their work depends on filling in their scantron correctly IN PENCIL with NAME, U**-ID, and SECTION NUMBER. They should also sign and date their scantrons.
- Finally, ask students to silence and put away all cell phones. Cell phones may not be used as calculators for any exam.
PART 1: ASCI-Survey (little paper) NOTE: Use the larger scantron for this part

When you hand this first part of the exam, tell students to read the instructions. Tell them that they should not write on the exam, that they will have 10 minutes to complete it, and that when they have finished, they should sit quietly. Once you have instructed to start answering the first part, don’t forget Task One about the clock. At the end of 10 minutes, call time. Collect the exam by having them hand them forward or across. After you have the exam, COLLECT the scantrons in the same way.

PART 2: Science-Related Attitudes Survey

HAND OUT SCANTRONS. Instruct students to fill out their scantron correctly IN PENCIL with NAME, U**-ID, and SECTION NUMBER. Give them time to do this before handing out the second part of the exam. When you hand this second part out, tell students to read the instructions. Tell them that they should not write on the exam, that they will have 15 minutes to complete it, and that when they have finished, they should sit quietly. Once you have instructed to start answering the second part, don’t forget Task One about the clock. At the end of 10 minutes, call time. Collect the survey by having them hand them forward or across. They should KEEP their scantrons.

PART 3: Chemistry Content Questionnaire

When you hand the third part of the exam out, tell students to read the instructions. Remind them that these questions start with number 31. Once you have instructed to start the exam, don’t forget Task One about the clock. Give students 40 minutes to complete this part. At the end of 40 minutes, call time. Collect the exam by having them hand them forward or across. After you have the exam, collect the scantrons in the same way.

Task Four: RETURNING EVERYTHING NEATLY

Separate the exams into “written on” and “not written on” stacks. Place both stacks in the box. Separate the scantrons into neatly organized Parts 1 & 2 and Part 3 stacks, ready for scanning. Check that name, U**-ID, and section number have been bubbled in for each scantron. If not, place the problem scantrons at the top of the correct stack and notify *** when you drop off the box. Place both stacks in the box. Keep the students’ list in the box and return it to Keily Heredia.

Bring the box to SCA431 when the exam is over. Be prepared to check in with *** or ***, showing them what is in the box. You may not leave until you have been checked in.

Contact 1’s cell phone: ********

Contact 1’s cell phone: ********
### Appendix C: Supplement for Chapter 3

**Table A3.1.**

*Item Difficulty, Item-Total Correlation, and Discrimination Index by Semester*

<table>
<thead>
<tr>
<th></th>
<th>Fall 2008 (n = 1396)</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S8</td>
<td>S9</td>
<td>S10</td>
</tr>
<tr>
<td>Mean</td>
<td>0.85</td>
<td>0.82</td>
<td>0.86</td>
<td>0.90</td>
<td>0.82</td>
<td>0.79</td>
<td>0.80</td>
<td>0.64</td>
<td>0.54</td>
<td>0.55</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.36</td>
<td>0.43</td>
<td>0.32</td>
<td>0.32</td>
<td>0.42</td>
<td>0.34</td>
<td>0.33</td>
<td>0.44</td>
<td>0.35</td>
<td>0.28</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.28</td>
<td>0.39</td>
<td>0.25</td>
<td>0.20</td>
<td>0.38</td>
<td>0.37</td>
<td>0.37</td>
<td>0.58</td>
<td>0.54</td>
<td>0.45</td>
</tr>
<tr>
<td></td>
<td>S11</td>
<td>S12</td>
<td>S13</td>
<td>S14</td>
<td>S15</td>
<td>S16</td>
<td>S17</td>
<td>S18</td>
<td>S19</td>
<td>S20</td>
</tr>
<tr>
<td>Mean</td>
<td>0.66</td>
<td>0.70</td>
<td>0.66</td>
<td>0.71</td>
<td>0.62</td>
<td>0.59</td>
<td>0.5</td>
<td>0.41</td>
<td>0.43</td>
<td>0.29</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.37</td>
<td>0.37</td>
<td>0.44</td>
<td>0.45</td>
<td>0.44</td>
<td>0.42</td>
<td>0.35</td>
<td>0.27</td>
<td>0.3</td>
<td>0.26</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.51</td>
<td>0.48</td>
<td>0.58</td>
<td>0.56</td>
<td>0.63</td>
<td>0.61</td>
<td>0.55</td>
<td>0.43</td>
<td>0.49</td>
<td>0.39</td>
</tr>
<tr>
<td>Spring 2009 (n = 690)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S8</td>
<td>S9</td>
<td>S10</td>
</tr>
<tr>
<td>Mean</td>
<td>0.84</td>
<td>0.81</td>
<td>0.83</td>
<td>0.87</td>
<td>0.77</td>
<td>0.75</td>
<td>0.77</td>
<td>0.58</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.29</td>
<td>0.36</td>
<td>0.3</td>
<td>0.34</td>
<td>0.35</td>
<td>0.33</td>
<td>0.35</td>
<td>0.44</td>
<td>0.37</td>
<td>0.3</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.27</td>
<td>0.35</td>
<td>0.29</td>
<td>0.28</td>
<td>0.38</td>
<td>0.37</td>
<td>0.39</td>
<td>0.62</td>
<td>0.57</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>S11</td>
<td>S12</td>
<td>S13</td>
<td>S14</td>
<td>S15</td>
<td>S16</td>
<td>S17</td>
<td>S18</td>
<td>S19</td>
<td>S20</td>
</tr>
<tr>
<td>Mean</td>
<td>0.61</td>
<td>0.62</td>
<td>0.6</td>
<td>0.62</td>
<td>0.58</td>
<td>0.56</td>
<td>0.46</td>
<td>0.34</td>
<td>0.38</td>
<td>0.27</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.34</td>
<td>0.35</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.42</td>
<td>0.32</td>
<td>0.23</td>
<td>0.32</td>
<td>0.28</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.50</td>
<td>0.50</td>
<td>0.63</td>
<td>0.59</td>
<td>0.64</td>
<td>0.62</td>
<td>0.50</td>
<td>0.37</td>
<td>0.48</td>
<td>0.36</td>
</tr>
<tr>
<td>Fall 2009 (n = 1404)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S8</td>
<td>S9</td>
<td>S10</td>
</tr>
<tr>
<td>Mean</td>
<td>0.84</td>
<td>0.84</td>
<td>0.85</td>
<td>0.88</td>
<td>0.8</td>
<td>0.76</td>
<td>0.79</td>
<td>0.64</td>
<td>0.5</td>
<td>0.52</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.31</td>
<td>0.37</td>
<td>0.3</td>
<td>0.31</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.42</td>
<td>0.41</td>
<td>0.2</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.32</td>
<td>0.37</td>
<td>0.29</td>
<td>0.27</td>
<td>0.39</td>
<td>0.47</td>
<td>0.44</td>
<td>0.59</td>
<td>0.69</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>S11</td>
<td>S12</td>
<td>S13</td>
<td>S14</td>
<td>S15</td>
<td>S16</td>
<td>S17</td>
<td>S18</td>
<td>S19</td>
<td>S20</td>
</tr>
<tr>
<td>Mean</td>
<td>0.60</td>
<td>0.64</td>
<td>0.61</td>
<td>0.66</td>
<td>0.59</td>
<td>0.58</td>
<td>0.5</td>
<td>0.39</td>
<td>0.42</td>
<td>0.29</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.34</td>
<td>0.39</td>
<td>0.44</td>
<td>0.44</td>
<td>0.34</td>
<td>0.38</td>
<td>0.33</td>
<td>0.24</td>
<td>0.28</td>
<td>0.25</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.56</td>
<td>0.58</td>
<td>0.63</td>
<td>0.65</td>
<td>0.56</td>
<td>0.59</td>
<td>0.56</td>
<td>0.43</td>
<td>0.51</td>
<td>0.41</td>
</tr>
<tr>
<td>Spring 2010 (n = 678)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>S1</td>
<td>S2</td>
<td>S3</td>
<td>S4</td>
<td>S5</td>
<td>S6</td>
<td>S7</td>
<td>S8</td>
<td>S9</td>
<td>S10</td>
</tr>
<tr>
<td>Mean</td>
<td>0.81</td>
<td>0.77</td>
<td>0.78</td>
<td>0.83</td>
<td>0.7</td>
<td>0.73</td>
<td>0.73</td>
<td>0.57</td>
<td>0.48</td>
<td>0.47</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.34</td>
<td>0.47</td>
<td>0.36</td>
<td>0.36</td>
<td>0.38</td>
<td>0.34</td>
<td>0.33</td>
<td>0.45</td>
<td>0.31</td>
<td>0.24</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.32</td>
<td>0.37</td>
<td>0.29</td>
<td>0.27</td>
<td>0.39</td>
<td>0.47</td>
<td>0.44</td>
<td>0.59</td>
<td>0.69</td>
<td>0.41</td>
</tr>
<tr>
<td></td>
<td>S11</td>
<td>S12</td>
<td>S13</td>
<td>S14</td>
<td>S15</td>
<td>S16</td>
<td>S17</td>
<td>S18</td>
<td>S19</td>
<td>S20</td>
</tr>
<tr>
<td>Mean</td>
<td>0.59</td>
<td>0.61</td>
<td>0.56</td>
<td>0.62</td>
<td>0.54</td>
<td>0.56</td>
<td>0.47</td>
<td>0.40</td>
<td>0.41</td>
<td>0.25</td>
</tr>
<tr>
<td>Item-total Corr</td>
<td>0.23</td>
<td>0.44</td>
<td>0.45</td>
<td>0.43</td>
<td>0.39</td>
<td>0.42</td>
<td>0.27</td>
<td>0.24</td>
<td>0.23</td>
<td>0.20</td>
</tr>
<tr>
<td>D-Index</td>
<td>0.40</td>
<td>0.64</td>
<td>0.66</td>
<td>0.61</td>
<td>0.61</td>
<td>0.62</td>
<td>0.48</td>
<td>0.42</td>
<td>0.39</td>
<td>0.29</td>
</tr>
</tbody>
</table>
Appendix D: Supplement for Chapter 4

As shown from Table 4.1A, all variables exhibited normal distribution with skewness and kurtosis less than 1.2. The SAT math score is 601.6 out of 800 for our sample, ROT is 13.1 out of 20, ACS final exam is 28.4 out of 40, course grade is 2.71 for organic chemistry I and 2.7 for organic chemistry II (closed to letter grade C).

Table A4.1.

Descriptive Analysis for SEM Variables

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>Mean</th>
<th>Std Dev</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>SATM</td>
<td>3259</td>
<td>575.67</td>
<td>77.29</td>
<td>0.03</td>
<td>-0.10</td>
</tr>
<tr>
<td>ROT</td>
<td>3834</td>
<td>12.72</td>
<td>4.04</td>
<td>-0.39</td>
<td>-0.41</td>
</tr>
<tr>
<td>ACS</td>
<td>3239</td>
<td>25.18</td>
<td>6.48</td>
<td>-0.09</td>
<td>-0.45</td>
</tr>
<tr>
<td>Grd Organic I</td>
<td>1431</td>
<td>2.48</td>
<td>1.04</td>
<td>-0.68</td>
<td>0.32</td>
</tr>
<tr>
<td>Grd Organic I</td>
<td>1076</td>
<td>2.69</td>
<td>0.92</td>
<td>-0.60</td>
<td>0.73</td>
</tr>
<tr>
<td>SATM</td>
<td>876</td>
<td>601.63</td>
<td>77.07</td>
<td>0.04</td>
<td>-0.22</td>
</tr>
<tr>
<td>ROT</td>
<td>876</td>
<td>13.08</td>
<td>3.87</td>
<td>-0.38</td>
<td>-0.50</td>
</tr>
<tr>
<td>ACS</td>
<td>876</td>
<td>28.37</td>
<td>5.83</td>
<td>-0.29</td>
<td>-0.40</td>
</tr>
<tr>
<td>Grd Organic I</td>
<td>876</td>
<td>2.83</td>
<td>0.72</td>
<td>0.28</td>
<td>-1.16</td>
</tr>
<tr>
<td>Grd Organic I</td>
<td>876</td>
<td>2.72</td>
<td>0.89</td>
<td>-0.51</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Figure A4.1 presents parameter estimates for alternate models. From the left graph when general chemistry is removed, mental rotations still does not show significant effect. The effect of mental rotations is significant when all other predictors are removed, which is consistent with literature.
Figure A4.1 Alternate models with general chemistry achievement is removed (left) and with only one predictor of mental rotations ability and all other predictors removed (right). This graph only shows the regression part and the measurement part is shown for clarity. The ovals represent latent variables, and the rectangles indicate observed variables. Single headed arrow designates a causal relationship (from cause to effect). Double headed arrow defines a relationship without causality interpretation. The numbers on the line are path or correlation coefficients.
Appendix E: Supplement for Chapter 7

Figure Notes for Figures 7.2-7.6

The ovals represent latent variables, the rectangles indicate observed variables, and the small circles are residuals. Single headed arrows designate causal relationship (from cause to effect). Double headed arrows can be correlation/covariance/variance without causality interpretation. The numbers on the line can be path or correlation/covariance coefficients.

Additional analysis

In addition to the analysis in supplement B to examine the relationship of attitude toward chemistry, conceptual knowledge and chemistry achievement, we did two more analyses. One is to provide psychometric evidence with a focus on measurement invariance for instrument ASCIv2 in the similar way with we did for ROT. The other is to examine the unique contribution of spatial ability, as measured by ROT, when attitude is also included.

Measurement Invariance for ASCIV2 to Measure Attitude toward Chemistry

Our previous work as part of my master thesis elaborated on the refinement and evaluation of ASCIv2 (Xu & Lewis, 2011). Here, I focused on the strong measurement invariance test for using ASCIv2 for group comparison in future studies using the multiple-CFA approach (Raykov, Marcoulides, & Li, 2012).

The latent variable modeling (LVM) was widely used for test measurement invariance in psychology and educational research (Muthén, 2002). However, this traditional procedure has the limitation, e.g., it cannot test the invariance for the anchor items because all loadings of the first item for each factor each group are fixed to be 1 for model identification purpose. Raykov etc (2012) proposed a more straightforward approach, which was used for this ASCIv2 data set. First the 2-factor model was run within the females and males separately. The model fit for the females is: $\chi^2 (n = 504, df = 17, p < .001) = 110.7$, CFI=.95, SRMR=0.04, which indicates a
reasonable fit. The model fit for the males is: \( \chi^2 (n = 459, df = 17, p < .001) = 117.3, \text{CFI}=.94, \text{SRMR}=0.04 \), which indicates a reasonable fit. This suggests the configural invariance across sexes holds.

The 2-factor model of ASCIv2 with eight items grouped within two subscales was used to examine strong invariance between the sexes. Alternate Fit indices such as chi-square (\( \chi^2 \)), Comparative Fit Index (CFI), and the Standardized Root Mean Square Residual (SRMR) were used to assess the overall fitness of the model to data. The criteria of CFI value greater than 0.95 and SRMR value less than 0.05 were used to indicate a good model fit, while CFI > .9, and SRMR < .08 indicates an acceptable fit (Bentler, 1990; Hu & Bentler, 1995).

Table A7.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>( \chi^2 )</th>
<th>df</th>
<th>( p )</th>
<th>CFI</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>CFA for females only</td>
<td>110.7</td>
<td>17</td>
<td>&lt;.001</td>
<td>.95</td>
<td>0.042</td>
</tr>
<tr>
<td>CFA for males only</td>
<td>117.3</td>
<td>17</td>
<td>&lt;.001</td>
<td>.94</td>
<td>0.047</td>
</tr>
<tr>
<td>Strong invariance</td>
<td>266.5</td>
<td>50</td>
<td>&lt;.001</td>
<td>.94</td>
<td>0.056</td>
</tr>
</tbody>
</table>

Then a model was run for strong invariance, which constrains all intercepts and factor loadings to be equal across groups but not at 1. The model result is: \( \chi^2 (n = 963, df = 50, p < .001) = 266.5, \text{CFI}=.94, \text{SRMR}=0.056 \), which indicates strong invariance holds. Accordingly, sex difference can be compared using ASCIv2 in future studies without much concern of test bias.

**Unique Contribution of Spatial Ability When Attitude is Controlled**

Multiple regression analysis was performed to predict ACS score using SAT math, spatial ability, sex (0 for male and 1 for female), intellectual accessibility and emotional satisfaction subscales of attitude as predictors for 291 students who had complete data during the Spring 2009. The obtained \( R^2 \) value was .314, suggesting about 31.4% of the variance in the
ACS score was accounted by the set of predictors. The adjusted $R^2$ was .302. It appears that an acceptable proportion of the variability was accounted for. The root mean square error was 11.2, which indicated that predictions of ACS tended to be off by about 11.2.

The obtained prediction equation is:

$$\text{Predicted ACS} = -3.93 + 0.04\times\text{SATM} + 0.09\times\text{ROT} + 0.30\times\text{SEX} + 0.26\times\text{ACCESS} + 1.05\times\text{EMOTION}$$

The standardized form of prediction is:

$$\text{Predicted ACS} = 0.44\times\text{SATM (s.s.)} + 0.06\times\text{ROT (n.s.)} + 0.02\times\text{SEX (n.s.)} + 0.05\times\text{ACCESS (n.s.)} + 0.21\times\text{EMOTION (s.s.)}$$

Thus, if we consider students who have the same SAT math, sex and attitude score, those with a 1 point higher ROT would be predicted to have a 0.09 point higher in ACS. If we consider those with the same score other predictors, students that score 1 higher on emotional satisfaction will be predicted to have an ACS score that is 1.05 points higher. The regression coefficients for SAT and emotional satisfaction are statistically significant, whereas the regression coefficients for ROT, sex and intellectual accessibility are not. Squared semipartial correlation was also examined for each of the predictors. SAT math uniquely accounted for 16.2% of the variability in ACS, ROT for 0.28%, sex accounted for .05%, intellectual accessibility accounted for 0.1%, and emotional satisfaction uniquely accounted 2.5% of the variability in ACS. This suggested that SAT math and emotional satisfaction have unique contrition to the ACS, while contribution of ROT can be negligible.

**References for Additional Analysis**


