

Summer 8-10-2015

# The Design and Validation of a Group Theory Concept Inventory

Kathleen Mary Melhuish  
*Portland State University*

Let us know how access to this document benefits you.

Follow this and additional works at: [http://pdxscholar.library.pdx.edu/open\\_access\\_etds](http://pdxscholar.library.pdx.edu/open_access_etds)

 Part of the [Algebra Commons](#), [Higher Education Commons](#), and the [Science and Mathematics Education Commons](#)

---

## Recommended Citation

Melhuish, Kathleen Mary, "The Design and Validation of a Group Theory Concept Inventory" (2015). *Dissertations and Theses*. Paper 2490.

[10.15760/etd.2487](https://pdxscholar.library.pdx.edu/open_access_etds/10.15760/etd.2487)

This Dissertation is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.edu).

The Design and Validation of a Group Theory Concept Inventory

by

Kathleen Mary Melhuish

A dissertation submitted in partial fulfillment of the  
requirements for the degree of

Doctor of Philosophy  
in  
Mathematics Education

Dissertation Committee:  
Sean Larsen, Chair  
John Caughman  
Samuel Cook  
Andrea Goforth

Portland State University  
2015

© 2015 Kathleen Mary Melhuish

## Abstract

Within undergraduate mathematics education, there are few validated instruments designed for large-scale usage. The Group Concept Inventory (GCI) was created as an instrument to evaluate student conceptions related to introductory group theory topics. The inventory was created in three phases: domain analysis, question creation, and field-testing. The domain analysis phase included using an expert protocol to arrive at the topics to be assessed, analyzing curriculum, and reviewing literature. From this analysis, items were created, evaluated, and field-tested. First, 383 students answered open-ended versions of the question set. The questions were converted to multiple-choice format from these responses and disseminated to an additional 476 students over two rounds. Through follow-up interviews intended for validation, and test analysis processes, the questions were refined to best target conceptions and strengthen validity measures. The GCI consists of seventeen questions, each targeting a different concept in introductory group theory. The results from this study are broken into three papers. The first paper reports on the methodology for creating the GCI with the goal of providing a model for building valid concept inventories. The second paper provides replication results and critiques of previous studies by leveraging three GCI questions (on cyclic groups, subgroups, and isomorphism) that have been adapted from prior studies. The final paper introduces the GCI for use by instructors and mathematics departments with emphasis on how it can be leveraged to investigate their students' understanding of group theory concepts. Through careful creation and extensive field-testing, the GCI has been shown to

be a meaningful instrument with powerful ability to explore student understanding around group theory concepts at the large-scale.

## Dedication

To my parents for teaching me to have a strong work ethic, to live life with kindness, and to pursue a career that truly fulfills you.

## Acknowledgements

I cannot begin to thank all of the people that have helped me complete this dissertation. First, and foremost, I should thank my advisor, Sean Larsen. He always pushed me to do better. He sets a high bar for his students, and I am grateful for that and the continued help, insight, and whiskey he provided. (Sean, that oxford comma was for you!) I also want to thank the other members of my committee, each of whom brought a different expertise that helped this project become what it is. Sam Cook provided valuable insight into the statistics and quantitative analyses. John Caughman brought the view of an experienced mathematics instructor and provided generous feedback on the practitioner paper. I also must thank Andrea Goforth who stepped out of the chemistry world to patiently provide support as the outside committee member.

There are nearly a hundred current and former abstract algebra instructors who have helped me at some point in my process between the Delphi study, instructors who have given their classes the GCI, instructors that have tested GCI themselves, and those who have evaluated tasks. I especially want to thank Annie Selden who provided feedback on every one of my tasks. I also want to thank Tim Fukawa-Connelly, Keith Weber, Dorothy Zeiser, and Andrew Conner who all evaluated over twenty tasks.

Perhaps, I most need to acknowledge the people who helped me in small or large ways in a less formal manner. Carolyn James, Andy Riffel, and Dana Kirin have all been kind enough to read drafts of the papers. Annie Bergman aided me in a great deal of analysis throughout the process. I also could not have

done this without the support of my wife, Simone, who supported me through this entire stressful process. Finally, I need to thank Jodi Fasteen, my partner-in-crime this year. She has helped me during every step of the process and patiently waded through all of my papers. I had always thought of the dissertation as being on an island, but it so far from it. I have never been more grateful for all of the help, support and kindness that my friends, family, and colleagues provided.



## Table of Contents

Abstract .....	i
Dedication .....	iii
Acknowledgements .....	iv
List of Tables .....	viii
List of Figures .....	x
<b>Introduction and Rationale.....</b>	<b>1</b>
Importance of Assessing Concept Understanding .....	3
Purpose of Concept Inventories .....	7
<b>Abstract Algebra Literature Review.....</b>	<b>10</b>
Student Understanding in Abstract Algebra .....	11
Instruction of Abstract Algebra: Innovations .....	28
Conclusion .....	34
<b>Concept Inventories and Theoretical Background.....</b>	<b>36</b>
Concept Inventories .....	36
Conceptual Knowledge .....	37
Fundamental Topics in Group Theory .....	49
Assessments .....	52
Validity and Reliability .....	55
Conclusion .....	67
<b>Project Methodology.....</b>	<b>68</b>
Concept Domain Analysis .....	69
The Creation of the Assessment .....	82
Conclusion .....	88
<b>Paper 1: Going from nothing to something: A Multi-faceted approach to developing questions for concept inventories as applied to introductory group theory .....</b>	<b>90</b>
On Validity and Concept Inventories .....	93
The Methodology .....	101
Question Creation and Selection .....	120
Question Refinement and Field Testing .....	127
Psychometric Properties of the GCI .....	141
Conclusion .....	145
<b>Paper 2: Scaling up and out: Reconsidering how students think about concepts in Group theory .....</b>	<b>149</b>
The Need for Replication .....	150
Student Understanding in Group Theory .....	152
Methods .....	153
Replication Results and Discussion .....	156
Conclusion .....	191

<b>Paper 3: A tool for unearthing student conceptions in group theory .....</b>	<b>193</b>
Why Should We Care about Student Understanding of Concepts in Proof-Focused Courses?.....	195
How was the GCI Created?.....	199
How Can the GCI be Utilized? .....	201
Sample Items.....	203
Companion Website.....	206
Concluding Remarks.....	207
<b>Conclusion .....</b>	<b>209</b>
Directions For Future Research .....	211
References.....	<b>214</b>
Appendix A Textbook Analysis Codes.....	224
Appendix B Sample Analytic Report from Textbook Analysis .....	228
Appendix C Sample Literature Review Report .....	232
Appendix D Expert Evaluation Summary .....	233
Appendix E Item Analysis .....	234
Appendix F Principal Component Analysis - % of Variance .....	235
Appendix G Distractor Frequencies.....	236

## List of Tables

Table 1	<i>Cognitive Behavior Associated with Calculus Exam Items. Adapted from “A characterization of Calculus I final exams in US colleges and universities,” M.A. Tallman and M.P. Carlson, 2012, The 15th annual conference on Research in Collegiate Mathematics Education, p. 2-222.</i>	45
Table 2	<i>Item Classification Guide by Difficulty and Discrimination. Adapted from “Assessment in Health Professions Education,” S. M. Downing &amp; R. Yudkowsky, 2009. New York, NY: Routledge, p. 108.</i>	66
Table 3	<i>Timeline of GCI Development</i>	69
Table 4	<i>Abstract Algebra Textbook Analyzed</i>	74
Table 5	<i>Important Topics in Group Theory Selected Through a Delphi Study</i>	104
Table 6	<i>Representation Codes</i>	108
Table 7	<i>Percentages of Representations in Textbook Examples and Exercises</i>	108
Table 8	<i>Example Purpose Codes</i>	109
Table 9	<i>Types of Definitions in Textbooks</i>	113
Table 10	<i>Selectivity of Sample Institutions</i>	129
Table 11	<i>Open-ended Responses for Cyclic Question</i>	130
Table 12	<i>Open-ended responses for identity question</i>	133
Table 13	<i>Open-ended Responses for the Revised Quotient Group Question.</i>	139
Table 14	<i>Selectivity of Sample Institutions</i>	155
Table 15	<i>Previous Results: Is <math>Z_3</math> a Subgroup of <math>Z_6</math>? (Dubinsky et al., 1994); A Student Wrote in an Exam, ‘<math>Z_3</math> is a Subgroup of <math>Z_6</math>’. In Your Opinion is this Statement True, Partially True, or False? Please Explain Your Answer (Leron &amp; Hazzan, 1996).</i>	160
Table 16	<i>Open-Ended Round Results: Is <math>Z_3</math> a subgroup of <math>Z_6</math>?</i>	162
Table 17	<i>Closed-Form Pilot Results: Is the Set 0,1,2 a Subgroup of <math>Z_6</math>?</i>	162
Table 18	<i>Closed-Form Large-Scale Results: Is the Set 0,1,2 a Subgroup of <math>Z_6</math>?</i>	162
Table 19	<i>Results From Lajoie &amp; Mura’s (2000): For each of The Following Groups, Say Whether or Not It Is Cyclic and Justify Your Answer. (a) The Set <math>Z</math> under Addition.</i>	170
Table 20	<i>Results from Lajoie &amp; Mura’s (2000): The Following Remarks were Made by Students in an Algebra Course. For each One, Say Whether You Agree or Disagree and Explain Why.</i>	171
Table 21	<i>Open-ended Round Results: Is <math>Z</math>, the Set of Integers under Addition, a Cyclic Group?</i>	172
Table 22	<i>Closed-form Pilot Round: Is <math>Z</math>, the Set of Integers under Addition, a Cyclic Group</i>	172
Table 23	<i>Closed-form Large-Scale Round: Is <math>Z</math>, the Set of Integers under Addition, a Cyclic Group</i>	172
Table 24	<i>Weber &amp; Alcock’s (2004) Results: Prove or Disprove: <math>Q</math> is Isomorphic to <math>Z</math></i>	181

Table 25	<i>Open-Ended Survey Results: Are <math>\mathbb{Z}</math>, the Integers under Addition, and <math>\mathbb{Q}</math>, the Rationals under Addition, Isomorphic?</i>	185
Table 26	<i>Closed-Form Pilot Results Are <math>\mathbb{Z}</math>, the Integers under Addition, and <math>\mathbb{Q}</math>, the Rationals under Addition, Isomorphic?</i>	185
Table 27	<i>Closed-Form Large Round Results: Are <math>\mathbb{Z}</math>, the Integers under Addition, and <math>\mathbb{Q}</math>, the Rationals under Addition, Isomorphic?</i>	186

## List of Figures

- Figure 1.* How conceptual understanding interferes with proving. Reprinted from “Making the transition to formal proof,” by R.C. Moore, 1994, *Educational Studies in Mathematics*, 27, p. 252. Copyright 1994 by Kluwer Academic Publishers. 7
- Figure 2.* Conceptual coherence in physics. Reprinted from “The Force Concept Inventory as a measure of students conceptual coherence,” by A. Savininen and J. Viiri, 2008 *International Journal of Science & Mathematics Education*, 6, p. 722. Copyright 2007 by National Science Council. 39
- Figure 3.* An analytic framework for reasoning-and-proving in geometry textbooks. Reprinted from “The mathematical nature of reasoning-and-proving opportunities in geometry textbooks,” by S. Otten, N.J. Gilbertson, L.M. Males and D.L. Clark, 2014, *Mathematical Thinking and Learning*, 16, p. 58 Copyright 2014 by Taylor & Francis Group, LLC. 43
- Figure 4.* Equivalence classes represented by fibers. Adapted from “Mathematics undergraduates' responses to semantic abbreviations, ‘geometric’ images and multi-level abstractions in group theory,” by E. Nard, 2000, *Educational Studies in Mathematics*, 43, p. 175. Copyright 2001 by the Kluwer Academic Publishers. 46
- Figure 5.* A genetic decomposition of the group schema. Reprinted from APOS Theory: A framework for research and curriculum development in mathematics education by I. Arnon, 2013, Copyright 20013 by the Springer London. 48
- Figure 6.* The assessment triangle. Adapted from “Rigorous methodology for concept inventory development: Using the ‘assessment triangle’ to develop and test the Thermal and Transport Science Concept Inventory (TTCI),” by R.A. Streveler, R.L Miller, A. I. Santiago-Roman, M.A. Nelson, M.R. Geist and B.M. Olds, 2011, *International Journal of Engineering Education*, 27, p. 969. Copyright 2011 by TEMPUS Publications. 53
- Figure 7.* The target analogy for reliability (left) and validity (right). Adapted from Research Methods Knowledge Based by W. Trochim and J. P. Donnelly, 2007, Copyright 2008 by the Atomic Dog Publishing. 56
- Figure 8.* Scree plot. Factors vs Eigenvalues 61
- Figure 9.* Overall methodology. 68
- Figure 10.* Domain Analysis Triangle 69
- Figure 11.* Components of the domain. Reprinted from “Focus article: On the structure of educational assessments,” by R.J. MisLevy, L.S. Steinberg, and R.G. Almond, 2003, *Measurement: Interdisciplinary Research & Perspective*, 1, p. 6. Copyright 2003 by Lawrence Erlbaum Associates, Inc. 69

<i>Figure 12.</i> Example of formal and informal definition and descriptions of homomorphism from Fraleigh (2002).	76
<i>Figure 13.</i> Homomorphism exercise from Fraleigh (2002).	77
<i>Figure 14.</i> Homomorphism example from narrative in Gallian (2009).	77
<i>Figure 15.</i> Overview of GCI creation methodology.	101
<i>Figure 16.</i> Domain analysis triangle.	102
<i>Figure 17.</i> Percentages of Exercises with each type of the six most prevalent expected student activities.	111
<i>Figure 18.</i> Task creation and selection.	120
<i>Figure 19.</i> Open-ended (left) and multiple-choice (right) field testing.	127
<i>Figure 20.</i> Multiple-choice version of cyclic group question.	131
<i>Figure 21.</i> Open-ended version of the identity question.	132
<i>Figure 22.</i> Sample student work illustrating the problematic domain.	134
<i>Figure 23.</i> Multiple-choice question on identity.	134
<i>Figure 24.</i> Revised multiple-choice question on identity.	135
<i>Figure 25.</i> Original quotient group question.	136
<i>Figure 26.</i> Student with correct conception of elements in quotient groups.	137
<i>Figure 27.</i> Student with incorrect conception of quotient groups.	137
<i>Figure 28.</i> New open-ended quotient group question.	138
<i>Figure 29.</i> Scree plot illustrating the number of components in the GCI.	145
<i>Figure 30.</i> Bob's work on the quotient group question.	165
<i>Figure 31.</i> Student response to subgroup question with differing element notation.	167
<i>Figure 32.</i> Proof that having a solution for $x+x=b$ for all elements $b$ in a group is a structural property.	178
<i>Figure 33.</i> An ordering of the rationals to show they are countably infinite.	179
<i>Figure 34.</i> A bijective map from $\mathbb{Z}$ to $\mathbb{Q}$ .	179
<i>Figure 35.</i> Overview of GCI creation stages.	201
<i>Figure 36.</i> Question on Subgroups	204
<i>Figure 37.</i> Question of Quotient Groups	205
<i>Figure 38.</i> Group Concept Inventory instructor resource website	207

## **Introduction and Rationale**

Abstract algebra is a standard course required of mathematics and mathematics education majors across the United States. For many students, this course is the first time concepts are to be reasoned about formally based on their properties (Hazzan, 1999). Dubinsky, Dautermann, Leron and Zazkis (1994) noted, “mathematics faculty and students generally consider it to be one of the most troublesome undergraduate subjects” (p. 268). While this statement is largely accepted, student understanding in the course has not been empirically evaluated on a large scale.

When the Force Concept Inventory (FCI) was introduced into the physics community, the state of students’ conceptual understanding was similarly not understood (Hestenes, Wells, & Swackhamer, 1992). The inventory was created to evaluate students’ conceptions of force via a validated and quick multiple-choice test. As a result, the test became widely used at the collegiate level. A concept inventory distills concepts from procedural and other knowledge to probe at a deeper level than many traditional assessments. Results from the FCI showed that students might do procedures correctly while still maintaining fundamentally flawed understandings of Newtonian force.

Advanced undergraduate mathematics education finds itself where the physics community was decades ago when the FCI was created. There is a body of research related to student understanding about abstract algebra, but it remains small scale. In general, the research falls into the categories of student understanding and instructional innovation (Weber and Larsen, 2008). Currently

literature exists to describe mechanisms that could lead to concept acquisition (Dubinsky, 1997), student difficulties with proofs, abstraction and complexity (Dubinsky et al., 1994; Hazzan, 1999; Leron, Hazzan, & Zazkis, 1995; Selden & Selden, 1987; Weber & Alcock, 2004; Weber, 2001) and instructional innovations (Larsen, Johnson, & Weber, 2013; Dubinsky, 1997). However, these studies were primarily exploratory and often relied on in-depth interviews with small samples. Student conceptions in algebra have not been explored systematically or comprehensively. In fact, Larsen, Johnson, and Bartlo (2014) have recently called for the “creation of validated assessment instruments” (p. 709) in group theory for the purpose of evaluating instructional innovations.

Recently, Weber (2013) similarly called for an increase of quantitative studies in the Research in Undergraduate Mathematics Education (RUME) community noting that qualitative studies dominate the undergraduate portion of the mathematics education community. Ideally, quantitative and qualitative methods could be used in a complementary fashion to establish both meaningful and generalizable results. The Group Concept Inventory (GCI) was created to address this call. The GCI is a quantitative measure created through both qualitative methods (such as interviewing students) and quantitative methods (such as item analysis and reliability - see methods section.) Ultimately, the validated tool serves a powerful purpose to complement qualitative explorations and evaluate larger populations.



## **Importance of Assessing Concept Understanding**

The aim of the GCI is to assess conceptual understanding independent of proof skills. Proof is often the focus of assessment in advanced mathematics, ignoring the role of strong conceptual understanding. Attending to concepts is important for several reasons:

1. Understanding concepts in group theory is not trivial.
2. Understanding concepts is an essential part of advanced mathematics.
3. Understanding concepts is a necessary for a high degree of success in proving.

The small body of literature related to abstract algebra has unanimously documents the difficulty of complete conceptual understanding of various topics. Dubinsky (1997) and his colleagues established that students frequently lacked advanced conceptions of topics in group theory ranging from groups to normality and quotient groups. Hazzan (1999) found that students struggled with abstraction level when dealing with objects in algebra defined solely by their properties. Students might use only one element of a set to evaluate a class of objects, or students might substitute information about familiar groups, such as the real numbers, to reason about unfamiliar groups such as modular arithmetic groups. Leron, Hazzan and Zazkis (1995) documented the struggles students have when coordinating the complexity of isomorphism such as differentiating between the idea of isomorphic groups and an isomorphism map. For a more complete discussion of student understanding see the literature review chapter.

Concepts play a vital role both in the formal proving process and independent from it. While formal proof may dominate assessment at the

advanced level, it is not the only important activity. Raman (1999) warned, “competence in mathematics might readily be misperceived as synonymous with the ability to create the form, a rigorous proof” (p. 60). When Weber and Mejia-Ramos (2011) explored what mathematicians did when reading proofs, they found their activities were rarely centered on verifying the formal deduction in proofs. Rather, they were testing examples, determining the big ideas and considering methods for their own use. These activities all require an understanding of concepts, instantiations, and relationships between ideas.

Even if the focus of a course is entirely on producing formal proof, conceptual understanding can be a mitigating and necessary component in the creation of valid proofs (Tall & Vinner 1981; Moore, 1994; Hart, 1994). Through a series of interviews with both students and mathematicians, Weber and Alcock (2004) found that undergraduates often relied on purely syntactic (symbol-pushing) proving strategies. Their undergraduates were limited in their approaches to proof and were unable to provide valid proofs and correct evaluations a mathematical statement’s validity. In contrast, advanced graduate students would often utilize semantic strategies where they reasoned from instantiations. When prompted to assess whether groups were isomorphic, undergraduates were limited to trying to create a map and so failed to show groups were not isomorphic. Their graduate counterparts were able to reason about properties and easily determine when two groups were not isomorphic. Weber and Alcock identified several requirements related to concept understanding including instantiating rich and accurate reflections of “the object

and concepts they represent” (p. 229). Further students need to be able to connect the formal definitions with instantiations. Having the formal definition of group and isomorphism alone was not a sufficient condition for students to be able to either describe isomorphism informally or utilize intuition to explore conjectures.

Moore (1994) explained several ways that understanding of concepts becomes important when proving. He gathered data from observing a transition to proof course and interviewing students from the class. He found the students had a multitude of concept-based errors. These include not being able to generate examples, lacking an intuitive understanding, not being able to use concept images to create formal proofs, and not knowing the appropriate definitions (See Figure 1).

Melhuish and Larsen (2015) illustrated how understanding of a particular concept (that of function) might play into the proving process. Students were prompted to prove or disprove the following claim:

*Let  $\phi$  be a 1-1 homomorphism from  $(G, o)$  to  $(H, *)$ . If  $G$  is an abelian group, then  $H$  is an abelian group.*

In order to prove this statement, one must show that all elements in  $H$  commute.

The standard approach to the proof would be to begin by selecting arbitrary elements in  $H$ . However, Melhuish and Larsen found that many of the students began in  $G$ , ultimately failing to realize the claim was not true. During interviews, a student explained that functions must always map from the domain to range, and so starting in  $G$  was the proper way to begin the proof. This piece of their concept image seemed to be interfering with their ability to construct a valid proof.

Concepts play an important role in advanced mathematics. As objects are now defined solely on properties, students often struggle to achieve complete understanding of concepts. Understanding concepts plays a central role in mathematical activities such as using examples to make sense of formal statements or exploring new conjectures. Furthermore, one's concept image can interfere with the creation of valid proofs. Having syntactic strategies alone has been shown to be insufficient in many cases. For these reasons, concepts in advanced mathematics should not be neglected. Beyond their role in proof, understanding concepts themselves should be a major goal of introductory abstract algebra courses.

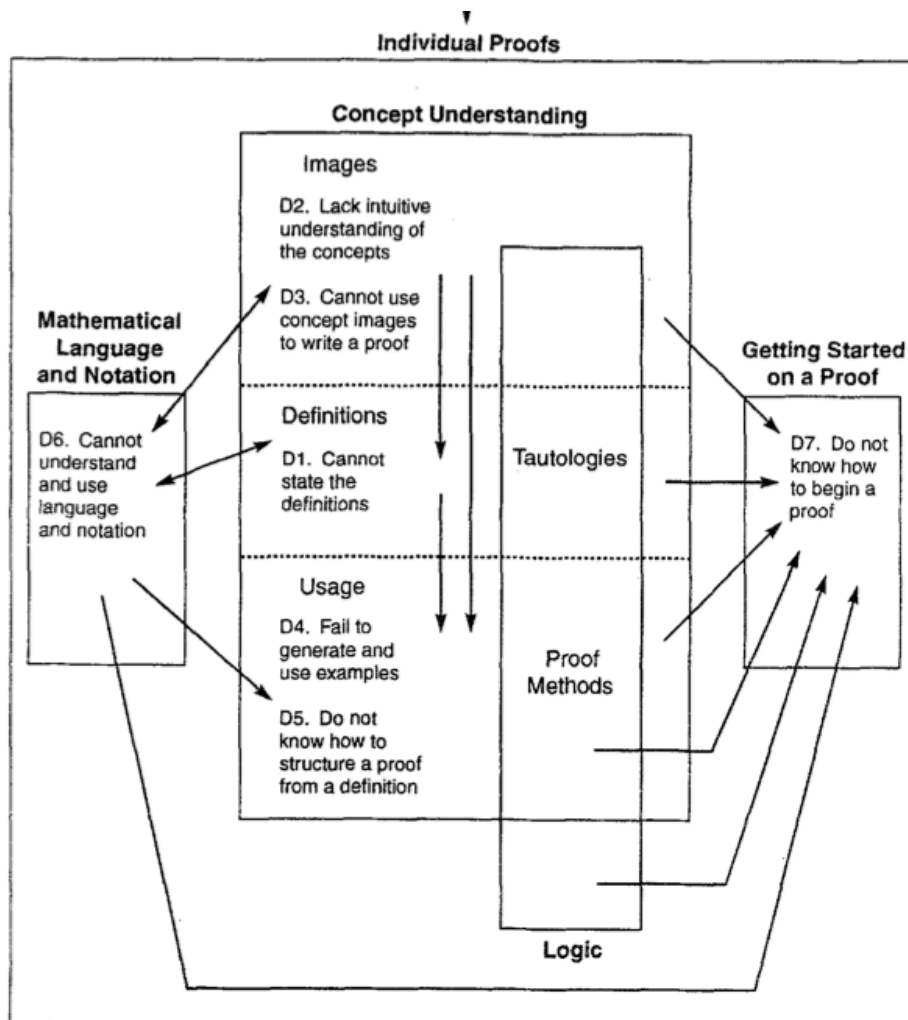


Figure 1. How conceptual understanding interferes with proving. Reprinted from "Making the transition to formal proof," by R.C. Moore, 1994, *Educational Studies in Mathematics*, 27, p. 252. Copyright 1994 by Kluwer Academic Publishers.

### Purpose of Concept Inventories

Concept inventories have traditionally been used for three purposes: diagnostics, placement, and evaluating instructional innovation (Hestenes, Wells, & Swackhamer, 1992). While placement is not typically relevant at this advanced level, the other two purposes remain significant.

**Diagnostics.** As a diagnostic tool, the inventory can serve a useful purpose to establish the current state of group theory concept understanding and to provide specific feedback for instructors. Since the Force Concept Inventory was released, it has been used repeatedly to collect information about what conceptions physics students have. The inventory has been credited with helping to instigate the reform movement in physics education (Savinainen & Scott, 2002a). In the same manner, the GCI was designed for ease and widespread usage to document how students currently conceive of concepts in group theory.

The diagnostic usage also makes the GCI useful for individual instructors. Concept inventories go through careful validation to ensure the distractors for all questions are meaningful. If an instructor administered the assessment to their class, they would gain feedback as to which concepts are understood and where additional attention should be put. Further, as Svainainen and Scott (2002a) explained about the FCI:

We believe that the very conception and design of the FCI can help the teacher to come to know and to understand the conceptual terrain of this field of instruction (in terms of both conceptual learning goals and student misconceptions) and to thereby be in a much stronger position to sustain effective teaching and learning dialogues (p. 51).

Concept inventories and similar assessments can be a powerful tool as the questions themselves provide an outline of alternate conceptions. The instrument can serve to bring awareness to various student conceptions that may differ from instructor conceptions. Instructors might find insight into unexpected student conceptions that may not have been captured by traditional proof-based assessment.

**Comparing groups and evaluation.** Innovation efforts can be evaluated meaningfully with the GCI, a validated instrument. GCI scores in classes with differing instruction can be compared for significant differences. Furthermore, performance on individual questions may reflect the impact of varying instruction.

With the FCI, this type of evaluation has been done at a larger scale to evaluate varying pedagogy and demographics. Docktor, Heller, Henderson, Sabella, & Hsu (2008) administered the FCI to 5,500 students to explore gender differences. Hoellwarth, Moelter and Knight (2005) compared traditional and studio classrooms looking at 6,000 students. The meaningful score from the FCI provides an easy way to compare groups. At a local level, an institution could evaluate conceptual understanding in their classes when implementing different types of instructional changes.

**Conclusion.** There are a few key facts that motivate the creation of this instrument. First, conceptual understanding in abstract algebra is not trivial, but it is essential. Yet, we often fail to assess concepts and instead present formal proofs as the only means for assessing in advanced mathematics. Furthermore, within the field of abstract algebra, there are no validated instruments to try to measure student understanding. Validated instruments can serve a central role in evaluating improvement efforts, researching instructional innovations and providing insight into the current landscape of student understanding in abstract algebra.

## **Abstract Algebra Literature Review**

Abstract algebra is traditionally considered a challenging course (Leron and Dubinsky, 1995). The difficulty with abstract algebra can be partially attributed to the nature of the course's content. Abstract algebra is a subject that is complex, abstract, and axiomatic. Many of the concepts such as isomorphism require a complex coordination of many ideas (Leron, Hazzan, & Zazkis, 1995). Additionally, the course is built around objects such as groups that are discussed and argued about abstractly (Hazzan, 1999, 2001). Abstract algebra is built on an axiomatic system leading to an elevated use of definitions (Edwards & Ward, 2004) and central role of proof and logic (Hart, 1994; Selden & Selden, 1987). As noted by Weber and Larsen (2008), "For most undergraduates this course is one of their earliest experiences in coping with the difficult notions of mathematical abstraction and formal proof" (p.139).

Research on the teaching and learning of abstract algebra has only begun in earnest over the last twenty-five years. Most early work in the area can be attributed to Dubinsky and his colleagues' who used their *APOS* (Action, Process, Object, Schema) framework to launch investigations into how students develop their understanding of major concepts of group theory (Dubinsky, 1997; Dubinsky, Dautermann, Leron, & Zazkis, 1994). Since this initial research, several other researchers have expanded upon student understanding in continued efforts to improve instruction (Weber & Larsen, 2008). The majority of research on teaching and learning abstract algebra falls into one of two categories: student misconceptions and instructional innovations. In this literature review, I present



an overview of research on student conceptions within the themes of algebra as *complex*, *abstract*, and *axiomatic*. The review concludes with attention to the literature related to instructional innovation efforts.

### **Student Understanding in Abstract Algebra**

**The abstract nature of Abstract Algebra.** As Hazzan (1999) pointed out, "...it is in the abstract algebra course that students are asked, for the first time, to deal with concepts which are introduced abstractly" (p. 73). Being able to reason about objects exclusively from properties requires students to have abstracted in some form. As Hazzan noted abstraction is discussed in a number of different manners in mathematics education. Abstraction can be understood in the development of objects. Tall (2004) addressed three different types of objects. Objects can arise through *empirical abstraction* where objects are studied in order to determine their properties, objects can arise through *reflective abstraction* where a process is encapsulated or compressed into an object, or objects can arise from the study of *properties* and their logical deductions.

**Reflective abstraction.** Reflective abstraction served as the underlying mechanism behind Dubinsky and colleagues' instructional innovations and APOS framework. While other researchers have mentioned processes and objects in algebra (Harel & Kaput, 1991; Sfard & Linchevski, 1994), the APOS framework was the only theory explored empirically.

APOS provides a development sequence, or genetic decomposition, students may use to make sense of mathematical concepts (Dubinsky et al., 1994). Dubinsky et al. summarize:

The essence of our theoretical perspective is that an individual, disequilibrated by a perceived problem situation in a particular social context, will attempt to reequilibrate by assimilating the situation to existing schemas available to her or him, or if necessary, use reflective abstraction to reconstruct those schemes at a higher level of sophistication (p. 269).

The levels of student conceptions fall into the categories of *action*, *process*, *object*, and *schema*. An action can be defined as, “any repeatable physical or mental manipulation that transforms objects in some way” (p. 270). Once the action can be understood as a whole (not each individual action needs to be taken), the action is *interiorized* and is now a process. A process can then be *encapsulated* into an object. This occurs when a process can be transformed by some action such as combining two processes. Equally important is the ability to *de-encapsulate* an object back to a process. An object should be more than a symbol but instead be able to be thought of as a process and object when appropriate. A set of processes and objects are *thematized* to form a schema or a collection of objects and processes that can be used together. (Schemas are discussed further in the complexity section of this chapter.)

Dubinsky et al. (1994) developed genetic decompositions of various algebra topics. Consider their breakdown of coset formation. Coset formation can be an action in familiar settings. At this stage, a student might list each individual element in a coset. Dubinsky et al. used this student explaining  $\mathbb{Z}_{18}/\langle 3 \rangle$ <sup>1</sup>, as an example of an action conception:

---

<sup>1</sup> The cyclic group of 18 elements modded out by multiples of 3.

<sup>2</sup> The eighteen topics were expanded to nineteen as cosets and quotient groups were separated to be consistent with topic treatment in curriculum. Modular groups were an example group and therefore are used within questions, but did not receive tasks independent of other concepts.

Well, the number in front is what you add to each element inside the set. So zero added to these six elements would keep the same six. One [the number] added to each, which is in the first column, would give you the 1,4,7,10 and then you add 2 to these first the  $H$  which is 0 through 6,9,12,15. Then you add 2 to each and you get 2,5,8,11,14, and 17 (p. 16).

This student began with the subgroup  $H=\{0,3,6,9,12,15\}$  and added elements from  $\mathbb{Z}_{18}$  one at a time to arrive at the elements in each coset. Coset construction can also be conceived of as a process where students associate each element with the subgroup without having to explicitly build the subsets. Dubinsky et al.'s process example student stated:

Ok, I interpreted this as  $0+H$ ,  $3+\dots$  Every third element beginning with 0. So  $1+H$  every third element beginning with 1 in  $\mathbb{Z}_{18}$  and every third element beginning with 2. And that would generate all the elements that are in  $G$  (p. 16).

These students frequently noted patterns for constructing which might limit their ability to deal with less familiar groups. This example student made no progress when trying to find a quotient group of a dihedral group. At the object level, a coset could be conceived of independent of the process, becoming an object for new actions such as comparing size of cosets. Dubinsky et al. presented example students who never wrote out all of the elements of cosets and worked entirely with their representatives. Dubinsky and his colleagues also provided this type of genetic breakdown for binary operation (Brown, DeVries, Dubinsky, & Thomas, 1997), normality (Asiala, Dubinsky, Matthews, Morics, & Oktac, 1997), coset multiplication (Asiala et al., 1997), permutations (Asiala, Kleiman, Brown, & Matthews, 1998), and symmetries (Asiala et al., 1998). The researchers worked to validate the framework through analyzing student work on exams and

interviewing students. Their analyses found evidence that students were working at the various levels of sophistication. However, as noted by Hazzan (1999), reducing abstraction such as using processes over objects does not necessarily equate to misunderstanding or even a lack of understanding. Students may work at various levels in problem-solving context as a means for sense making and developing intuition.

***Empirical abstraction.*** While reflective abstraction dominates the literature, empirical abstraction can also be found. Simpson and Stehlíková (2006) focused on the alternate transition, or as they label *shifts* in attention, as students moved from examples to general reasoning. The sequence includes five shifts where students begin by seeing the elements of a set as objects to be operated on, then attend to the interrelationships between set elements, followed by attending to names and symbols defining the abstract structure, seeing other examples of the general structure and finally moving to a formal system to derive consequences. The authors noted that while the first shift may involve the reification of objects (in the APOS sense), the final three shifts “involve relating the definitional properties of the (teacher- defined) abstract algebraic structure with the interrelationships noted in the initial example (and in subsequent instances of the structure)” (p. 352). This article focused primarily on the second shift defined as “*the shift of attention from the familiarity and specificity of objects and operations to the sense of interrelationships between the objects caused by the operations*” (p. 352). The researcher presented one case study where a university student was given a non-standard definition of  $\mathbb{Z}_{99}$  (a commutative ring). Through

extensive engagement with the context, the student was able to gain insight into structure. For example, the student confronted the idea of “opposite numbers” and was able to transition from subtraction to additive inverses arriving at  $99-x$  being the additive inverse in  $\mathbb{Z}_{99}$ . Further, the student used this in calculations treating  $3-7$  as  $3+99-7$ . Simpson and Stehlíková claimed these shifts represent a change from focus on operations and objects to obtaining new structural properties.

This type of exploration of examples and abstracting properties is not unusual in instructional sequences. For example, Larsen et al.’s (2013) curriculum has students reinvent the group concept through exploration of a symmetry group. (A more complete discussion of this process can be found in the instructional innovation section.)

***Reducing abstraction.*** Through interviewing algebra students from a traditional lecture-based class, Hazzan (1999) discovered that most students would work at a lower level of abstraction than what was introduced in class. Hazzan introduced a *reducing abstraction* framework to make sense of this activity. Within the framework are three interpretations for *levels of abstraction*.

The first level is *Abstraction level as the quality of the relationship between the object of thought and the thinking person*. This interpretation takes an individualistic view where each person has a relationship with a concept based on a given level of familiarity. In the case of groups, students might work with familiar objects, such as real numbers, instead of the unfamiliar concept of a newly introduced group. For example, in Hazzan’s study, a student decided that  $\mathbb{Z}_3$  was not a group based on the fact that the inverse of 2,  $\frac{1}{2}$ , was not found in  $\mathbb{Z}_3$ .

The student changed the operation of modular arithmetic to the familiar operation of real number multiplication. A similar error occurred when students commonly confuse normality with commutativity (Asiala et al., 1997). This type of reduction is consistent with Selden and Selden (1987) and Hart (1994) who addressed the potential errors in proofs caused by assuming that groups (or other structures) behave like real numbers or otherwise confused operations from unfamiliar to familiar.

The second interpretation is *Abstraction level as reflection of the process-object duality*. Hazzan (1999) asserted that conceiving of a concept as a process was at a lower abstraction level than conceiving a concept as an object. She suggested the use of first person as an indicator that the student was working in this lower level of abstraction. For example, a student explored the definition of quotient group stating, “I take all the elements [of  $H$ ] and multiply them on the right with some element from  $G$ ” (p. 81). In this way, she was not capturing quotient groups as a single object, but reduced the abstraction level to thinking about one step at a time.

The final interpretation is *Abstraction level as the degree of complexity of the concept of thought*. This type of reduction in abstraction occurred when a student would use one group when evaluating a set of groups. A non-generic example might be used to evaluate general statements about a class of objects leading the potential for overgeneralization. This type of reduction was also found in Selden and Selden’s (1987) taxonomy where students would argue about

elements when sets are appropriate such as arguing for  $gh=hg$  versus  $gH=Hg$  to establish normality.

*Navigating abstraction level.* Hazzan's (1999) framework provided insight into three different ways that students may lower abstraction levels in group theory. Hazzan cautioned that reducing abstraction "should not be conceived as a mental process which necessarily results in misconceptions or mathematical errors" (p. 75). Rather, these activities serve as coping mechanisms prior to students developing the mental constructs needed to work with the abstract objects.

Lowering abstraction has also been illustrated when students reason using Cayley tables (Findell, 2002; Hazzan, 2001) and students leveraging diagrams to make sense of cosets (Nardi, 2000). Findell presented an interview case where a student used an operation table when determining if  $\mathbb{Z}_3$  was a subgroup of  $\mathbb{Z}_6$ . The student began by creating what she called the "total table" (p. 1). After realizing she was unsure of the operation, the student made a table for multiplication, then ruled it out based on the 0 row. She then made the correct table and isolated the 0, 1, and 2 rows to correctly realize that this set was not closed. She even spontaneously used the table to find subgroups. However, using the table limited the student to verifying one axiom at a time in an external manner (such as having to first hunt down the identity element from the table.) The use of this instantiation might have been inefficient, but it served the purpose of reducing the abstraction level and making the properties of the group visible.

Weber and Alcock (2004) presented another important aspect of dealing with abstract objects. In their interview study with undergraduates and graduate students, they found graduate students flexibly navigating between formal, abstract representations, and concrete representations. Their undergraduate counterparts struggled to leave the formal plane. When determining whether groups were isomorphic, the undergraduates attempted to build maps using a symbol-pushing strategy labeled *syntactic proof production*. In contrast, doctoral students examined the groups for structural properties. The undergraduates had no intuition for what it meant for groups to be isomorphic. When asked, they stated the formal definition as equivalent to their intuition. All the experts espoused the intuition that isomorphic groups are essentially the same groups with renaming. When determining if  $\mathbb{Q}$  and  $\mathbb{Z}$  were isomorphic, the undergraduates (upon determining the cardinality) attempted and failed to build an isomorphism. In contrast, the doctoral students immediately determined the statement as false and largely used the cyclic nature of  $\mathbb{Z}$  and not  $\mathbb{Q}$  as their reasoning. The authors labeled the graduates' attempts as *semantic proof production* where "the prover uses instantiation(s) of the mathematical object(s) to which the statement applies to suggest and guide the formal inferences that he or she draws" (p. 210). Weber and Alcock showed navigating between the formal language and instantiations can provide needed intuition and illustrated that undergraduates may not have developed robust images of abstract concepts to do so successfully.

Zazkis, Dubinsky and Dauthermann (1996) presented a dichotomy similar to the semantic and syntactic proof production. The researchers used student



exploration of  $\mathbf{D}_4$  to differentiate between *visual* and *analytic* approaches. They defined visualization as “an act in which an individual establishes a strong connection between an internal construct and something to which access is gained through the senses” (p. 441). In contrast, an analytic act “is any mental manipulation of objects or processes with or without the aid of symbols” (p. 442). While not restricted to proof productions, the authors use analytic and visual in a parallel way to Weber and Alcock. Any diagrams or drawings constitute visual acts. In contrast, when “symbols are taken to be markers for mental objects and manipulated entirely in terms of their meaning or according to syntax rules, then we take the act to be one of analysis” (p. 442). In this way syntactic proof production could be labeled analytical. Zazkis et al. interviewed students with the prompt to find the elements of  $\mathbf{D}_4$ . They found that some students took a visual approach: using a square. Other students took analytic approach: writing down all permutation possibilities. Zazkis et al. challenged the dichotomy explaining analysis required both visual (connecting to the square) and analytical (labeling vertices). The authors conclude by presenting a model where students flexibly move between visual and analytic acts advocating for students to move flexibly between concrete and abstract.

**The complex nature of Abstract Algebra.** In addition to being abstract, concepts in abstract algebra have a high level of complexity. By complexity, I mean the number of aspects needed to be coordinated to understand and use a concept. In APOS terminology, this aligns with schemas. A person’s schema for a topic is “all of her or his knowledge connected (explicitly or implicitly) to that

topic.” A schema could be roughly equivalent to a *concept image* in the sense of Tall and Vinner (1981).

Even the fundamental concept of group requires an understanding of nontrivial concepts: set, binary operation (see Zaslavsky & Peled, 1996), associative property (see Larsen, 2010), identities, and inverses. Dubinsky et al. (1994) discussed some understandings and misunderstandings of groups and subgroups based on exams and student interviews from classes using their APOS-informed curriculum. Similar to students focusing on one simple structural property preserved by isomorphism as discussed above, Dubinsky et al. found students would often only attend to the set aspect of groups. Several students thought any set of three elements was the same group and that any subset of a group would be a subgroup. The authors attributed this difficulty to students trying to relate the unfamiliar concept of group with the familiar concept of set. Students with this conception were not yet coordinating the operation and set aspects of group. Dubinsky et al. also provided a genetic decomposition for quotient groups, which required coordination of new concepts: normality, cosets, and coset operation (via representatives). Students struggled to make sense of normality (confusing it with commutativity) and struggled to build cosets even with an understanding of normality. Siebert and Williams (2003) considered some of the complexity of cosets and quotient groups in terms of modular arithmetic groups. They found students had three conceptions around the cosets in this quotient group. The cosets could be viewed as infinite sets, they could be viewed as a representative element *and* a set, or just as the representative element.

Dubinsky et al. concluded their paper by commenting, “It is clear from our interview transcripts that an individual’s understanding of the elementary concepts connected with mathematical groups is quite complex” (p. 294). They continued stating that even in the beginning of abstract algebra, constructing understanding is a major cognitive development.

Leron, Hazzan, and Zazkis (1995) explored the issue of complexity in the context of isomorphism. They attested that for students in their first abstract algebra course, “isomorphism is a complex and compound concept, composed of and connected to many other concepts, which in themselves may be only partially understood” (p. 53). At the basics of the definition, students must understand quantifiers, functions, and groups. There are also some important distinctions such as the concept of two groups being isomorphic and the isomorphism itself, the formal definition and the naive definition, and proving groups are isomorphic versus proving groups are not isomorphic. When asked if two groups are isomorphic, students used the *order type*. The order type is the respective orders of the elements in a group, so the order type of  $\mathbb{Z}_4$  would be 1, 2, 4, 4. This approach would only be valid to show two groups with different order type are not isomorphic. Two groups that are not isomorphic could have the same order type. Leron, et al. attributed this confusion to one of three causes: neglecting all other properties for the simple and comfortable one, the fact that order type is sufficient for many small finite groups, or a confusion between a statement and its converse.

Leron, et al. (1995) went on to differentiate between two types of complexity: *syntactical* and *computational* complexity. When the students were asked what properties were preserved with isomorphism, they tended to list commutativity first and order of elements last. However, when asked to actually determine if two groups are isomorphic, students began by testing order of the elements and ended by determining if the groups are commutative. Commutativity was identified as syntactically simpler meaning it has simpler definition. However, computationally, showing all elements commute involves a global property, which takes more work the “step-by-step nature of order-type calculations” (p. 168).

Leron, et al. (1995) presented one more aspect of the complexity of isomorphism concept by investigating how students construct isomorphisms. They found students desired a canonical procedure and would get stuck if there were more than one way to proceed. This may reflect the quantifier in the definition where “there exists a function” could be interpreted as “there exists a unique function” or simply a lack of comfort with uncertainty. Additionally, “there exists a function” was expressed in three ways: “I can find a function,” “it is possible to find a function,” and “there exists a function” (p. 170). These phrasings may represent different levels of development for quantifiers. In a parallel manner, students may express a process or object view of the map with phrases ranging from, “Each element of  $G$  to each element of  $G$ ” (mapping each element individually) to a map from “ $G$  to  $G$ ” (seeing the map as an object

relating the sets.) The authors established some of the many coordinations students need to make in order to properly deal with isomorphism.

Nardi (2000) also tackled some of the complexities in group theory addressing the concept of order and the First Isomorphism Theorem. Through observation of tutoring sessions, Nardi found that order applying to both groups and elements was a potential point of confusion. When working with a prompt requiring order, the student explained, “I don’t understand how an element can have an order” (p. 173). While the student was comfortable with order of a group, the concept did not make sense applied to an element even after the tutor provided the definition from class (smallest integer  $k$  such that  $g^k=e$ ). Further, the tutor attempted to connect the two types of order by showing that  $|\langle g \rangle|=|g|$ . To make sense of this statement, the student had to use the group generated by  $g$  evoking additional concepts such as a cyclic group and using the operation correctly. (In this case, the student struggled to transition between additive and multiplicative notation, a struggle termed *notational inflexibility* by Selden and Selden (1987).) At each stage, the tutor seemed to have to unpack more and more concepts that were hidden in the simple statement:  $|\langle g \rangle|=|g|$ .

Nardi (2000) then identified the First Isomorphism Theorem “as a container of compressed conceptual difficulties.” The students had to coordinate multiple mappings, concepts of isomorphism, homomorphism, quotient groups, and kernel. They struggled at nearly every stage of proving this theorem from notation (the meaning of  $\sim$ ) to the definition a kernel. The complexity of the order

of an element and the complexity involved in the First Isomorphism Theorem presented barriers for the students in Nardi's study.

***Strategic knowledge.*** Schemas are also utilized in problem solving.

“When a person confronts a mathematical problem situation, he or she evokes a Schema and makes use of some of its components and some relations among them to deal with the situation” (Arnon et al., 2013, p. 112). Weber (2001) dealt with this issue in terms of strategic knowledge. Strategic knowledge can be defined as, “heuristic guidelines that they can use to recall actions that are likely to be useful or to choose which action to apply among several alternatives” (p. 111). Weber identified three types of strategic knowledge used in abstract algebra: *Knowledge of the domain's proof techniques*, *Knowledge of which theorems are important and when they will be useful*, and *Knowledge of when and when not to use 'syntactic' strategies*. With each concept comes a set of appropriate techniques. If one wants to show something is a group, then the group axioms should be invoked. If one wants to show groups are not isomorphic, then structural properties should be examined. Through interviewing graduate and undergraduates, Weber found undergraduates often lacked the knowledge of when to use theorems. When evaluating a prompt about surjective homomorphisms, the four doctoral students all made use of the First Isomorphism Theorem immediately. Only two of the four undergraduates did and only after substantial struggle. The undergraduate students in Weber's study largely reverted to syntactic strategies that did not prove fruitful. These strategic knowledge issues were consistent with Hazzan (1999) who found students misapplying Lagrange's

theorem. Strategic knowledge is intimately tied to the complexity of the subject where when and how to use a theorem or various strategies needs to be integrated into students' schema.

These episodes highlight just how complex some of the topics are in abstract algebra. The compactness of formal mathematics can act to obscure the complexity of concepts and theorems in introductory courses where students have yet to develop the necessary mental constructs.

**The axiomatic nature of Abstract Algebra.** For many students, abstract algebra is one of their first encounters with axiomatic systems and formal proofs. Much of the research related to abstract algebra has been in the context of proof.

**Definitions.** Edwards and Ward (2004) reported on the role definitions play in formal mathematics. The authors distinguish between *extracted* and *stipulated* definitions. The everyday use of definition is typically extracted. The definition describes something that already exists. In contrast, stipulated definitions can serve to create new concepts via stipulating defining properties. Through task-based interviews, the authors found students often fail to understand the role of definitions in mathematics where they are stipulated rather than extracted. Using Tall and Vinner's (1981) *concept image* and *concept definition* framework, Edwards and Ward illustrated that students rarely use the idealized model of development: building the concept image from the concept definition. Using the context of coset multiplication, the authors found that even when students knew the formal definition, they proceeded to operate using an incorrect concept image (such as the operation as unioning.) The students both professed an

explicit desire for definitions to describe something and failed to use definitions instead, appealing to informal concept images.

Lajoie and Mura (2000) addressed this issue in the context of cyclic groups. They found that their students were not using the formal definition to reason about cyclic groups, but rather their intuitive idea about cycles. The students largely explained the idea of generating as repeatedly operating an element with itself. As a result, students did not correctly address infinite groups. Of 28 student responses, they found 18 students thought  $\mathbb{Z}$  was not cyclic. In  $\mathbb{Z}$ , if one begins with an element such as 1 and operates 1 with itself, only the positive numbers are generated. The element will not “cycle” back around. Many students concluded that all cyclic groups are finite. If the students had leveraged the formal definition, which includes taking all positive and negative powers of one, they may have been able to appropriately deal with this infinite case.

This is consistent with Moore (1994) who found students were not able to use concept images to appropriately inform proofs and were unable to use definitions to determine the structure of proofs. Students who could state the definition of one-to-one were still unable to see how to use the definition to prove that a function was one-to-one. The statement of the formal definition alone was not sufficient to make sense and leverage the definition. This result was echoed by Wheeler and Champion (2013) who found students largely unable to prove one-to-one and onto results.

Hart (1994) presented a similar definition-type proof. The prompt was to show a set and operation form a group which Hart deemed a *satisfy axioms proof*.



Similar to the other researchers, Hart found that beginning undergraduate students often struggled with this basic proof type. This was not the case for more advanced students indicating that at some point a transition occurred where students do gain the requisite understanding to deal with these proofs (or potentially, these are just the students that survived to graduate studies and advanced courses).

**Logic and quantifiers.** Beyond not understanding the role of definitions, some of the failure to use them correctly might be attributed to a lack of comfort with logic and deduction. Selden and Selden's (1987) taxonomy explored many of the errors their abstract algebra students made during the proving process. Some errors included starting with the conclusion and being unaware that two differently named elements might not be different (an important distinction when showing that cosets form a partition). Additionally, the Seldens included the use of the converse of a theorem. This mistake is frequently associated with Lagrange's Theorem (Hazzan & Leron, 1996) where students make claims about the existence of subgroups of various orders. Another major error found in abstract algebra proofs is *weakening the theorem*. Selden and Selden had observed students frequently adding assumptions such as a group being finite or cyclic without an awareness of the alteration to the conjecture they are proving.

While most of these errors have not been studied extensively, Selden and Selden (1987) mentioned a major source of error that has been researched: quantifiers. As discussed in the complexity section, quantifiers are a necessary prerequisite to even make sense of definitions. Students are known to struggle

with quantifiers (Dubinsky & Yiparaki, 2000). Selden and Selden (1995) illustrated that students were often unable to *unpack* informal statements- convert statements from their informal (but typical) statement of a theorem to a formal style that reveals the quantification scheme. This is a key step to selecting a *proof framework* which Selden and Selden (1995) define as, “the ‘top-level’ logical structure of a proof.” This issue of quantification is relevant both for its necessary role in proving, and maybe more substantially, its necessary role in understanding definitions. The concept of one-to-one discussed above is often defined: A function  $f: A \rightarrow B$  is 1-1 if for all  $x \neq y$  in  $A$ ,  $f(x) \neq f(y)$ . While one might have a concept image of one-to-one (such as the horizontal line test), to use and understand this definition in a formal way requires a firm understanding of quantification. Novotná, Stehlíková, and Hoch (2006) presented a framework addressing some quantification issues related to binary operation understanding. They addressed the role of quantification in understanding the definition of identity elements. An element  $e$  is an identity for a given operation in a set if it fulfills the property  $ex=xe=x$  for *all*  $x$  in the set. They provided an example operation where students may identify the identity element as an expression depending on  $x$ . An identity must be the same for all elements in the set. Without an understanding of quantification, this difference may not be apparent.

In group theory, students are not only grappling with complex, abstract objects, but often are struggling to work in a formal system.

### **Instruction of Abstract Algebra: Innovations**

Prior to the 1990s, there were a few scattered articles on instructional innovation related to abstract algebra. Notably, Lesh (1976) provided a quantitative analysis comparing the influence of two types of organizers for the topic of finite groups. Through videos, four groups of students were exposed to finite groups in one of four combinations: either before the formal definition (advanced organizers) or after (poster organizers) and the organizers were either examples or counterexamples. Students did better when given advanced organizers and when given counterexamples. Lesh argued that the typical approach of providing just examples and not counterexamples was not ideal for students to make abstractions.

**APOS theory and programming.** The genetic breakdowns discussed above served as a driving force to create curriculum that would transition students through the stages to achieve advanced understanding of concepts. Dubinsky and his colleagues have proposed an instructional approach utilizing computer programming. This instructional approach was developed through the creation of genetic decompositions that were refined through two cycles of implementations of programming-based classes and reflections on student interviews (Dubinsky 1997; Dubinsky et al., 1994).

The classes utilized the *ACE* teaching cycle (Activities, Class discussion, and Exercises). Students constructed computer programs to explore mathematical concepts and then worked in groups for problem solving and discussion. The computer programming language ISETL allowed for the computer to do the

processes and then for processes to be used as inputs, ideally encouraging the transition from a process conception to an object conception.

Early in the course, students programmed group axioms where a set and an operation are inputs and output would be true or false for closed, associativity, identity and inverses. Then a program (*is\_group*) would input the set and operation into the four previous programs for each axiom to determine if they set and operation form a group (Leron & Dubinsky, 1995). Students then explored  $\mathbb{Z}_n$  under multiplication and addition (with and without zero). Leron and Dubinsky argued that their programming language was similar to mathematical definitions where a set of axioms must be met. The quotient group concept was similarly developed through programming. Students would create a program  $oo:=PR(G,o)$  where a group and operation are inputted. Then *oo* would take inputs that are group elements or subgroups. Inputs of an element and subgroup would calculate a coset (Asiala et al., 1997).

Through programming, Leron and Dubinsky claim there is “a good chance that a parallel construction will occur in the their mind” (p. 230). However, the data comes from exams and interviews without analysis of what students were doing when engaging with the programming. Furthermore, although students are reported to have performed marginally better than a control group of students from a lecture class, there was no attempt to document any significant differences (Brown et al., 1997).

**Guided reinvention and emergent models.** Larsen has spearheaded the development of a group theory curriculum where students reinvent formal

concepts beginning with their own informal knowledge (Larsen, Johnson, & Weber, 2013). Larsen designed his curriculum based on the Realistic Mathematics Education design heuristics of guided reinvention and emergent models, as well as proofs and refutations (Larsen & Zandieh, 2008). Larsen aimed to leverage student strategies in such a way that they transition from using models of situated activity to models for more general activity. (See Cook (2014) for a case where RME design heuristics are used in a Ring Theory design experiment.)

The curriculum was built through iterations of development and implementation. A series of small-scale design studies with pairs of students were conducted with the goals of discovering students' informal knowledge and strategies that can anticipate formal concepts. Additionally, Larsen worked to develop instructional activities to evoke these strategies and activities to help students use strategies to transition to formal concepts. The first iteration led to the development of a local instructional theory for group and isomorphism. After implementation in the classroom, a new second small-design experiment was conducted to build local instructional theories for quotient group. At this point Larsen and his team developed a full curriculum and worked to implement the curriculum in a multitude of group theory classes (Larsen, et al., 2013).

The instructional sequence begins with students working to identify the symmetries of a triangle leading to the creation of an operation table (Larsen et al., 2013). While searching for relationships and patterns to calculate combinations of symmetries, the students discovered the group axioms and discovered the property that each symmetry appears once in each row and column

(a motivating factor to establish the inverse axiom (Larsen, 2009).) Through investigating other groups for common structure, the students arrive at the definition for group that is then used for deductive activity. A similar process occurs with isomorphism where students are given “mystery group” tables and must determine if they are the same as the order 6 dihedral group (the symmetries of a triangle.) This activity culminates in a formal definition for isomorphism.

The quotient group sequence begins with students exploring the sets of even and odd numbers. The students leverage this understanding to find partitions of the dihedral group that would behave like even and odd numbers. Dubinsky et al. (1994) illustrated that students struggle with quotient groups and their set elements. Larsen and Lockwood (2013) showed how students might overcome this challenge. Consider the following episode from Larsen’s initial teaching experiment. At this point the students are making sense of a table with two elements: flips and rotations (of a square.) Initially, the student said it had eight elements.

Teacher/Researcher: What if I said it had two?

Rick: Then it wouldn’t be a group.

Sara: Well if you want to make meta-groups [student name for quotient groups]

Teacher/Researcher: Ok, let’s make meta-groups

Rick: No, wait, maybe it would be a group it would be a group now for me (p. 732).

Sara quickly agreed and when the teacher/researcher asked what the elements of the group were, Rick replied, “Rotations and flips. My identity would be rotations” (p. 732). This exchange represented one of the important transition

points illustrating how Larsen's curriculum directly aims to support some of the difficulties established in prior studies.

Each concept reinvention is followed by a deductive phase where students prove conjectures and arrive at important consequences such as the normality requirement for the creation of quotient groups. Unlike Dubinsky (1997), each of the curriculum stages was illustrated with corresponding episodes providing insight into how students engage in the mathematical activities (Larsen, Johnson, & Weber, 2013).

As in the case of Dubinsky (1997), the effect of this instructional method is not well-documented. Larsen et al. (2013) did compare student responses on a quotient group survey administered to eight classes using the curriculum and six classes that did not. Students in the experimental classes were significantly more likely to correctly assess if a subgroup could be used to form a quotient group from a table. Furthermore, the students from traditional classes only used normality as justification; the experimental class's students were equally likely to address the set operation not being well-defined. As noted by Larsen et al., there is still a lot of evaluation work to be done.

**A note on other research in Abstract Algebra classrooms.** Beyond this instructional innovation research, there also exists a body of research where the abstract algebra setting was used as a means to explore other aspects of teaching or learning. For example, Cnop and Grandsard (1998) and Grassl and Mingus (2007) implemented cooperative work in the abstract algebra classroom but did not report on the content or student learning. Other objectives such as teacher

listening (Johnson & Larsen, 2012), sociomathematical norms (Fukawa-Connelly, 2012), and student-tutor relationships (Jaworski, 2002) have been studied in an abstract algebra context. However, these papers largely lie beyond the scope of this literature review because their focus was not on student learning of algebra concepts either through exploration of their understanding or content-related instructional innovations.

## **Conclusion**

The body of educational research in abstract algebra focuses largely on incorrect student conceptions and instructional innovations. Several studies focus on proof in the context of abstract algebra (Hart, 1994; Selden & Selden, 1987; Weber & Alcock, 2004; Weber, 2001). Another batch of studies explores aspects of student difficulty such as with abstraction (Hazzan, 1999), complexity (Leron et al., 1995), or process-object duality (Dubinsky et al., 1994). There is also literature sharing research-grounded instructional innovations, most notably Larsen et al. (2013) and Dubinsky (1997).

This leaves a lot of areas for continued research. First, *APOS* theory is the only thoroughly explored mechanism for concept acquisition. The sequence action, process, object, and schema may not be the only means of object construction (Tall, 1999). Of note, empirical abstraction may have a more substantial role than has been explored. The APOS framework also has the limitation of being purely cognitive. Although, recently Johnson (2013) has contributed a theoretical paper for analyzing learning in classrooms informed by



the design heuristic of guided reinvention which may provide an alternate way to view learning in the subject.

Even in terms of research on student conceptions, majority of prior studies have small sample sizes. A validated instrument, such as the GCI, can serve to test the generalizability of smaller studies and make a broader argument about the general state of student understanding in abstract algebra. Further, in the absence of a validated measuring tool, claims of successful instructional innovations lack a strong foundation. This is reflected in the cautious claims made by both Larsen et al. (2013) and Dubinsky's (1997) teams.

Group theory is typically believed to be a traditional stumbling block for mathematics majors (Dubinsky et al., 1994). As a field, we need to develop a coordinated picture of what it means for a student to understand group theory. The themes of complexity, abstraction, and formal logic provide a lens to continue towards this goal.

## **Concept Inventories and Theoretical Background**

In this section, I briefly discuss some of the theoretical considerations for this study. This section begins with an introduction to concept inventories. The meaning of conceptual understanding is explored with attention to various components of it. The conceptual boundaries in a given domain are discussed in terms of a domain analysis. I briefly address various tools that can help map the conceptual domain including curriculum (textbooks) and expert opinion. Additionally, I provide a brief overview of theory relevant to assessment design. Within each section, I return to the goals of the GCI to situate the discussion relative to this study.

### **Concept Inventories**

A concept inventory is a multiple-choice assessment instrument with a focus on conceptual understanding. While many assessments aim to capture procedural understanding, conceptual understanding is often neglected (Pegg, 2003; Tallman & Carlson, 2012). Stone et al. (2003) described the creation of this sort of instrument for statistics. Their questions were based on, “the basic concepts of statistics and probability. The questions are non-computational in nature and focus on conceptual understanding rather than problem-solving ability (p. 3).” In a similar manner, the GCI is aimed to capture conceptual understanding centered on set of concepts rather than computational or proving ability.

The development of concept inventories can trace its roots to Hestenes et al.'s (1992) *Force Concept Inventory* (FCI). The name of the instrument lends

itself to two interpretations: the *Force Concept* Inventory or the *Force Concept* Inventory. The authors described their instrument using the former interpretation: they are aiming to assess the students “belief systems” (p. 2) about one concept: force. Hestenes et al. attempted to determine if students have a “coherent conceptual system” (p. 14) underlying their understanding of Newtonian physics. However, within the construction of their instruments, the authors created items to capture a multitude of incorrect conceptions related to various concepts integral to the force concept. In this way, they also treat the instrument as a *Force Concept* Inventory, assessing various concepts related to the subject.

Similarly the GCI could be understood flexibly as a *Group Concept* Inventory or a *Group Theory* Concept Inventory. The former requires an underlying model. The fundamental concepts in group theory are all intimately tied into one’s understanding of group. A well-developed system around group would include understanding of concepts necessary to understand groups (such as sets) as well as concepts that are informed by one’s understanding of groups (such as normal subgroups.) In the following section, I expand on various aspects that might contribute to a robust model of group starting with a discussion of the nature conceptual understanding.

### **Conceptual Knowledge**

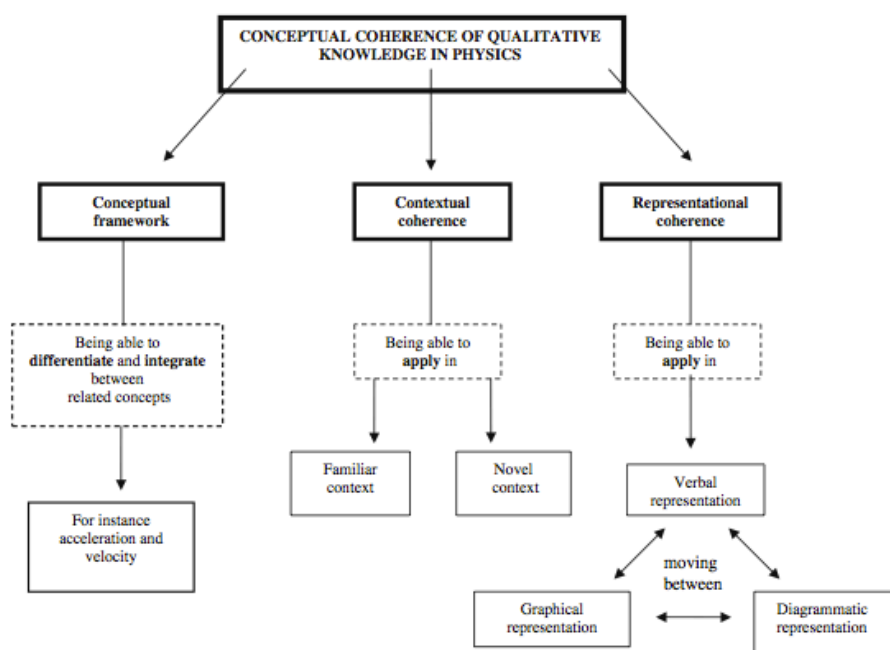
What one means by *conceptual knowledge* is not consistent within the mathematics education field. Hiebert and Lefevre's (1986) seminal work in the area introduced the definition of conceptual knowledge as “knowledge that is rich in relationships. It can be thought of as a connected web of knowledge, a network

in which the linking relationships are as prominent as the discrete pieces of information” (pp. 3-4). Star (2005) argued that this is actually a way concepts can be known, claiming the term conceptual knowledge encompasses both this way of knowing but also “what is known (knowledge of concepts)...” (p. 408). Star classified knowledge of concepts and relationships as *deep* conceptual knowledge. It is this definition that best reflects the goals of concept inventories. The FCI did not assess the definition of force, but rather assessed the underlying model and relationships. Similarly, the goal of the GCI is to assess not just the concepts but the properties and relationships amongst and within concepts. Understanding a concept requires both an understanding of the relations to surrounding concepts, as well as the concept itself.

This idea is consistent with Tall and Vinner’s (1981) *concept image* and *concept definition* constructs. They explain that the “total cognitive structure which colours the meaning of the concept” (p. 152) is far greater than just the symbol or name used to represent it. They use the term *concept image* to capture the total cognitive structure associated with a concept, which includes “all mental pictures and associated properties and processes” (p. 152). A *concept definition* is then the “words used to specify a concept” (p. 152). Tall and Vinner note that concept images need not be coherent and that frequently only portions are evoked at given times. It is this lack of coherence associated with Newtonian force concepts that Hestenes et al. (1992) set out to capture with their inventory.

A complete understanding of a topic would include related concepts forming some sort of coherent system. These relationships are a component of

Star's (2005) deep conceptual knowledge. Tall (2007) discussed the need for “compressing knowledge into thinkable concepts” in order “to build a coherent connected mental structure to put significant ideas together” (p. 153). In the context of the Force Concept Inventory, Savininen and Viri (2008) used the idea of *conceptual coherence* to frame this structure. In their theory, conceptual coherence in physics could be thought of as a conceptual framework (relating different concepts), contextual coherence (being able to apply knowledge in appropriate contexts), and representational coherence (moving between verbal, graphical and diagrammatic representations.) (See Figure 2.)



*Figure 2.* Conceptual coherence in physics. Reprinted from “The Force Concept Inventory as a measure of students conceptual coherence,” by A. Savininen and J. Viiri, 2008 *International Journal of Science & Mathematics Education*, 6, p. 722. Copyright 2007 by National Science Council.

In the context of mathematics, Biehler (2005) presented a parallel breakdown of three aspects that contribute to mathematical meaning of concepts.

He identified *the domain of application, relations* to other concepts within mathematics and *representations* available for working with the concept. Biehler used functions to illustrate his breakdown. The domain of application for functions included applications like curve fitting and modeling motion. Representations included types such as equations, tables, and graphs. Relations to other concepts included relationships to variables and mappings.

In a group theory context, groups would be related to other concepts such as binary operations, group axioms individually, and group isomorphisms. Relations to other concepts are part of a deep conceptual knowledge. The tools and representations for working with the concepts bring an additional dimension. A group might be understood symbolically (such as a set with a rule or a more general notation), tabularly (via a Cayley Table), or visually (as a set of symmetries). The domain of application might include describing the symmetries of a shape or describing the structure of our number system. As noted by Savininen and Viri (2008), the representations and applications aspects overlap with the relations aspect. “In order to apply a concept in a variety of contexts, the student must relate (integrate) a concept to other concepts. The student also needs to differentiate that concept from related concepts” (p. 723). The relationships between concepts underlie any assessment. It is nearly impossible to isolate a concept fully and still ask a meaningful question.

Utilizing the ideas of concept image, conceptions, and conceptual coherence, various components of conceptual understanding can be explored more in depth. The following section includes a discussion of examples, tasks and

student activity, representations, and formal/informal understanding of definitions.

**Examples.** Sinclair, Watson, Zazkis, and Mason (2011) addressed examples as an important part of a *concept image*. They introduced the construct of a *personal example space*. A personal example space was defined as “the set of mathematical objects and construction techniques that a learner has access to as examples of a concept while working on a given task” (p. 291). An example consists of a specific instantiation of a more general concept. Aspects of personal example spaces include *population* (dense or scarce), *connectedness* (connected to other examples), and *generality* (representiveness of a class of objects). Whereas a personal example space is individual, a *conventional example space* is the example space “as generally understood by mathematicians and as displayed in textbooks” (Watson & Mason, 2006, p. 15). It is the conventional example space that can be analyzed to bolster the representative nature of the concept inventory and provide insight into the domain of application.

Examples can serve a multitude of purposes within a field. For example, Watson and Mason (2006) identified *reference examples*. Reference examples capture the properties of a class of object and can then be utilized to test conjectures and in other applications. Within abstract algebra, Fukawa-Connelly and Newton (2014) have recently investigated how an instructor utilized examples of groups. They used variation theory to explore the types of properties a presented example group did or did not have. These properties included whether example groups were commutative or non-commutative, infinite or finite, as well

as what property was missing in the case of non-examples. Further, they identified various functions of examples in instruction including: illustrating a *definition*, instantiating a *statement*, *introducing* a topic, or *motivating claims*. While this analysis was conducted on examples from a single class, both the types of examples and their function provide a starting point for describing the usage and types of examples related to group. Exploring the function and nature of examples in curriculum can provide insight into the types of examples to which students have access as well as what examples might be utilized towards certain ends.

**Tasks.** Associated tasks provide another way to address the domain of application. Tasks can be understood in terms of their expected student activity. In this way, the standard activity associated with various topics can be understood. In Thompson, Senk, & Johnson (2012) and Otten, Gilbertson, Males, and Clark's (2014) textbook analyses, they investigated student activities associated with proof in algebra and geometry textbooks respectively. Some of the activities they found in their textbooks include making conjectures, investigating statements, constructing proofs, and finding counterexamples. (See Figure 3 for a complete list of codes.) These types of explorations can help identify the domain of application portion of conceptual understanding.



	Exposition	Student Exercises	
	Properties, Theorems, or Claims	Related to Mathematical Claims	Related to Mathematical Arguments
Mathematical Statement or Situation	<ul style="list-style-type: none"> <li>• General</li> <li>• Particular</li> </ul>	<ul style="list-style-type: none"> <li>• General</li> <li>• Particular</li> <li>• General with particular instantiation provided</li> </ul>	<ul style="list-style-type: none"> <li>• General</li> <li>• Particular</li> <li>• General with particular instantiation provided</li> </ul>
Expected Student Activity		<ul style="list-style-type: none"> <li>• Make a conjecture, refine a statement, or draw a conclusion</li> <li>• Fill in the blanks of a conjecture</li> <li>• Investigate a conjecture or statement</li> </ul>	<ul style="list-style-type: none"> <li>• Construct a proof</li> <li>• Develop a rationale or other non-proof argument</li> <li>• Outline a proof or construct a proof given an outline</li> <li>• Fill in the blanks of an argument or proof</li> <li>• Evaluate or correct an argument or proof</li> <li>• Find a counterexample</li> </ul>
Justification (or environment for exploration)	<ul style="list-style-type: none"> <li>• Deductive</li> <li>• Empirical</li> <li>• Outline</li> <li>• Past or future</li> <li>• Left to student</li> <li>• None</li> </ul>	<ul style="list-style-type: none"> <li>• Deductive (explicit)</li> <li>• Empirical (explicit)</li> <li>• Implicit</li> </ul>	<ul style="list-style-type: none"> <li>• Deductive (explicit)</li> <li>• Empirical (explicit)</li> <li>• Implicit</li> </ul>
	• Statements <i>about</i> reasoning-and-proving	• Exercises <i>about</i> reasoning-and-proving	

*Figure 3.* An analytic framework for reasoning-and-proving in geometry textbooks. Reprinted from “The mathematical nature of reasoning-and-proving opportunities in geometry textbooks,” by S. Otten, N.J. Gilbertson, L.M. Males and D.L. Clarck, 2014, *Mathematical Thinking and Learning*, 16, p. 58 Copyright 2014 by Taylor & Francis Group, LLC.

Within a course such as abstract algebra, proof activity often dominates the tasks. For that reason, it makes sense to further subdivide types of proofs. In Hart’s (1994) analysis of proof-writing in abstract algebra, he identified several key types of proofs in group theory including *satisfy axioms proof* (show a given example is a group), *set-definition proof* (show a given subset is a subgroup), *uniqueness proof* (the existence of a unique idempotent element), *syntactic proof* (use a procedural or syntactic approach to show that a given group is Abelian), and *non-routine proof* (such as showing that a group with even number of elements has one element that is its own inverse). Hart’s list is not exhaustive, but does illustrate some of differences that exist within the proof category of activity.

While much of the activity in algebra textbooks may be limited to proof, other activity should also be explored. Tallman and Carlson (2012) adapted Anderson and Krathwohl's (2001) seven intellectual behaviors to categorize cognitive behavior associated with tasks on calculus exams. Their categories include: *Remember*, *Recall and apply procedure*, *Understand*, *Apply understanding*, *Analyze*, *Evaluate*, and *Create*. Remembering involves addressing a rote factual question stemming from some prior knowledge. In group theory, this might appear as stating or checking a definition for group or isomorphism. Recall and apply procedure involves being prompted to apply a known procedure. This might look like "calculate the order of an element" in a group theory context. Several of their activities are less procedural including evaluating and applying understanding. In a group theory context, a student may be asked to evaluate if a given statement is true, such as: "All order 12 groups have a subgroup of order 6." Students may also need to apply understanding of various important theorems in order to answer questions. For complete descriptions of Tallman and Carlson's categories see Table 1.

Together, tasks and examples provide a relatively complete view of the domain of application (at least within the context of introductory group theory classes.)

**Representations.** Representations also provide essential information about concepts. In Lesh's (1981) article on problem-solving, he introduced the notion of *representational systems* including spoken symbols, written symbols, pictures, manipulative models, and real world situations. Broadly, representations

can be thought of as “two entities that are taken, by an actor or an observer, to be in some referential relation to one another, one taken to ‘represent’ the other”

(Kaput, 1998, p. 266).

Table 1

*Cognitive Behavior Associated with Calculus Exam Items. Adapted from “A characterization of Calculus I final exams in US colleges and universities,” M.A. Tallman and M.P. Carlson, 2012, The 15th annual conference on Research in Collegiate Mathematics Education, p. 2-222.*

Cognitive Behavior	Description
Remember	Students are prompted to retrieve knowledge from long-term memory (e.g., write the definition of the derivative).
Recall and apply procedure	Students must recognize what knowledge or procedures to recall when directly prompted to do so in the context of a problem (e.g., find the derivative/limit/integral of $f$ ).
Understand	Students are prompted to make interpretations, provide explanations, make comparisons or make inferences that require an understanding of a mathematics concept.
Apply understanding	Students must recognize when to use (or apply) a concept when responding to a question or when working a problem. To recognize the need to apply, execute or implement a concept in the context of working a problem requires an understanding of the concept.
Analyze	Students are prompted to break material into constituent parts and determine how parts relate to one another and to an overall structure or purpose. Differentiating, organizing, and attributing are characteristic cognitive processes at this level.
Evaluate	Students are prompted to make judgments based on criteria and standards. Checking and critiquing are characteristic cognitive processes at this level.
Create	Students are prompted to put elements together to form a coherent or functional whole; reorganize elements into a new pattern or structure. Generating, planning, and producing are characteristic cognitive processes at this level.

As in *personal* and *conventional* example spaces, representations exist as both *external* systems and *internal* systems (Goldin, 2002). External representation systems are the concrete systems available and observable. Internal representation consists of the personal systems that students construct. These are unobservable mental models. External systems are those existing in documents such as textbooks. Within the group theory context, researchers have addressed visual representations of groups such as utilizing an equilateral triangle for symmetries (Almeida, 1999) and student use of Cayley tables (Hazzan, 2001).

Nardi (2000) reported on some of the challenges of a student utilizing a visual representation of equivalence classes (cosets) (See Figure 4). Understanding and relating representations is a key piece of conceptual understanding.

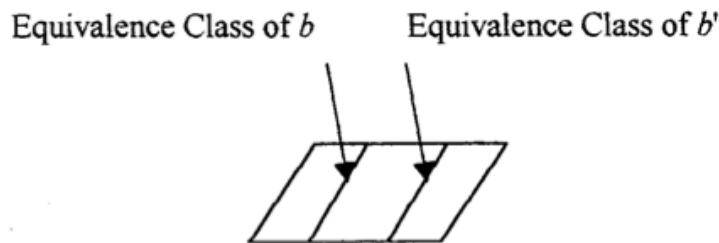


Figure 4. Equivalence classes represented by fibers. Adapted from “Mathematics undergraduates’ responses to semantic abbreviations, ‘geometric’ images and multi-level abstractions in group theory,” by E. Nard, 2000, *Educational Studies in Mathematics*, 43, p. 175. Copyright 2001 by the Kluwer Academic Publishers.

**Formal and informal definitions.** Closely related to representations are definitions. A definition could be one type of representation. Tall and Vinner (1981) refer to a *formal concept definition* as a mathematical definition that is accepted by the mathematics community. Raman (2002) noted in mathematics there is often one accepted formal definition such as the delta-epsilon definition for continuity. However, a multitude of informal definitions might exist such as continuous functions being characterized as “functions as ones whose graphs have no breaks” (p. 136). In her study, she investigated both the formal and informal ways that topics can be discussed in textbooks.

Both formal and informal characterizations are an essential aspect of understanding a concept. For example, in Leron, Hazzan and Zazkis’s (1995) discussion of isomorphism, they differentiated between a naïve and formal understanding of isomorphism. A formal understanding of isomorphism would be

via the typical definition: two groups are isomorphic if there exists a one-to-one and onto homomorphism between the groups. A naïve view of isomorphism would be the non-formal characterization of groups such as being essentially the same groups just with elements renamed. The informal characterizations of concepts provide another dimension of conceptual understanding.

**Group schema.** In the broadest sense, the concepts (as well their examples, applications, representations, and characterizations) in group theory are all part of a well-developed group *schema* (Dubinsky, 1994; Amon, et al., 2013). A schema for a mathematical topic can be defined as, “all of [an individual’s] knowledge connected (explicitly or implicitly) to that topic” (p. 110). An individual’s schema includes objects, processes, and other schemas “which are linked by some general principles or relations to form a framework in the individual’s mind that may be brought to bear upon a problem situation involving that concept” (Amon, et al., 2013, p. 110). The group schema requires the coordination of an *axiom schema*, *binary operation schema* and *set schema* as illustrated in Figure 5. This genetic decomposition illustrates some of the coordination and fundamental concepts required to have a well-developed group schema.

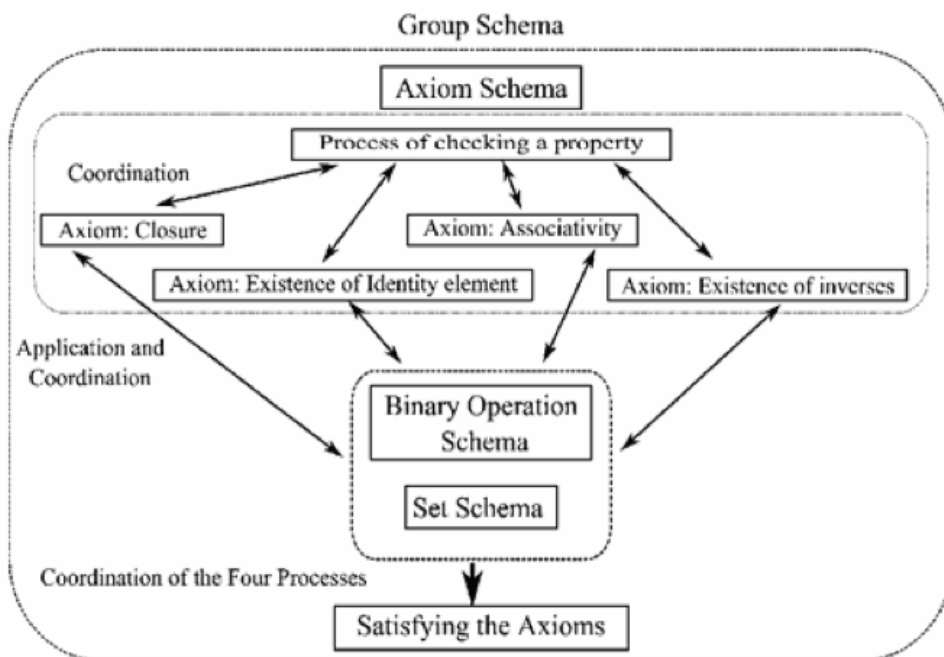


Figure 5. A genetic decomposition of the group schema. Reprinted from *APOS Theory: A framework for research and curriculum development in mathematics education* by I. Arnon, 2013, Copyright 20013 by the Springer London.

However, using this decomposition does not capture the scope of schema as defined above which most closely resembles a concept image. The genetic decomposition consists of the mental constructions needed to make sense of group. A schema may also contain related conceptions such as quotient groups (which requires coordination of cosets, binary operation and group (Asiala, Dubinsky, Mathews, Morics, & Oktaç, 1997)), as well as example groups and how they may be leveraged. A group concept inventory needs to address groups in both directions: concepts that are coordinated to understand groups and concepts that require groups to make sense of them. “When a person confronts a mathematical problem situation, he or she evokes a Schema and makes use of some of its components and some relations among them to deal with the

situation” (Arnon et al., 2013, p. 112). These relationships play a central role in mathematical activity. Conceptual understanding is not knowledge of what a concept is alone. Rather, conceptual understanding represents a broad network of relationships that include relationships amongst topics, formal and informal understanding of a concept, representations of the concepts, and how the concepts may be utilized. In order to create a full picture of understanding a concept, these various components of conceptual understanding should be addressed.

### **Fundamental Topics in Group Theory**

Most university group theory courses start in one of two directions: beginning with symmetry and then building group axioms or beginning with group axioms with symmetries as just a passing example (Almeida, 1999). These differing schools of thought may reflect differences in what constitutes the fundamental concepts of group theory. In fact, Dubinsky et al., (1994) reported “On Learning the Fundamental Concepts of Group Theory” where they discussed groups, subgroups, cosets, coset products, and normality, sparking a discussion on this matter. Burn (1996) challenged that isomorphism, closure, associative, identity, inverses, sets, functions, and symmetry were all overlooked as fundamental concepts. Dubinsky, et al. (1997) conceded that the title should have stated “on some fundamental concepts of group theory” (p. 251) where functions, sets, permutations, symmetries and the four group axioms are also fundamental to group theory. Johnson, Caughman, Fredericks, and Gibson (2013) discussed the important topics in introductory group theory as, “the axioms, basic properties, orders of elements, subgroups, examples, permutations, isomorphisms,

homomorphisms, cosets, normal subgroups, quotient groups, and the isomorphism theorems” (p. 750). Caughman, an experienced group theory instructor, provided this list when explaining the tensions between coverage and developing deep conceptual understanding. Suominen (2014) provided a slightly different list of topics when having graduate students zoom out on abstract algebra more holistically. Three graduate students were asked to identify the important topics in abstract algebra. The students agreed only on group, ring, field, Galois Theory, isometries and geometric applications. Suominen noted that there is no consensus as to what topics are essential in the subject area. These discussions highlight the fact that what constitutes the fundamental concepts has not been explored in any concrete manner and any number of topics may be argued as essential without further exploration.

**Using experts to determine fundamental topics.** The above studies and conversations represent one potential source of determining essential aspects of a domain: using experts in the field. For an ill-defined domain with subjective attributes, often experts in the domain provide the optimal resource for making decisions about what content is fundamental. In previous concept inventories and conceptual-driven assessments, experts served a multitude of roles. In the PCA, experts evaluated tasks after their creation to determine if they were appropriate and accurate (Carlson, Oehrtmen, & Engelke, 2010). In the development of the Calculus Concept Inventory, the creators leveraged a panel of experts to identify relevant topics and create and evaluate tasks (Epstein, 2007). Experts were leveraged in a more formal manner in the creation of the Comprehensive



Assessment of Outcomes in a first Statistics Course, a conceptual assessment of introductory statistics (CAOS). A team was assembled that served to both make decisions of where to focus, as well as create multiple-choice tasks. The team provided several rounds of feedback where they addressed the content validity (relevance to what is being targeted), and identified concepts that were not being targeted on the test (Delmas, Garfield, Ooms, & Chance, 2007). This iterative process is one potential model for arriving at a set of tasks and important topics related to a domain. Another option is to employ a panel consensus protocol in the front-end of a project. The creators of the Thermal and Transport Science Inventory did this through a Delphi study where a panel of experts arrived at ratings for importance and difficulty of a series of topics (Streveler, Olds, Miller, & Nelson, 2003). A more detailed discussion of Delphi studies can be found in the Methodology Chapter.

**Curriculum and textbooks.** Beyond leveraging the knowledge of experts, artifacts of the field can also serve as a basis for determining the important and valued aspects of domain. For introductory group theory, textbooks provide essential information about what is valued in the classrooms. In regards to textbooks, Zhu and Fan (2006) noted:

...textbooks are a key component of the intended curriculum, they also, to a certain degree, reflect the educational philosophy and pedagogical values of the textbook developers and the decision makers of textbook selection, and have substantial influence on teachers' teaching and students' learning (p. 610).

Textbooks contain narratives that introduce the important topics in a domain as well as examples, and exercises related to these topics. Textbook analysis

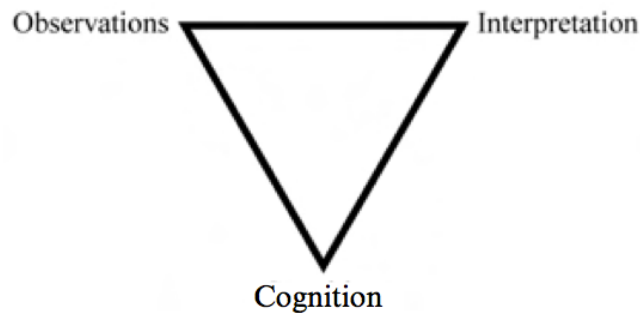
frequently consists of analyzing the narratives or exercises alone (Li, 2000). However, Thompson, Senk, and Johnson (2012) argued, “The two aspects together provide a more complete and coherent picture of opportunities to learn than that which can be obtained from analyses of either narrative or exercises alone” (p. 256). In order to arrive at a complete picture of what is valued, both should be explored. A better picture of the conceptual domain surrounding a topic can be developed through exploring what examples textbooks provide, the types of activities the textbooks present to students, and the types of representations frequently used.

**Conclusion.** Prior to developing items, an understanding of the targeted subject area is vital. Experts and artifacts provide essential information on the domain. In the case of education, instructors and textbooks are two essential parts of the domain. Textbooks provide insight into the intended curriculum and typical tasks to which students are exposed. Experts provide a deeper level of insight and through careful processes such as a Delphi study, a meaningful consensus on the valued and important aspects of a domain can be established.

### **Assessments**

There are three key components underlying assessment: observations, interpretation and cognition (see *Figure 6*). We are trying to measure cognition via observations and arrive at assessment of cognition via interpreting the observations (Pellegrino, Chudowsky, & Glaser, 2001). Cognition is “a theory or set of beliefs about how students represent knowledge and develop competence in a subject domain” (p. 44). The observation corner “represents a description or set

of specifications for assessments tasks that will elicit illuminating responses from students' about the target domain to be measured" (p. 48). The interpretation corner "encompasses all the methods and tools used to reason from fallible observations" (p. 48). In order for an instrument to be valid, it must coordinate these aspects.



*Figure 6.* The assessment triangle. Adapted from "Rigorous methodology for concept inventory development: Using the 'assessment triangle' to develop and test the Thermal and Transport Science Concept Inventory (TTCI)," by R.A. Streveler, R.L. Miller, A. I. Santiago-Roman, M.A. Nelson, M.R. Geist and B.M. Olds, 2011, *International Journal of Engineering Education*, 27, p. 969. Copyright 2011 by TEMPUS Publications.

**A quick note on test theory.** When evaluating and creating an instrument, a test theory is needed to inform the process. There are two major branches of test theory often referred to as classical test theory and modern test theory. The classical test model assumes a simple linear relationship where:

$$\text{Test score} = \text{True score} + \text{Error}.$$

In a valid instrument, true score would represent what the test is intended to measure or equivalently the score across all parallel forms even for an invalid instrument (Hambleton & Jones, 1993). Whereas error would be a random

uncorrelated component accounting for the difference between true score and test score.

In contrast, modern test theories, often categorized as latent trait theories or item response theories, take a more nuanced approach to modeling items. Instead of a test focus, each item is modeled with an item characteristic curve. These curves are generated by various functions that look like:

$$P_i(\theta) = c_i + \frac{(1 - c_i)}{1 + e^{-Da_i(\theta - b_i)}} \text{ (Hambleton \& Jones, 1993).}$$

where  $\theta$  is equivalent to true test score, or the latent trait being measured. The above model is the most complex with three parameters: item difficulty ( $b$ ), item discrimination ( $a$ ) and a guessing factor ( $c$ ). Simplified versions based on just item difficulty, or item difficulty and discrimination also exist.

Overall, item response models have some benefits over classical test models. The models are more theory grounded, and do not have the limitations of being sample (of selected items) and sample (of examinees) dependent (Fan, 1998). Additionally, item difficulty can obscure item discrimination in classical test theory (Hambleton & Jones, 1993). A particularly easy item might be non-discriminating because most students get it right using a classical test theory approach, however, the item might do an excellent job discriminating between low scorers which would only be captured in the more complex models of modern test theories. Yet, classical test theory continues to underlay the majority of concept inventories (see Carlson et al. (2010) and Hestenes et al. (1992) for examples.) From a practical standpoint, classical test models require a smaller

sample size and less intensive analysis (such as goodness-of-fit studies) (Hambelton & Jones, 1993). Furthermore, there is no consensus in the psychometrics field as “which one is preferable is a matter of continued debate” (Haladyna, 2004, p. 164). There have been several attempts to empirically differentiate between item statistics derived from the different theories (Fan, 1998; Macdonald & Paunonen, 2002) and to see if test creation using item response theory and using classical test theory produced parallel forms (Lin, 2008). In each of these cases, the theoretical choice did not yield significant differences, further validating the idea that the ease of using a classical test theory trumps the stronger theoretical basis for item response theories. For these reasons, I took a classical test theory approach during the development of the GCI.

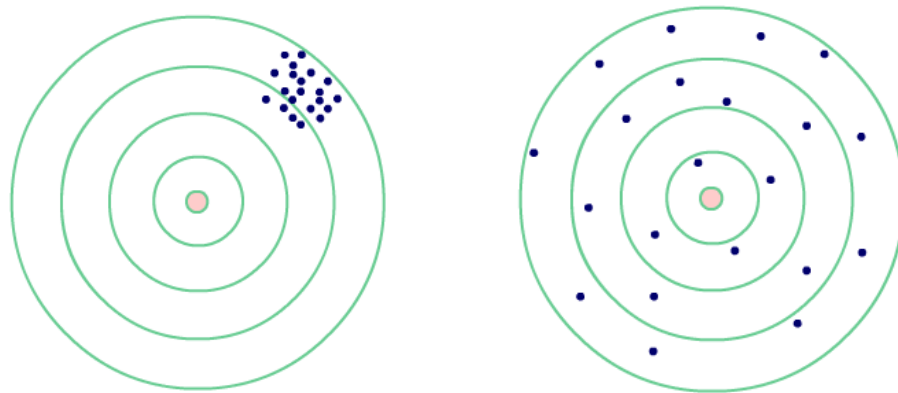
### **Validity and Reliability**

The creation of any sort of measurement instrument involves the dangerous challenge of converting a complex situation into a distilled score. These scores must convey some sort of interpretable meaning whether at the item level or at the instrument level. When creating an assessment, two questions must be addressed:

1. Can this assessment be used repeatedly with consistent results?
2. Are we measuring what we purport to be measuring?

These questions can be addressed using the constructs of reliability and validity respectively. A reliable instrument can be used repeatedly with the same results expected. Reliability could be thought of analogous to precision whereas

validity could be thought of as accuracy. Consider the target in Figure 7. The left target provides a metaphor for a reliable instrument. The scores are not hitting the target, but they are all in the same area and measuring something consistently. The right target would be an instrument that is valid (the scores are centered around the target) but not reliable as the scores are scattered. In the target metaphor, both of these images represent problematic assessment tools.



*Figure 7.* The target analogy for reliability (left) and validity (right). Adapted from Research Methods Knowledge Based by W. Trochim and J. P. Donnelly, 2007, Copyright 2008 by the Atomic Dog Publishing.

**Validity.** The question of validity is a complex one. Within the field of measurement and psychometrics, types of validity are varied and inconsistently named (Trochim & Donnelly, 2008; Lissitz & Samuelsens, 2007). Trochim and Donnelly (2008) defined validity as: “the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations are based” (3-1 Construct Validity, para. 2). An operationalization is the translation of a construct into the concrete assessment. As my goal was to create an assessment that measures group theory understanding, then the concept inventory assessment was the operationalization

of this construct. Validity is then the degree to which this assessment actually represents group theory understanding.

Generally, authors take one of two approaches, either treating types of validity as independent categories or as subcategories of construct validity (Trochim & Donnelly, 2008). I treat validity as consisting of multiple categories where construct validity is of narrower focus.

Since the introduction of the Force Concept Inventory (Hestenes et al., 1992) and Mechanics Baseline Test (Hestenes & Wells, 1992), concept inventory creation has permeated through many subject areas at the university level. However, methods for establishing validity vary greatly and are often underreported (or perhaps just underexplored) (Lindell, Peak, & Foster, 2007). The most common forms of validity explored in concept inventory fall into three categories: content, criterion-related, and construct validity (Lindell, et al., 2007). For the remainder of this discussion I adopt this terminology from Lindell, et al.'s meta-analysis of various physics-related concept inventories. Each type of validity is discussed with respect to their analysis as well as their relationship to Messick's (1995) and Lissitz and Samuelsen's (2007) validity frameworks.

***Content validity – relevance and representativeness.*** Lindell et al. (2007) defined *content validity* as “the degree to which an inventory measures the content covered in the content domain” (p. 15). Messick (1995) captured this idea as *content relevance and representativeness*. Messick argued the need to determine the knowledge that is to be revealed by the assessment tasks. This involves determining the “boundaries and structure of the content domain” (p.

745) as discussed in the prior section. A major purpose is to understand the “construct-relevant sources of task difficulty” (p. 745). The assessment items need to be relevant to the identified domain and its structure. Beyond relevance, the tasks should also be *representative*. All of the important parts of the domain should be covered. Tasks should be selected based on functional importance, what Brunswik (1955) called *ecological sampling* (as cited in Messick (1995), p. 745). A task with functional importance is one that is essential to the domain and is rich enough to successfully unearth potential alternate conceptions. In the GCI, the tasks reflect the important aspects of group theory and target potential areas that would differentiate experts from novices. Messick advocated for the use of experts to appraise representativeness and relevance.

***Construct validity – substantive.*** Lindell et al. (2007) defined construct validity as, “The degree to which scores can be utilized to draw an inference on the content domain” (p. 15). This is roughly translated to what Messick (1995) labeled *substantive* validity and what Lissitz and Sameulsen (2007) labeled *latent process* validity. This type of validity is characterized as “the need to move beyond traditional professional judgment of domain content to accrue empirical evidence that the ostensibly sampled processes are actually engaged by respondents in task performance” (Messick, p. 745). Additionally, processes such as factor analysis can serve to validate that items with expected commonalities are indeed correlated. Whereas content validity may often resemble face validity, substantive construct validity provides the key information needed to have confidence in one’s operationalization. This is particularly true for multiple-



choice tests where complicated cognition is simplified down to a lettered answer. For creating a concept inventory, interviewing students about their choices provides construct validity. By having them share their thought processes and probing them about how their selection choice is related to their conception of various topics, an argument is made that their selection is meaningful and that students are in fact choosing answers for the reasons hypothesized.

*Factor analytic methods.* The meaningfulness of factor analysis in concept inventory creation is controversial. Before discussing the controversy, I introduce factor analysis. The reader is cautioned that this is a brief conceptual overview and not a detailed mathematical breakdown.

Factor analytic methods aim to leverage correlations amongst items to identify underlying or latent variables. An instrument might have 25 items, but they might not all be contributing unique information. Test item outcomes might be a linear combination of other factors. Perhaps there are 25 items, but there might be five underlying factors that account for most of the variation.

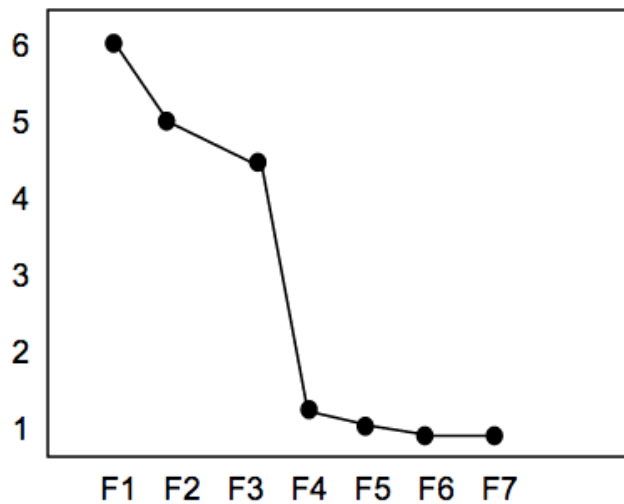
Factor analysis can serve several purposes:

1. Identifying how many latent variables underlie a set of items.
2. Condensing information – eliminating items that might be attributed to the same latent variable.
3. Identifying groups of items that co-vary and attributing meaning to an underlying factor.
4. Eliminating items that do not fit into a factor category or fit into more than one category (Devellis, 2011).

Factor analysis can be confirmatory or exploratory. Confirmatory factor analysis is used when an instrument is designed with specific factors in mind. However, in

the case of the GCI, exploratory factor analysis was more appropriate. Factor analysis can be done to explore unanticipated structure that might exist.

Principal component analysis (PCA) is the method utilized by mainstream statistical software (Tryfos, 1998). Essentially, PCA is an iterative process where factors are extracted. A correlation matrix can be created for all individual items. A first factor (the sum of the items) is used to try and recreate observed correlations. A residual matrix is computed by subtracting the single-factor predictions from the original correlations. This process can be repeated on the residual matrices to extract additional factors. There are various rules for determining how many factors to extract. One rule of thumb is to only find factors that contribute to more than a single item. (This is done through finding the eigenvalues associated with each factor.) Alternately, a scree plot can be used where factors and their corresponding eigenvalues are graphed. The number of factors can be determined by looking for a sudden drop or “elbow” in the graph. See *Figure 8* that illustrates a situation with three underlying factors. Once the number of factors is found, the matrix can be rotated to maximize interpretability of the factors.



*Figure 8.* Scree plot. Factors vs. Eigenvalues

If an instrument is designed with expected factors, then this type of analysis can serve to verify that the expected factor structure did exist. In the case of the Force Concept Inventory, Heller and Huffman (1995) challenged the validity of the instrument based on their factor analysis. The FCI taxonomy has six dimensions. Heller and Huffman attempted to confirm this structure in the assessment instrument through factor analysis. They found no factors accounting for a substantial amount of variance arguing that, “from the students’ point of view, the FCI does not appear to test for a coherent, universal force concept, dimensions of a force concept, or any organized alternative beliefs” (p. 503). Hestenes and Halloun (1995) challenged that this is not a flaw, but in line with the established incoherence in students’ systems. The population of students does not have a coherent Newtonian idea of force and therefore, the underlying factor structure would not be apparent based on this population.

Heller and Huffman (1995) further argued that without a factor structure, students with low scores might have a lack of coherent understanding, but that students with high scores do not necessarily equate with having a coherent system. If there is no evidence that the test itself reflects the coherent system, then score-based claims are limited. If you have a coherent system, you should score well on the test; however, scoring well on the test does not guarantee you have a coherent system. Not arriving at a claim of coherence does not detract from the meaningfulness of the assessment. In fact, a concept inventory is significantly more nuanced than what a factor analysis can account for. The meaningfulness is not in the simplified right and wrong answers, but rather that each multiple-choice question targets a multitude of conceptions around a given task and concept. Factor analysis treats the questions as dichotomous and loses the richness of diagnostically meaningful answer choices. A successful concept inventory targets areas of incoherence and that may also reflect less than optimal factor structure. As argued by Lissitz and Sameulsen (2007) the meaningfulness of an assessment comes first and foremost from establishing that the tasks are eliciting the expected behaviors. If an assessment is deemed relevant and representative and each distractor is explored qualitatively to associate its selection with student thinking, the assessment is undoubtedly meaningful.

***Criterion validity.*** Lindell, et al. (2007) defined *criterion validity* as, “The degree to which scores on an inventory predict[s] another criterion” (p. 15). Messick (1995) referred to this type of validity as *external* where an instrument is correlated with some sort of external behavior or measurement. Lissitz and

Samuelsen (2007) argued that criterion validity is not a type of validity, but rather a measure for *utility*, an external factor. Their argument was that a high correlation between measure A with an external measure B might indicate that measure A has high utility for predicting measure B rather than they measure similar constructs. Often grades are correlated with concept inventories. However, a lack of correlation between the two does not invalidate measure A. If measure A has strong construct and content validity, then lack of criterion validity would not invalidate the instrument. Lindell, et al. found that only one of the twelve concept inventories in their analysis reported on criterion validity, perhaps strengthening Lissitz and Samuelsen's argument that criterion validity is an indirect measure of validity.

***Other validity related considerations.*** Both Messick (1995) and Lissitz and Sameulsen (2007) presented additional aspects of validity. Of particular note, Messick cautioned that *construct-irrelevant variance* can occur because of undue *difficulty* and *easiness*. Unintended difficulty can emerge when non-related issues such as poor-reading abilities impede a student's ability to demonstrate knowledge in another domain. Unintended easiness can emerge when context clues might lead students to the correct answer without having the domain knowledge. These are concerns that may be addressed during the follow-up interviews with students. Messick (1995) also included an aspect of validity based on consequences and potential biases. In the statics concept inventory, the authors partially addressed this issue by making sure there were no significant differences in mean scores based on gender and race. While what is required to establish

validity varies (Lissitz and Sameulsen, 2007), the more validity components addressed, the better the argument that an instrument is unbiased and meaningful.

**Reliability.** Trochim and Donnelly (2008) defined reliability as, “repeatability or consistency. A measure is considered reliable if it would give you the same result over and over again” (3-2c Theory of Reliability, para. 2). Reliability can be estimated in several manners such as test- retest, parallel forms, and internal consistency. Parallel forms reliability is useful with large samples and item banks where two parallel tests could be administered to half of the samples. However, this is not a practical form of reliability with the work-intensive item development in concept inventories. Test- retest reliability is determined by given a test to a sample at one point in time and then again (to a similar) sample at another point in time. A reliable test will produce similar scores on both occasions (all else being equal.) While this has been done occasionally in concept inventories (Olds, Streveler, Miller, & Nelson, 2004; Smith, Wood, & Knight, 2008), an internal consistency approach is more common as it does not require separate administrations.

As explained in Trochim and Donnelly (2008), internal consistency is frequently estimated by Cronbach’s alpha. Cronbach’s alpha is equivalent to using a split-half approach. A split-half approach takes the set of items and splits them into two halves. The scores on each half should be correlated. For a given sample, if you split a test into all possible halves and then correlate each pair, the average provides a measure of internal consistency. If students score dramatically different on different halves of the test, there is a likely a reliability problem

indicating error (and not the latent trait measured by test) is contributing significantly to scores. However, there are limitations in this type of reliability exploration. The underlying assumption is that the test measures a single underlying trait. The use of an internal consistency estimate such as Cronbach's alpha can be used to explore reliability, but the dimensionality of the test mitigates the degree that the test should be internally correlated.

**Increasing reliability.** In classical test theory, the defining characteristics of an item are difficulty and discrimination. The item difficulty is defined as the percentage of test-takers who answered correctly. The acceptable range of difficulty varies. If the proportion of students answering a question correctly is less than chance, the item should be reviewed. The question might be misleading or the answer key is wrong. If the proportion of students answering a question correctly is high (such as greater than 85%), then the question is likely too easy and may be answerable through some other clue based on the wording as opposed to their knowledge of the topic.

The second major factor to address is item discrimination. The item discrimination is defined as, "the item's ability to sensitively measure individual differences that truly exist among test-takers" (Haladyna, 2004, p. 166). If we operate under the assumption that the GCI is assessing a related set of concepts or more generally the group concept, students with incorrect answers on an item should correlate with lower overall scores (with the item removed). Students who answer the item correctly should correlate with higher scores on the assessment. An item would have high discrimination if the group of students who answered

the item correctly averaged 90% on the test and the group of students who answered incorrectly averaged 40% on the test. If the averages are about equal in the two groups, the item is not discriminating. If the average is higher for students answering incorrectly, the item is discriminating negatively which might indicate the answer key is incorrect.

Table 2

*Item Classification Guide by Difficulty and Discrimination. Adapted from "Assessment in Health Professions Education," S. M. Downing & R. Yudkowsky, 2009. New York, NY: Routledge, p. 108.*

Item Class	Item Difficulty	Item Discrimination	Description
Level I	0.45 to 0.75	+0.20 or higher	Best item statistics; use most items in this range if possible
Level II	0.76 to 0.91	+0.15 or higher	Easy; use sparingly
Level III	0.25 to 0.44	+0.10 or higher	Difficult; use very sparingly and only if content is essential – rewrite if possible
Level IV	<0.24 or >0.91	Any Discrim.	Extremely difficult or easy; do not use unless content is essential

The point-biserial correlation is a standard measure for item

discriminations where a student's correctness on an item (1 or 0) is correlated with their overall score. This correlation coefficient can range from -1 to 1. The higher an item's discrimination, the better the item correlates to the overall test. A negative score might indicate an item is scored incorrectly. Table 2 presents some overall guidelines.

By attending to item discrimination and difficulty, items that are not consistent with the overall test can be altered or removed. If an item is too difficult or easy, it is not contributing useful information to the score. If an item fails to discriminate, the information it contributes is not consistent with the test.



## **Conclusion**

For an assessment to be meaningful it must be valid and reliable. By determining what it means to have conceptual understanding, the scope of a concept inventory was identified generally. However, situating this scope in the context of group theory involves several measures to ensure validity. Domain analysis plays a central role in content validity where textbooks, literature, and experts can help assure the representativeness and relevance of items. Further, surveying and interviewing students provides the means for establishing construct validity. This qualitative analysis can lend evidence that each multiple-choice distractor is in fact reflecting the associated conception. Furthermore, quantitative analyses such as correlating scores with external factors can bolster validity. The essence of a strong assessment is identifying what you want to assess, mapping the related domain, than ensuring meaning by establishing validity and reliability.

## Project Methodology

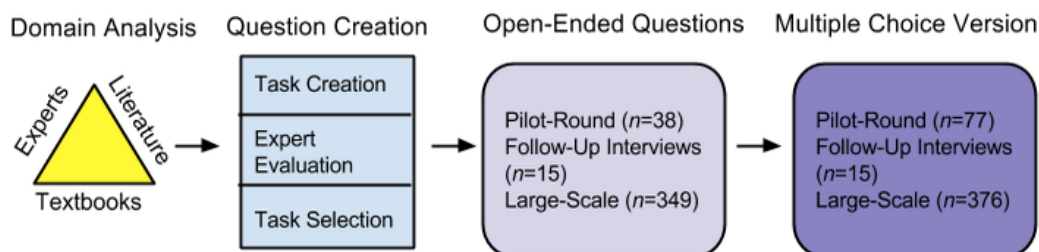


Figure 9. Overall methodology.

The steps involved in the creation of concept inventories stem from the work of Hestenes and his colleagues on the Force Concept Inventory and Mechanics Baseline Test (Hestenes et al., 1992; Hestenes & Wells, 1992). In Lindell et al.'s (2007) meta-analysis of concept inventories, they outlined the general steps that Hestenes and his successors took when designing the instruments:

1. Identify purpose
2. Determine the concept domain
3. Prepare test specifications
4. Construct initial pool of items
5. Have items reviewed – revise as necessary
6. Hold preliminary field testing of items – revise as necessary
7. Field test on large sample representation of the examinee population
8. Determine statistical properties of item scores – eliminate inappropriate items
9. Design and conduct reliability and validity studies (p. 15).

While the apparent linearity may be misleading, the general framework presents a guideline for creating a concept inventory. The first part of the GCI methodology focused on determining the concept domain. The second part focused on going from a concept domain analysis to the creation and validation of a concept assessment. Table 3 presents an overview of GCI creation timeline.

Table 3

<i>Timeline of GCI Development</i>						
Delphi Study/ Textbook Analysis/ Literature Search	Develop/ Evaluate Tasks	Pilot Open- Ended Tasks	Large-scale Open-ended Round	Follow-Up Interviews	Pilot Multiple- Choice Version (and follow-up interviews)	Large-Scale Multiple- Choice (continued follow-up interviews)
Summer 2014	September 2014	October 2014	December 2014	January 2015	March 2015	April-May 2015

### Concept Domain Analysis

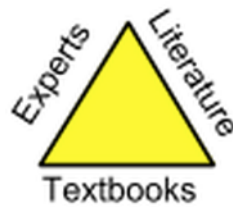


Figure 10. Domain Analysis Triangle

Typically, the creation of a concept inventory relies on a taxonomy of foundational understandings about the area to be covered (Carlson, Oehrtman, & Engelke, 2010; Hestenes, Wells, & Swackhamer, 1992). However, such taxonomy did not exist for group theory. A domain analysis was used to achieve the taxonomy goals.

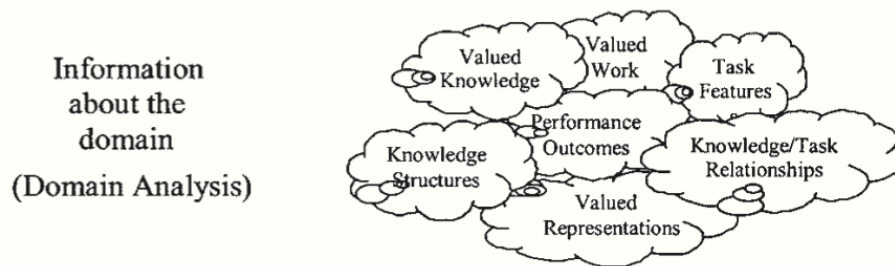


Figure 11. Components of the domain. Reprinted from “Focus article: On the structure of educational assessments,” by R.J. MisLevy, L.S. Steinberg, and R.G. Almond, 2003, *Measurement: Interdisciplinary Research & Perspective*, 1, p. 6. Copyright 2003 by Lawrence Erlbaum Associates, Inc.

A domain analysis aims to identify the “boundaries and structure of the content domain” (Messick, 1995, p. 745) needing to be determined. This is an essential part of creating an instrument to measure understanding of any domain. Prior to creating any tasks, assessment developers should determine what is important and relevant within the domain to be measured. This helps to develop an argument for content validity (the scope and relevance of questions to the targeted domain.) MisLevy, Steinberg, and Almond (2003) explained domain analysis as:

...marshaling substantive information about the domain—bringing together knowledge from any number of sources and then beginning to organize beliefs, theories, research, subject-matter expertise, instructional materials, exemplars from other assessments, and so on (p. 7).

The analysis of the domain can be broad and first requires identifying what is to be analyzed. In the case of the GCI, the conceptual domain surrounding group was the intended domain. However, identifying what it meant to *understand* a concept and what are the fundamental related concepts was not a trivial task. The domain’s boundaries were established via expert surveying, textbook analysis, and literature consultation.

**Delphi study and expert panel.** The first step in developing the GCI was to conduct a Delphi study. A Delphi study is a “widely used and accepted method for achieving convergence of opinion concerning real-world knowledge solicited from experts within a certain topic” (Hsu & Sanford, 2007, p. 1). Through a series of iterative rounds, experts arrive at a consensus on a topic. A Delphi study allows

for participants to maintain anonymity from one another while interacting with one another's ideas.

A panel of 13 experts was gathered with the goal of heterogeneity. Each member of the panel had taught the course at least two times. I selected the panel to be heterogeneous in order to best represent the variety of views that may exist relevant to introductory group theory. Aligned with Streveler et al.'s (2011) recommendations, the panel was geographically dispersed and contained experts with varying backgrounds including textbooks authors. The panel had four abstract algebra textbook authors, eight mathematicians (with a range of specialties including math history, analysis and several group theory specialists), and five mathematics education researchers who have published related to abstract algebra pedagogy.

The panel was surveyed four times:

**Pass 1.** The panelists listed what they considered to be the essential topics in group theory. I then compared these responses and selected all topics that appeared on at least two lists.

**Pass 2.** The panelists were presented with the list of topics from the first pass. The panelists then rated each topic on a score from 0-10 for importance and 0-10 for difficulty. I then compiled the ratings in order to present the median, mode, and interquartile range (IQR).

**Pass 3.** During the third pass, the panelists reconsidered their ratings based on the descriptive statistics provided. The experts assigned a new point value from 0-10 in the two categories. If a panelist assigned a number outside of the

IQR, I asked them to provide justification for their responses. At the end of the round, I compiled the same descriptive statistics.

**Pass 4.** The panelists rated each topic in terms of importance and difficulty one final time. They rerated after considering the justifications and descriptive statistics from pass 3.

At this point, I selected all topics that had a mean score of at least 9 out of 10 to compile a list of fundamental topics in introductory group theory.

The Delphi process provides some advantages over other methods of eliciting expert opinions. Unlike a round table discussion, responses are anonymous and so experts are not going to switch opinions based on perceived importance of individuals. However, unlike individual interviews, this process does allow for experts to consider and reflect on their peers' opinions (Sterveler et al., 2011).

**Textbook analysis.** The next phase involved analyzing textbooks in order to analyze the treatment of the topics selected through the Delphi Study. These topics were: the First Isomorphism Theorem, homomorphism, associativity, inverses, identity, quotient groups, cosets, subgroups, normal subgroups, kernel, abelian groups, isomorphism, cyclic groups, binary operations, group, Lagrange's theorem, and the order of an element. Representative textbook analysis has been used frequently in the development of concept inventories (Lindell et al., 2007; Stone et al., 2003; Wage, Buck, Wright, & Welch, 2005). However, exactly how textbooks informed the creation of the tool is largely unaddressed in the literature. This textbook analysis had purposes aligned with domain analysis goals.

Particularly, the analysis consisted of coding various aspects of concept-understanding including: related tasks, representations, formal and informal definitions, and examples. Mesa (2004) explained the textbook treatment of topics as the *intended* conceptions of topics.

In order to identify the typical group theory textbooks, I selected a random sample of United States colleges and universities that offer mathematics majors. The sample size was 294 representing the 1,244 US institutions with mathematics majors in order to achieve a 95% confidence interval about the proportion of institutions (+/-5%) using each textbook. The textbook usage was identified based on what text is used in the current term or most recent term the course was offered. I first consulted online resources such as syllabi or online bookstores. When unavailable, I contacted instructors and mathematics departments directly. This led to a degree of response bias, however, there is no reason to believe that the responding institutions were noticeably different.

Initially, each textbook used by at least 20 institutions was included for analysis. However, this resulted in three textbooks. The fourth textbook, the only other textbook with substantial usage, was also included. For these purposes, different editions of the same texts were treated as the same textbook. The textbooks analyzed can be found in Table 4.

Table 4

<i>Abstract Algebra Textbook Analyzed</i>		
Title	Edition	Author
<i>Contemporary Abstract Algebra</i>	7 <sup>th</sup>	Joseph Gallian
<i>A First Course in Abstract Algebra</i>	7 <sup>th</sup>	John B. Fraleigh
<i>Elements of Modern Algebra</i>	7 <sup>th</sup>	Linda Gilbert & Jimmy Gilbert
<i>Abstract Algebra, An Introduction</i>	3 <sup>rd</sup>	Thomas W. Hungerford

**Pass 1.** The first textbook pass was to identify the relevant sections of the text. For the purposes of this analysis, I selected all sections of the text that introduced the relevant topics. In general, this represented the first several sections of the textbooks, or in the case of one text, a subsection of the book dedicated to groups.

**Pass 2.** The next pass was a more detailed exploration aimed to align with the various aspects of conceptions related to each of the topics. My categories for analysis consisted of several types: intuitive/formal definitions (Raman, 2002), representations (Mesa, 2004), expected student activity (Otten, Gilbertson, Males, & Clark, 2014; Tallman & Carlson 2012), and examples (Fukawa-Connelly, & Newton, 2014). (See the Theoretical background for a more thorough discussion of these aspects of conceptual understanding.) The textbooks were coded using a thematic analysis approach (Braun & Clarke, 2006). Initially, I began with a set of codes in each category adapted from the respective sources. However, I refined the codes in order to better capture representations and activities related to group theory topics, and I changed example purposes to better capture the range found



in textbooks. All exercises, examples and definitions were also tagged with the relevant topic from the Delphi study. A complete set of codes and descriptions can be found in Appendix A.

The textbook analysis goals were two-fold: determine what is valued in the field in terms of representations, activities, and examples and determine which examples and representations typical introductory students would have access to. The goal is to present a snapshot of the textbooks while also identifying warrants for selecting tasks later.

Figure 12 illustrates a formal and informal discussion of a topic. I coded the sections with the codes *formal* and *informal*. I then created analytic notes to further parse the informal discussion. These notes can then be compared across textbooks to look for commonalities in informal discussion of topics.

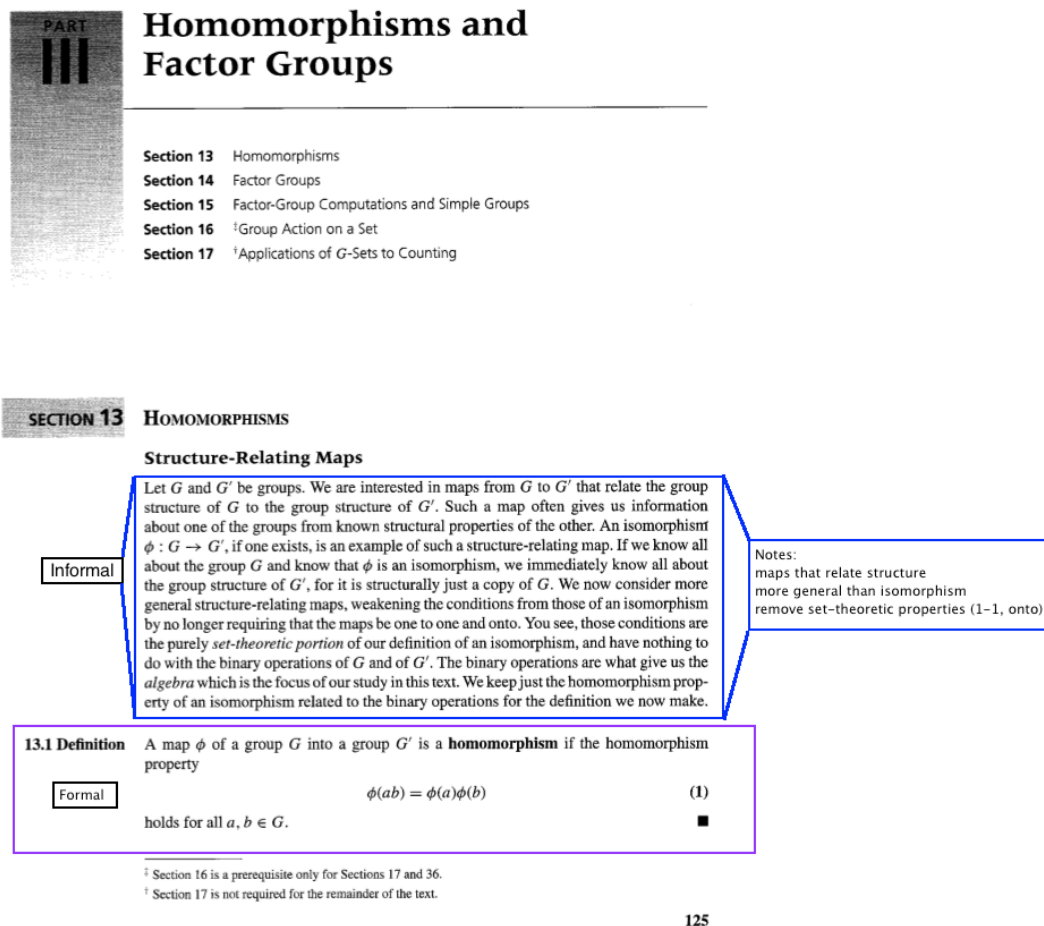


Figure 12. Example of formal and informal definition and descriptions of homomorphism from Fraleigh (2002).

Figure 13 includes an exercise that was coded during this process. The task was first coded for relevant topics: *group* and *homomorphism*. The exercise received representation codes: *group-verbal* (“the additive group of functions mapping  $\mathbf{R}$  into  $\mathbf{R}$ ”), *function-verbal* (“the second derivative”), and *function symbolic* ( $\Phi(f)=f''$ ). The expected student activity was *evalaute* as the students were prompted to determine if the map was a homomorphism. Beyond the initial codes, analytic notes identified exactly what the activity is: *determine if a given map is a homomorphism*.

In Exercises 1 through 15, determine whether the given map  $\phi$  is a homomorphism. [Hint: The straightforward way to proceed is to check whether  $\phi(ab) = \phi(a)\phi(b)$  for all  $a$  and  $b$  in the domain of  $\phi$ . However, if we should happen to notice that  $\phi^{-1}[[e']]$  is not a subgroup whose left and right cosets coincide, or that  $\phi$  does not satisfy the properties given in Exercise 44 or 45 for finite groups, then we can say at once that  $\phi$  is not a homomorphism.]

9. Let  $F$  be the additive group of functions mapping  $\mathbb{R}$  into  $\mathbb{R}$  having derivatives of all orders. Let  $\phi : F \rightarrow F$  be given by  $\phi(f) = f''$ , the second derivative of  $f$ .

Figure 13. Homomorphism exercise from Fraleigh (2002).

Figure 14 contains an example instance of homomorphism. First, the topic was identified: *homomorphism*. Then the particular example was identified:  $S_n$  to  $\mathbb{Z}_2$ . The representations involved were: *group-table*, *group-name*, *function-diagram*, and *function-verbal*. The homomorphism was described in words, but then illustrated with a visual of two Cayley tables with lines indicating mapping. The example purpose was: *example following a definition*. (Note: this example would also have an informal code as it serves to illustrate “telescoping nature of the mapping.”)

■ **EXAMPLE 11** The mapping from  $S_3$  to  $\mathbb{Z}_2$  that takes an even permutation to 0 and an odd permutation to 1 is a homomorphism. Figure 10.2 illustrates the telescoping nature of the mapping. ■

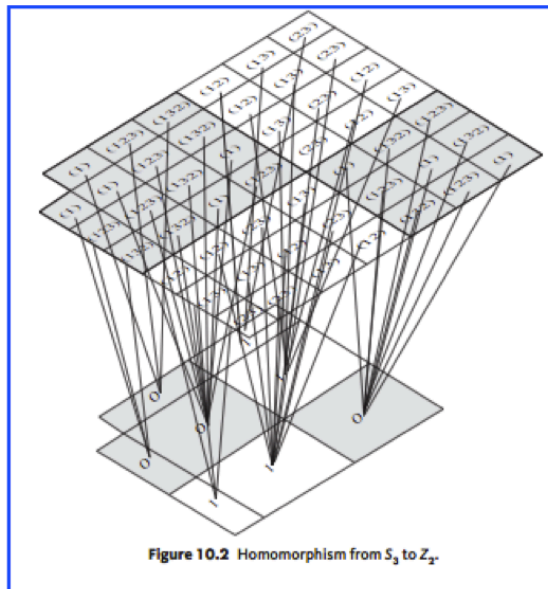


Figure 14. Homomorphism example from narrative in Gallian (2009).

A second coder, a mathematics education PhD student, coded one complete section from each textbook. She coded a section related to a different topic for each text in order to try and cover a variety of codes that may emerge. A measure for inter-rated reliability was determined. For both the exercises and examples, the agreement between coders was over 80%. That is, a second coder identified at least 80% of the same codes.

The textbook analysis served to inform the creation of analytic reports on each topic. The reports contained frequencies of representations, types of examples and for what purpose they were used, formal and informal definitions, and expected student activity around exercises. An example analytic report can be found in Appendix B.

I ultimately leveraged these reports for purposes of access, value, and representativeness. In terms of access, the curriculum or textbook used in an introductory text should not prevent a student from being able to make sense of an item in the GCI. The GCI only made use of representations found throughout all texts in a relatively frequent basis. For example, groups are presented symbolically and verbally or through the usage of a Cayley Table. A lesser-used representation, such as left-hand representation, was not incorporated into the GCI. Further, examples were limited to those treated frequently throughout textbooks, notably common groups (such as real number or integers) as well as modular groups (a typical example used to illustrate or motivate a number of topics across texts.) These choices also reflected what appeared to be valued in this curriculum. Because tables were a consistent representation, Cayley tables

were used in items. This is not to be confused with creating a GCI that reflects textbooks in nature. In fact, many of the textbook activities were procedural, involved advanced-proofs, or were otherwise not conducive to the ultimate goal of creating tasks that captured *conceptual understanding*. This analysis also served to illuminate certain activity-types that were procedural in nature. The biggest asset of the textbook analysis was to create an awareness of what students have the opportunity to learn.

**Literature consultation.** After the Delphi study and initial textbook analysis, the literature base served as the third leg of triangulation. Particularly, topics and tasks identified as essential were explored in terms of known *student conceptions* in the literature. This was done in as comprehensive of a manner as possible. In the first pass, I conducted a full text search for each concept in the 35 journals identified by SIGMAA on RUME as research-based and relevant to undergraduate mathematics (RUMEonline!, 2011). I then conducted a title search within relevant conference proceedings including conferences held by the International Group for the Psychology of Mathematics Education, the Special Interest Group of the Mathematical Association of America on Research in Undergraduate Mathematics Education, the Congress of European Research in Mathematics Education, the International Congress on Mathematics Education and the Southern Hemisphere Conference on the Teaching and Learning of Undergraduate Mathematics. Finally, I searched for “Group Theory” and “Abstract Algebra” in the ERIC database to assure I had not overlooked relevant literature.

For each essential topic, I compiled reports that included both attention to student conceptions and what tasks may be relevant for the creation of the GCI.

An example report can be found in Appendix C.

**Task creation.** After the completion of the textbook analysis and literature search, I created an initial set of 42 open-ended tasks. I directly adapted tasks from the literature whenever possible. However, many of these tasks contained example groups that were not treated consistently across textbooks. In these cases, the tasks were altered or new tasks aimed to target the same conceptions were developed. Further, several tasks were created to target informal understanding of concepts identified in the textbooks. The majority of the tasks were structured to be “evaluate” type problems. This was done to allow for the maximum number of conceptions to emerge. By leaving the truth-value of the statement unknown, students may have a number of ways to agree or disagree with the question. The examples and representations used were adapted from those common to textbooks.

**Expert task evaluation.** Every task was evaluated by at least four experienced group theory instructors. Two of the evaluators primarily focused on research in mathematics education and two of the evaluators focused on mathematics research. They were asked to evaluate:

1. Is this task relevant to an introductory group theory course?
2. Does this task represent an important aspect of understanding [relevant concept]?
3. If you said NO to either of the above, indicate why. (For example, the task might be too advanced, capture prerequisite knowledge or be unrelated to the targeted concept.)

4. What student conceptions might this task capture? (You could speak generally, or provide samples of how you think students might respond.)

This evaluation served the purpose of eliminating tasks that may be inappropriate for the GCI for a lack of importance or relevance. The expert feedback was also used to refine questions before piloting them.

I analyzed the open-ended portions of the survey responses using a thematic analysis approach (Braun & Clark, 2006). This was done in a series of passes.

Pass 1: All responses were read for initial ideas. Initial yes/no responses to the importance and relevance questions were identified. The primary goal of pass 1 is to become comfortable with the data.

Pass 2: The evaluations were open-coded. This took the form of analytic notes capturing the underlying ideas in each.

Pass 3: These notes were organized into themes and number of categories emerged within each question. Overall categories of codes included:

1. Minor wording alterations
2. Content modifications (often suggestions of changing example used)
3. Concerns of eventual ability to be multiple choice question
4. Concerns of difficulty
5. Potential Incorrect or Incomplete Conceptions

Appendix D includes a sample analysis of expert evaluations on two questions related to Abelian groups.

After I completed the coding, I first attended to the relevance and importance of issues to select tasks. Any task that more than one evaluator rated

as unimportant or not relevant was eliminated. I then selected one for each concept using the following hierarchy:

1. Task has been studied in literature.
2. Task can be connected to literature.
3. Experts identified the tasks as having the power to unearth student conceptions.
4. Little concerns of difficulty, scope or ability to convert to multiple-choice.

I also altered and refined many of the tasks that were ultimately selected based on the expert feedback. In general, these alterations were minor and involved slight changes to wording or context. However, for the question on kernel, none of the questions evaluated particularly well, and so the question was replaced with one suggested by an evaluator.

### **The Creation of the Assessment**

The creation of the assessment roughly followed Carlson, Oehrtman and Engelke's (2010) *Four-Phase Instrument Development Framework*. Following the domain analysis, I conducted several more phases to convert open-ended questions to multiple-choice questions, and then ultimately refine the multiple-choice questions. First, I administered the open-ended questions to a small pilot group of students to test out wording and that the questions appeared to be differentiating between conceptions. Then, I made alterations and field-tested updated versions of the open-ended questions. I interviewed fifteen students about their open-ended responses. I piloted an initial version of the multiple-choice instrument with some pilot classes. I then refined the instrument and gave it to a



large sample of diverse students. After this phase, I conducted additional psychometric analyses to address reliability and structure of the GCI.

**Phase 1- stem development.** The first step towards the creation of the GCI was designing the open-ended survey. These questions reflected known areas of difficulty from the literature (such as a lack of an object view of cosets (Asiala et al., 1997) or misapplication of Lagrange (Hazzan & Leron, 1996)) while remaining representative of the domain.

I tested the open-ended questions in two stages. First, the questions were given to 38 students who had recently completed group theory or had finished covering the relevant topics in their current course. These students came from five different institutions. This sample was one of convenience. The classes were local to the Pacific Northwest, or came from classes with instructors I knew (from the Mid-Atlantic and Southeast.) The pilot allowed for four early interviews and initial testing of the questions to make sure they did not have wording issues and that they had the power to target the conceptions intended. At this point, I altered some questions in minor ways and one question was replaced (see Paper 1 for detailed explanation.)

After the pilot, I administered the open-ended survey to 349 students from 30 classes at 29 institutions. I intentionally aimed to make this sample representative. Institutions with mathematics majors were selected randomly with respect to selectivity (least selective ( $>75\%$  admitted), mid-level selective (50-75% admitted), more selective (25-50% admitted), and most selective ( $<25\%$  admitted)) and geographic regions (West, Midwest, Southeast, Northeast, and

Mid Atlantic). I then conducted fifteen interviews with local students to assure that I was interpreting their answers correctly. The students were prompted to explain their thinking for each answer they provided on the open-ended survey. Furthermore, if their responses did not connect to their understanding of a relevant concept, they were asked to explicate their understanding of the intended concept.

**Phase 2 – the creation of the multiple-choice instrument.** Based on the results from the open-ended round, I created a multiple-choice version of the assessment. A multiple-choice item consists of a stem, a key, and distractors. A stem is the command or question, a key is the correct choice, and distractors are plausible incorrect choices (Haladyna, 2004). For the GCI, the stems came from the open-ended questions with minor refinements. Refinements included removing justification prompts. The open-ended version of questions included these additional prompts in order to assure that students shared their reasoning. However, in multiple-choice questions, this is an unnecessary concern as the closed-form allows the degree of detail to be determined by the test creator.

I developed the distractors for the multiple-choice items by analyzing the 349 open-ended responses. For each question, I developed a set of codes that represented the various answer types that emerged. Examples of these codes can be found in Paper 1 and Paper 2. The codes began open, but eventually became a stable set of codes after an initial coding of 200 responses. I then applied the stabilized codes to all of the data. I used frequency reports to determine which

responses were most common and converted those responses into the multiple-choice versions of the questions.

During this process, I also attended to item-writing guidelines when possible. In Haladyna, Downing, and Rodriguez's (2002) analysis of textbooks and studies on item-writing guidelines, they found many guidelines, but only a few that have been empirically studied. The empirically studied guidelines include avoiding "all of the above" options as students pick them frequently even if they only agree with a subset of the other options. This makes it impossible to differentiate between students who agree with different subsets of responses. The "all of the above" option dilutes the meaningfulness. They also found inconsistent results about the effect of the number of distractors on difficulty and discrimination. The optimal number of distractors lay somewhere in the two to four range. In most tests analyzed, only two distractors tended to represent genuine misconceptions even when other distractors were included. For the GCI, I decided to create only the number of distractors that appeared to genuinely connect to student conceptions. This meant that the number of distractors varied depending on the question. These guidelines informed the creation of the multiple-choice items to a great extent; however, remaining faithful to distractors arising from the open-ended surveys remained the primary concern. The first version of the multiple-choice instrument also contained "none of the above" options in order to allow for new conceptions to emerge.

**Phase 3- piloting the multiple-choice instrument.** At this point, I piloted the multiple-choice assessment with 77 students from 8 classes. During January of

2015, I attempted to contact all instructors of current undergraduate group theory courses at institutions with mathematics majors. This included all classes titled: Group Theory, Abstract Algebra, Modern Algebra, or Algebraic Systems. If course schedules were available, I used them to identify courses that were running. If not, I (or a member of my team), called the institution. From this process, 39 instructors agreed to have their classes participate. The classes that covered the relevant material by March were included in the pilot round. While the overall sample was representative in terms of selectivity, the pilot round classes overrepresented highly selective schools.

At this stage, I analyzed the students' responses in terms of item difficulty, item discrimination, and the effect of various questions on reliability. (See Appendix E for these results.) If any items had failed to discriminate or had too low of difficulty, I would have removed or altered them from the GCI. However, that did not occur at this point.

***Item refinement.*** During this time, I also refined distractors. Following Carlson et al.'s (2010) suggestions, the optimal way to do this and increase validity is through student interviews. I interviewed students who selected various distractors to confirm or disconfirm that their thinking matched the distractor intentions. Carlson et al. also suggest deleting distractors when less than 5% of students select them. I calculated the frequencies of all responses and deleted the "none of the above" response if selected by less than 5% of students. I also used the interviews to address the "none of the above" responses that remained. When possible, I converted to a concrete distractor. Alternately, if a student selected

“none of the above” for a reason not related to the questions’ intentions, I altered the questions to try and eliminate unintended variance. At this point, the updated GCI was ready to be field-tested.

**Phase 3 – field-testing.** The remaining sample classes used the refined version of the GCI. At this point, 376 students from 30 institutions, representing all levels of selectivity, took the GCI. I then completed the remaining of the 15 follow-up interviews. I also analyzed the GCI’s psychometric properties in terms of reliability, item analysis, factor analysis, and criterion-related validity.

Each of the items was evaluated in terms of discrimination and difficulty with this more diverse sample. The results can be found in Appendix E. Any questions with too low of difficulty or discrimination were marked for further exploration in the next round of development.

At this stage, I correlated the instrument with student self-reported grades on a 4.0 scale. Grades in abstract algebra should ideally be positively correlated with a measure of group understanding. I utilized Pearson’s  $r$  to determine correlation. If my goal was predictive, some of the nuances such as variability attributed to classes would need to be addressed. For establishing correlation, Pearson’s  $r$  is both sufficient and established as the standard approach in concept inventory literature (Carlson, Oehrtman, & Engelke, 2010; Stone et al., 2003).

I then calculated Cronbach’s alpha an estimate of reliability in order to verify that there is some level of internal consistency in the instrument. As this is a low-stakes test that is not developed to be entirely unidimensional, Cronbach’s alpha provided rough information about the reliability of the test. Any alpha less

than 0.5 is unacceptable, between 0.5 and 0.6 is poor, between 0.6 and 0.7 is acceptable, between 0.7 and 0.9 is good, and greater than 0.9 is excellent (George & Mallery, 2003).

**Additional analysis.** I used principal component analysis to explore the dimensionality of the GCI. Epstein (2007) used exploratory factor analysis with the Calculus Concept Inventory. Since I did not enter the development of this assessment with a set of predetermined factors, I similarly used an exploratory approach. By exploring the correlation between items, some underlying relationships might have been made apparent. In the Calculus Concept Inventory, this analysis revealed two factors: functions and derivatives. In the GCI, the questions were created to intentionally target specific concepts that were loosely related to some underlying understanding of introductory group theory. Ultimately the analysis served to illustrate this structure.

## **Conclusion**

The creation of the GCI involved careful attention to developing a meaningful instrument. Through domain analysis using textbooks, expert opinions, and literature on student understanding, an initial exploration of the domain was made in order to make an argument for content validity. Ultimately, I created or adopted tasks that were closely tied to student conceptions surrounding topics, were accessible to the general population of group theory students, and used representations and examples that were valued and available in curricula. Through wide-scale surveying and follow-up interviews, genuine student answers served as the foundation for the multiple-choice options. The multiple-choice

version of the assessment was further refined and tested through item analysis, reliability calculations, and continued follow-up interviews to provide the strongest case for validity and meaning.

**Paper 1: Going from nothing to something: A Multi-faceted approach to developing questions for concept inventories as applied to introductory group theory**

At the heart of any measure is validity. We want to be able to claim that we are in fact measuring what we purport to be measuring. While this may seem like an obvious goal, the reality is that establishing validity is often a complex and challenging aspect of measure design. In undergraduate science and mathematics fields, the concept inventory style of assessments represent a large step forward in terms of strengthening the connection between student conceptions and what is actually measured in an assessment. Concept inventories have flourished in undergraduate sciences and mathematics since the early work on the Force Concept Inventory (Hestenes, Wells, & Swackhamer, 1992) and the Mechanics Baseline Test (Hestenes & Wells, 1992). The growing number of concept inventories all share certain qualities: (1) multiple-choice format, (2) concept-focused rather than procedurally focused assessment, and (3) connected to known student conceptions. If a concept inventory is valid, it should be shown to address concepts over procedures and elicit student responses that reflect the conceptions the measure was designed to target. However, beyond the stated goals, there exist implicit aspects related to validity. Are the questions related to the subject domain being targeted? Are the questions important to that subject area? A measure cannot cover all aspects of a given subject area and so the questions in the instrument must aim to be as representative and reflective as possible. This article presents a methodology developed to maximize validity through attention to



question selection, meaningfulness of responses and minimizing construct-irrelevant variability.

The methodology shared in this paper was developed in order to maximize validity during the creation of an introductory group theory concept inventory, the Group Concept Inventory (GCI). The current concept inventories in mathematics exist within subject areas typically aligned with service courses. In fact, only the Calculus Concept Inventory (CCI) (Epstein, 2006) and Precalculus Concept Assessment (PCA) (Carlson, Oehrtman, & Engelke, 2010) were developed within mathematics departments. Calculus and Precalculus both have significant literature on student conceptions in their respective domains. In a non-service course such as a group theory, the literature base is less extensive leaving a large gap that needs to be traversed to establish a valid question set. This paper introduces a methodology created for the development and refinement of questions in the GCI. The methodology is divided into three parts:

1. Domain analysis
2. Question creation and selection
3. Question refinement, replacement and elimination

Examples from the GCI creation illustrate each of the steps with the overall goal of being transparent and systematic to best connect the methodology to validity goals.

Within these stages, I leveraged prior concept inventory creation methodologies when possible. For the domain analysis, I used an expert consensus protocol that has been previously used in the creation of a thermal and transport concept inventory (Streveler, Olds, Miller, & Nelson, 2003). This

protocol led to the selection of fundamental group theory topics. After a list of basic topics to cover, I explored them in two ways: through reviewing literature related to student conceptions and analyzing curricula using textbooks as a proxy. Both textbooks and literature have been used consistently to inform the process of creating concept inventories. However, in general their use is often described vaguely with textbooks serving more to verify the existence of topics in curriculum rather than concretely analyzing both how concepts are approached and to what specific examples and representations students may have access. The GCI creation outlines a careful approach to textbook analysis, a comprehensive approach to reviewing literature, and discussion as to how these steps directly informed the creation of the question set.

The second half of the process is shared in a manner to complement the creation process of the PCA (Carlson, et al., 2010). Experts are often used to vet questions in concept inventories. I leveraged experts to specifically address concerns of content validity, evaluate the questions for technical issues, and share their knowledge about how students may approach the questions. After ultimately selecting the questions with the greatest potential to unearth known student conceptions, I followed the process of going from open-ended questions to closed-form questions outlined in the PCA creation. I present the evolution of four questions with various paths ranging from direct adaptation from open-ended to closed-form (similar to the PCA question-creation process) to complete deletion and replacement of questions. Finally, I conclude with discussion of

traditionally used psychometric approaches and their limitations and strengths in the context of a diagnostic instrument such as a concept inventory.

### **On Validity and Concept Inventories**

In this section, I provide a background on validity concerns and test creation, as well as concept inventories in general and their creation processes.

**What does it mean for a test to be meaningful?** Underlying any assessment are a multitude of subjective components that together determine the meaning of a given instrument. Pellegrino, Chudowsky, and Glasser (2001) presented these components in terms of three dimensions: *observation*, *interpretation* and *cognition*. If we are measuring cognition, we do so via observing students engaged in a set of tasks and interpreting what we are seeing. The cognition aspect relies on a “theory or set of beliefs about how students represent knowledge and develop competence in a subject domain” (p. 44). In concept inventories, this component frequently comes from prior studies and relevant theory for the educational research. We cannot access a student’s cognition directly, but we can use evidence to develop theories about how they conceive of various concepts. The observation dimension “represents a description or set of specifications for assessment tasks that will elicit illuminating responses from the students about the target domain to be measured” (p. 48). The tasks, multiple-choice questions in the case of concept inventories, must succeed in two manners: unearthing student conceptions and representing the relevant domain. The tasks must connect to the theories of cognition, but then be interpreted correctly. The interpretation corner “encompasses all the methods

and tools used to reason from fallible observations” (p. 48). It is the interpretation of the observation that is at the heart of validity. This is the place where meaning is added to what is observed. A valid instrument must coordinate these aspects.

Due to its complex nature, validity is often subdivided into a number of types. Broadly validity can be thought of in terms of external and internal factors (Lissitz and Samuels, 2007). Internal validity comes from the test itself. For example, think-aloud protocols can be utilized to add meaning to students’ written responses to bolster claims of their interpretations. External measures of validity come not from the test itself, but from its relationship to other measures and external usage. As argued by Lissitz and Samuels, this latter category is not essential to the meaningfulness of an instrument. An instrument with internal validity can stand independently and remain meaningful. For example, if group theory grades do not correlate with the GCI scores, the reasons may be more closely related to differences in what is measured rather than any validation issues with the GCI. The GCI targets concepts while a traditional course may only ever assess students’ ability to produce formal proofs. A lack of correlation would not lessen the meaning of the GCI, but rather only raise questions about the differences and connectedness of traditional assessment and conceptual understanding in the course.

Internal validity can be divided into three categories: *content*, *construct*, and *generalizability*. Before continuing, it is worth noting that the names given to types of validity vary considerably. I adopt Lindell, Peak, and Foster’s (2007) usage of content and construct validity from their meta-analysis of concept

inventories. Content validity is the “degree to which inventory measure the content covered in the content domain.” The content domain is whatever area is meant to be assessed. In the case of the GCI, the content domain is introductory group theory. In order to argue that an instrument has content validity, the test developer must determine what exists in a given content domain, as well as how that measure aligns with the domain. Construct validity is defined as, “[t]he degree to which scores can be utilized to draw inference on the content domain.” This validity stems from the degree a student’s answer can be interpreted. The answer should connect to a student conception related to some aspect of the domain. Typically, student interviews are utilized to establish construct validity. This type of validity is consistent with what Messick (1995) terms the *substantive aspect*. Generalizability is the extent to which the test generalizes across time and place. Traditional measures of reliability such as parallel tests or internal consistency estimates fall under this umbrella. Variation in test scores should only be related to the construct the test intends to measure. If responses are affected substantially by other error, the test will not be reliable. Lissitz and Samuelson (2007) address all aspects of generalizability in their reliability factor of internal validity. A test must necessarily be reliable to be valid. While the constructs of validity and reliability are often treated as dual constructs, a more natural way to address reliability is to include it as a subcomponent of validity such as seen in both Lissitz and Samuelson, and Messick’s validity frameworks.

Validity generally fails in one of two ways: *construct underrepresentation* and *construct-irrelevant variance* (Messick, 1995). Construct underrepresentation

occurs when a measure does not include essential facets of the domain being assessed. Construct underrepresentation is directly related to content validity. In contrast a second issue can occur, construct-irrelevant variance, where extra pieces not related to the intended domain are being measured. This could be as simple as confusing wording causing students to pick an answer that is not representative of their underlying understanding of the concept. This could affect construct validity (ability to connect observed student responses to cognition) as well as generalizability (where certain samples of students may be more likely to make sense of a certain wording based on their instruction.) Avoiding validity pitfalls involves attending to each of the internal factors: content, construct, and generalizability.

**Concept inventory creation methodology.** The methodologies for creating concept inventories vary from instrument to instrument. Lindell, Peak, and Foster (2007) conducted a meta-analysis of 12 physics-related concept inventories (including the Force Concept Inventory) to compare their respective methodologies. They found that typically the inventories followed a nine-step sequence:

1. Identify a purpose
2. Determine the concept domain
3. Prepare test specifications
4. Construct initial pool of items
5. Have items reviewed - revise as necessary
6. Hold preliminary field testing of items - revise as necessary
7. Field test on large sample representation of the examined population
8. Determine the statistical properties of item scores – eliminate inappropriate items
9. Design and conduct reliability and validity studies (p. 15).

While these steps existed in some form across inventories, their explicitness and methods at each stage varied substantially. Lindell, Peak, and Foster noted that student conceptions (either from literature or through qualitative studies) were used to determine the concept domain in only six of the concept inventories they compared. Frequently, the researchers determined the concept domain based on their own views rather than connecting to student conceptions or other outside resources. This was the case in five of the concept inventory methodologies analyzed by Lindell, Peak, and Foster. This stage determines a great deal of the test's content validity.

Messick (1995) referred to *domain analysis*, as the essential aspect in step 2. Domain analysis involves identifying foundational concepts, essential tasks and established areas of difficulty. MisLevy, Steinberg, and Almond (2003) describe domain analysis as:

...marshalling substantive information about the domain--bringing together knowledge from any number of sources and then beginning to organize beliefs, theories, research, subject matter expertise, instructional materials, exemplars from other assessments, and so on (p. 7).

As noted by Messick, domain analysis provides the argument for task selection based on *relevance* and *representativeness*. Each task should be relevant to the domain at hand. However, it is insufficient for a measure to merely be relevant to the domain. It needs to also have a degree of representativeness. As it is impossible to cover every aspect of a given domain, tasks need to be selected to be functional - that is, reflect the types of activity actually associated with the area and to address expertise in the area. Tasks should serve to differentiate the novice and the expert. A domain analysis can be a rather daunting process depending on

how well explored a given area is. In cases like introductory group theory, without consensus on important topics (Burn, 1996; Dubinsky, Dautermann, Leron, & Zazkis, 1997; Suominen, 2014) and a relatively small education literature base, a systematic domain analysis is be a necessary first step. The domain analysis is essential for achieving content validity, but also matters in terms of other validity types. For example, the domain analysis includes determining what content students have access to. A test is not generalizable if it covers content that is unrepresented and not accessible to all of the population. This type of analysis allows for the test creator to determine what aspects of the domain should be represented. In the case of group theory, certain examples might exist in some curricula and not others. If the goal is to assess a student's conceptions around isomorphism, but one of the groups used is the dihedral group, a student's knowledge of the dihedral group likely supersedes their knowledge of isomorphism. As some textbooks do not address this example in detail, a student would not have access to this question despite the fact they might have formidable knowledge of isomorphism. This would immediately impact construct validity and generalizability.

Currently, the CCI and PCA exist in the of field mathematics education. The researchers who created the inventories modeled their development after the Force Concept Inventory to varying degrees. Generally, the test creation followed the path of a beginning pool of open-ended items that were converted to multiple-choice versions based on student responses. In both cases, interviews or "clinical labs" were used to refine questions and assure construct validity. In the CCI,



statistical measures (such as item discrimination) were used to determine if an item needed be discarded, altered, or replaced. In the PCA, items were refined based primarily on student interviews with a high emphasis placed on construct validity. In both methodologies, the final stage included estimating reliability and correlating the test to external measures.

Besides question refinement differences, the largest divergence in methodology occurred in the initial question creation. In the CCI, a panel of experts determined the content to be covered and how the content should be measured. The panel then reviewed each item. The questions in the PCA were created based on extensive studies aimed to probe student understanding around concepts that are needed in calculus. This came from both the pre-existing literature and the researchers' own studies.

The methodology for the GCI creation combined the emphasis on student conceptions and usage of experts in the field. Experts provide important insight into what is valued and representative in a field. However, this knowledge must be connected to the areas of difficulty and genuine student conceptions about topics. This methodology aims to complement the work done by Carlson, et al. (2010). Their contribution carefully laid out a methodology for beginning with a set of assessment items, a taxonomy of student conceptions, and arriving at a valid multiple-choice measure. I used a similar methodology to develop multiple-choice questions from open-ended versions. However, my methodology for developing the initial question set is centered in a careful domain analysis where experts are used in a systematic manner and paired with curriculum analysis and

literature on student conceptions to develop questions. Further, my focus is on how to better achieve generalizability and content validity with explanations of not just how questions were converted from open-ended to multiple-choice, but how and why questions were eliminated or completely replaced.

**The purpose: a discussion of conceptual understanding.** As noted by Lindell et al. (2007), the precursor to domain analysis is identifying a purpose. For the case presented here, the goal is to capture a measure of conceptual understanding in group theory. I adopt Star's (2005) approach where conceptual knowledge includes both Hiebert and Lefevre's (1986) knowledge rich in relationships and "knowledge of concepts" (p. 408). Concepts in advanced mathematics have often been discussed in terms of *objects*. Understanding of objects can arise in various ways including compressing of processes, abstracting structure from experience, and working from a formal definition (see Tall's (2004) three worlds of mathematics for a thorough discussion.) Knowledge needs to be compressed into *thinkable concepts* in order to "build a coherent connected mental structure to put significant ideas together" (Tall, 2007, p. 153).

Tall and Vinner's (1981) seminal work on concept image and definition unveiled that student understanding of concepts extends far beyond their knowledge of definition and includes "all mental pictures and associated properties and processes" (p. 152). Further, concept images do not require coherence and often only portions are evoked at a given time. Concept inventories aim to capture a students' degree of coherence. Savinainen and Viiri (2008) introduced a framework to reflect conceptual coherence as found in the Force

Concept Inventory which included three dimensions: relating different concepts, being able to apply knowledge in appropriate contexts and being able to move between representations. This framework is similar to Biehler's (2005) meanings of mathematical concepts where he identified the *domain of application*, *relations* to other concepts and *representations* available for working with the concept. These factors all serve to inform the eventual domain analysis. The rest of the paper focuses on the creation of the GCI with the underlying intention of capturing conceptual understanding.

### The Methodology

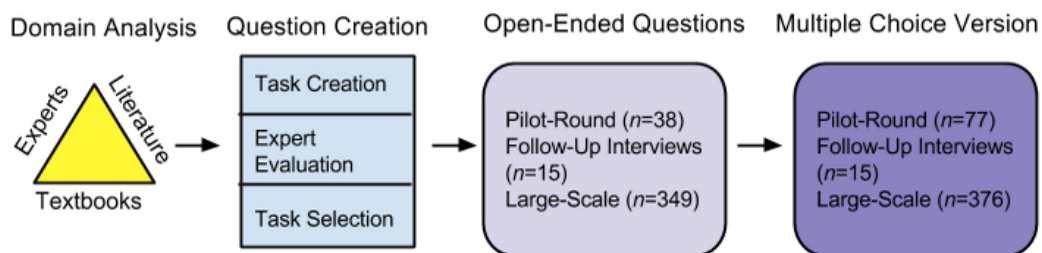
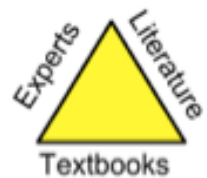


Figure 15. Overview of GCI creation methodology.

The creation and validation process for the GCI is illustrated below with focus on systematic domain analysis and the question creation and refinement process. The domain analysis included: (1) using an expert consensus protocol (Delphi Process) to arrive at a set of fundamental concepts; (2) a textbook analysis focused on conceptual aspects of the selected concepts; and (3) a literature search focused on student understanding of the selected concepts. This process then informed the task creation and selection process which included: (1) the creation of an initial set of questions related to the tasks; (2) expert evaluation of questions; and (3) question selection for piloting. The latter half of the results

focus on how open-ended questions were converted into multiple-choice questions. I used four examples to illustrate different paths that questions can take including examples of direct adaptation, minor refinements, complete replacement, and deletion. For each stage of the question-creation process, I connect the methodology to the various validity goals and discuss which aspects can be generalized to all concept inventories. In the next sections, I elaborate on each of these stages of the GCI creation. Furthermore, I connect each stage to its validity purposes. Finally, I discuss how other test creators can leverage the methodology.



*Figure 16.* Domain analysis triangle.

**Domain analysis.** Many sources can provide information about what is important, valued, and relevant to a field. I used a triangulation of experts, literature on student conceptions, and textbooks to analyze the introductory group theory domain.

**Delphi study.** Experts can be leveraged to address subjective questions such as what is important or difficult in a field. One model for arriving at consensus for an ill-defined problem is to utilize a Delphi study (Dalkey & Helmer, 1963). This technique has been used in a small number of concept inventories including the creation of the thermal and transport science concept inventory (Streveler, Olds, Miller, & Nelson, 2003). In a Delphi study, experts

provide opinions on a given question through a series of rounds where they are able to consider each other's responses after each round.

*In the GCI.* My rounds closely followed those of Streveler, et al. (2003).

The process consisted of four passes:

Pass 1: Experts were asked to compile a list of concepts they think are essential in introductory group theory.

Pass 2: A list was compiled of all concepts mentioned by at least two experts. The experts then rated each topic on a scale from 1-10 for difficulty and importance.

Pass 3: The experts were provided with the 25th, 50th, and 75th percentile scores for both categories and asked to rate, again. During this pass, the experts provided justifications for any rating outside of the 25th-75th percentile range.

Pass 4: Experts were provided with the same numerical information, as well as the justifications from pass 3 and asked to provide a final rating.

The panel for the GCI had thirteen members who participated in at least three rounds including eleven members who provided final ratings. As the domain was introductory group theory, all panel members had taught the course at least two times. The eleven panel members providing final rankings included four algebra textbook authors, eight mathematicians (with a range of specialties including math history, analysis and several group theory specialists), and five mathematics education researchers who have published related to abstract algebra pedagogy.

I selected all topics with a mean importance of nine or greater for further analysis. This narrowed the original thirty-three topics to eighteen.<sup>2</sup> Table 5 provides the selected topics and their importance mean and difficulty mean.

Table 5

<i>Important Topics in Group Theory Selected Through a Delphi Study</i>		
Topic	Importance Mean	Difficulty mean
1. First Isomorphism Theorem	9.91	8.82
1. Homomorphism	9.91	6.82
1. Associativity	9.91	2.25
1. Inverses	9.91	2.09
1. Identity	9.91	1.18
1. $\mathbb{Z}_n$	9.91	3.55
1. Quotient groups, Cosets	9.91	8.64
1. Subgroups	9.91	4.18
1. Normal Subgroups	9.91	8.09
10. Kernel	9.82	5.09
11. Abelian Groups	9.64	2.64
11. Isomorphism	9.64	6.82
13. Cyclic Groups	9.55	5.27
14. Binary Operations	9.45	2.27
15. Group	9.36	5.00
16. Lagrange's Theorem	9.18	5.64
17. Order of an element	9.09	3.55
18. Order of a group	9.00	1.64

<sup>2</sup> The eighteen topics were expanded to nineteen as cosets and quotient groups were separated to be consistent with topic treatment in curriculum. Modular groups were an example group and therefore are used within questions, but did not receive tasks independent of other concepts.

*Discussion.* Experts are a powerful resource for establishing content validity. They have knowledge about what is valued in a given field. A Delphi study is unique in its ability to allow for experts to reflect on each other's opinions while mitigating for perceived hierarchy. A round-table discussion may become biased because certain experts may have more status than others. A standard survey, while protecting anonymity, does not allow for experts to consider the opinions of other participants. The Delphi process both protects anonymity and allows for experts to consider the input of their fellow panelists.

Beyond using the Delphi structure, the success of the protocol relied on having a heterogeneous panel. By choosing a panel representing various backgrounds, their opinions were more likely to be representative and therefore strengthen content validity. If the panel consisted of only people with similar backgrounds, they may value a particular aspect of the domain more than is generally valued amongst those active in the field. In addition, the instructors had a variety of approaches to teaching the introductory group theory course. By having instructors rate and discuss topics, I could explore the differences that existed across classrooms in terms of content coverage. If I used questions on content that is not universally covered, subsections of the population would not have access to those questions and subsequently score lower on the GCI than they may have otherwise.

In other assessment development, the topics are often selected by the researchers themselves (Lindell, Peak, & Foster, 2007) or through a group of

experts (Delmas, Garfield, Ooms, & Chance, 2007; Epstein, 2007). Delphi studies provide a means for leveraging expert knowledge in a more universal manner. However, because of the nature of the survey, test developers may be limited in their ability to probe deeply into instructors' opinions leaving decisions based on numbers alone. Furthermore, the process can be time-consuming for the panelists due to the number of rounds. This raises the possibility of high attrition rates. Despite these limitations, the Delphi process remains a powerful tool for determining a consensus on content in a way that maximizes validity.

***Textbook analysis.*** Textbooks frequently serve the purpose of informing concept inventories (Lindell et al., 2007; Stone et al., 2003; Wage, Buck, Wright, & Welch, 2005). However, their use is often limited to verifying whether a set of topics or sections exists or is otherwise vaguely described. Textbook analyses can be powerful for determining how a concept is treated in the written curriculum. One model for this is Mesa's (2004) textbook analysis where she analyzed the *intended* conceptions of function. She explored function conceptions by attending to representation types, contexts and practices. Typically, textbooks are analyzed through an iterative coding process to identify the trends across books with relation to categories of interest. I introduce a systematic textbook analysis as a way to help determine the concept domain for an assessment targeted towards certain classes.

***In the GCI.*** During the creation of the GCI, I analyzed textbooks to determine typical activity and tasks, valued examples, valued representations, and formal and informal definitions of topics. I identified the most commonly used



textbooks in order to analyze the most relevant curriculum. Of the 1,244 schools with a mathematics major, a random sample was surveyed with 294 schools responding. This sample<sup>3</sup> allowed for a 95% confidence interval with error of +/- 5%. If textbook information was not available through the course website, I contacted the mathematics department or most recent instructor to find out what textbook was used in their introductory group theory course (alternately called modern algebra, algebraic structures, or abstract algebra). In schools where the textbook was not uniform, I included the textbook most recently used. I then selected any textbook used by at least 20 schools for further analysis. This number was eventually lowered to include the fourth most popular textbook. Of the institutes responding, 32% used Gallian (2009), 15% used Fraleigh (2002), 8% used Gilbert and Gilbert (2008) and 6% used Hungerford (2012). There were a total of 32 textbooks in use, but no textbook was used at more than nine institutions beyond those top four.

For each textbook, I analyzed the sections that introduced the eighteen topics identified through the Delphi study. I used a thematic analysis (Braun & Clark, 2006) where each example and exercise was coded. They were coded within the categories of: (1) relevant concept, (2) representation, and (3) example purpose or expected student activity, and (4) group/map being used. The codes initially aligned with existing frameworks, but I created and adapted codes as appropriate for this context. The representation category was adapted from

---

<sup>3</sup> As in any random sample, there are limitations due to response bias. Particularly, institutions that did not have their textbooks available via their website were likely underrepresented as not all instructors responded to email requests for current textbook.

Mesa's (2004) function representation codes with several codes remaining unaltered (symbolic, table, diagram) and many additional codes emerging related to groups. Maps were also coded similarly. Table 6 represents the stable set of representation codes after several rounds of adapting and refining. The trends for commonly used representations were overwhelming with a group being represented by a named symbol (such as  $\mathbb{Z}$  for integers) or a verbal description (the set of integers) well over half of the time. A complete breakdown of the most common group representations can be found in Table 7.

Table 6

*Representation Codes*

Group – Verbal Description	Map – Symbolic Rule
Group – Symbolic Name	Map – Defined Element-wise
Group – Table	Map – Function Diagram
Group – Elements and Operation	Map – Defined on Generating Set
Group – Set Builder Notation	Map – Verbal Description
Group – Cayley Digraph	Map – Visual Other
Group – Geometric Representation	
Group – Defined by Generating Set	

Table 7

*Percentages of Representations in Textbook Examples and Exercises*

	Gilbert & Gilbert (2008)	Hungerford (2012)	Fraleigh (2003)	Gallian (2009)
Name	69.1%	72.6%	89.3%	73.9%
Verbal Description	27.0%	14.6%	21.3%	27.4%
Table	8.2%	6.7%	7.4%	4.7%
Set of elements with operation	27.6%	19.1%	6.1%	15.8%

The example purpose codes were similarly adapted from Fukawa-Connelly and Newton's (2014) study of example purposes in a group theory course. The example purpose codes can be found in Table 8. Certain examples served a lot of purposes throughout the textbooks. The modular addition groups, of the form  $\mathbb{Z}_n$ , were one of the most leveraged types of example groups. This example either motivated or illustrated the definition of group in all four textbooks. In fact, a group of this type fulfilled an example purpose in nearly every section of every text. This was unsurprising in light of the results from the Delphi study where  $\mathbb{Z}_n$  was the *only* example group that had a mean importance score above a 9.0. Other example groups were treated differently depending on the textbook. For example, the dihedral group ranged from being the group that motivated the definition of group (Fraleigh, 2002), to being a specific example that was rarely used for any example purposes (Gilbert & Gilbert, 2008). In fact, the group was not even named consistently across texts with the majority calling  $D_4$  (or  $D_8$ ) the dihedral group of order 8, whereas Gilbert and Gilbert label this group the Octic Group.

Table 8

---

*Example Purpose Codes*

---

Example motivating a definition (EMD)  
 Example of a concept following a definition (EFD)  
 Example illustrating a specific property an object does or does not have (EP)  
 Example illustrating how to calculate or determine something (EC)  
 Example illustrating a proving technique (ET)  
 Example motivating a theorem (EMT)  
 Example illustrating a theorem (EIT)  
 Example using a theorem (EUT)  
 Example illustrating a notation (EIN)

---

In addition to representation type and topic, I coded each exercise in terms of *expected student activity* (Otten, Males, & Gilbertson, 2014). While the expected activities ranged from concept to concept, the trends were consistent. Figure 17 illustrates the percentages of the most common concept related activities across textbooks. The most frequent expected student activity fell into the super category of proving an established statement. I further subdivided proof activities into a number of categories including: (1) Proving a Definition is Satisfied (PS), (2) Proving a Direct Consequence of a Theorem (PT), (3) Proving a Direct Consequence of a Definition (PD), (4) Proof by Counterexample (PC), and (5) Proof Advanced (PA). The Proof Advanced category captured expected proof activity that relied on advanced proof techniques, nontraditional arguments or pulling together many concepts in a way that was not immediately obvious from theorems and definitions alone. While proving established statements far outnumbered evaluation (EV) activities, evaluating statements did exist across textbooks. Evaluating statements related to single concepts often had students determine if a given instance satisfied the definition of a concept. Determining if a subset forms a subgroup or determining if a group is cyclic falls into this category. Other common activity types across textbooks include *Determining or Calculating* (DC) such as finding the order of an element and changing representations such as moving from a set of elements and operation to a Cayley Table.

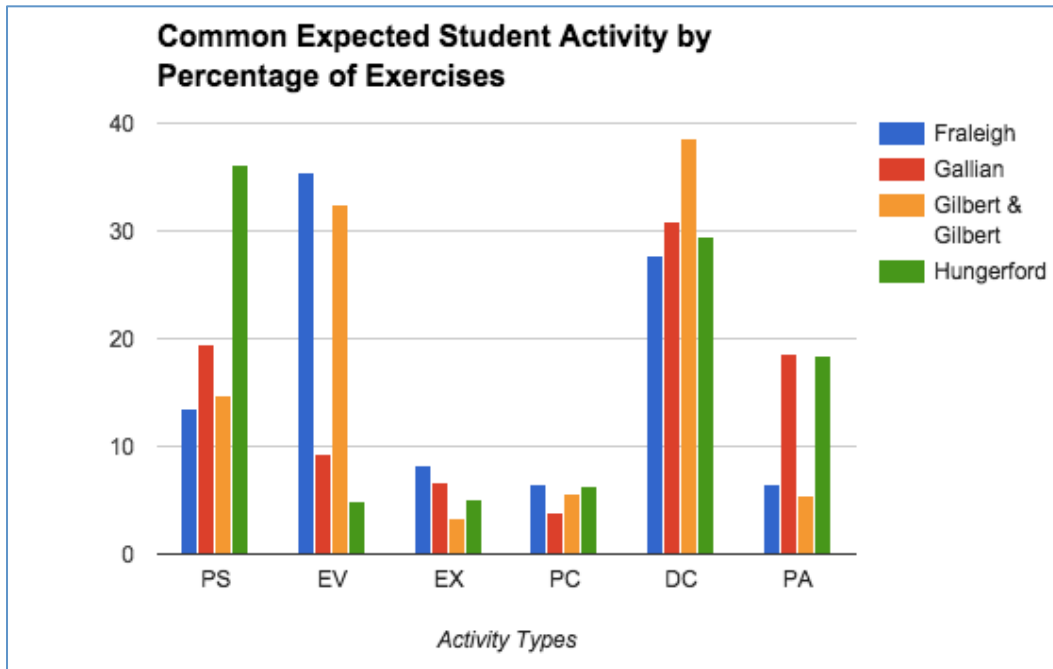


Figure 17. Percentages of Exercises with each type of the six most prevalent expected student activities.

For reliability purposes, a mathematics education graduate student coded one section (on a different topic) from each book. They coded both exercises and examples. The inter-rater reliability was 83.4% for exercises and 81.4% for examples. These numbers represents the total number of codes identified by the second coder and first coder over the number of codes.

In addition to coding exercises and examples, I analyzed the formal and informal definitions for each concept. Generally, textbooks did not provide informal definitions to accompany formal definitions beyond what was labeled a *translation*. I defined a translation as an unpacking of a definition into less formal language. In contrast, an informal definition was a description that provided some intuitive explanation of the concept. Isomorphism was atypical in having informal

definitions in all texts. Table 9 presents the formal, informal, and translation definition for isomorphism.

While informal definitions were scarce and inconsistent, formal definitions, while also presented, were also occasionally inconsistent across texts. For example, the order of an element was defined in one of two ways:

- (1) **(Gilbert & Gilbert, 2008; Fraleigh, 2003)** “The order  $o(a)$  of an element  $a$  of the group  $G$  is the order of the subgroup generated by  $a$ . That is,  $o(a) = o(\langle a \rangle)$ ” (Gilbert & Gilbert, 2008, p. 167).
- (2) **(Gallian, 2009; Hungerford, 2012)** “The order of an element  $g$  in a group  $G$  is the smallest positive integer  $n$  such that  $g^n = e$ . If no such integer exists, we say that  $g$  has infinite order. The order of an element  $g$  is denoted by  $|g|$ ” (Gallian, 2009, p. 60).

Differences in formal definitions also existed for modular arithmetic and normal groups. Generally, the alternate definition was a theorem or otherwise listed as a consequence of the definition. (See Paper 2 for a discussion of the differing modular arithmetic definitions.)

For each topic, I compiled an analytic report that summarized definitions, representation types, examples and their purposes and expected student activity.

*Discussion.* The textbook analysis provided a way to explore intended conceptions around topics. The analysis served the dual purpose of exploring what was valued in the field and what students have access to. The examples found in books represent the *conventional example space*, the set of examples generally understood by the mathematics community (Zazkis & Leikin, 2008). The conventional example space is an important aspect of the concept domain highlighting what the community values. The conventional example space also

gives insight into what examples may be part of a students' personal example space. A personal example space is "the set of mathematical objects and construction techniques that a learner has access to as examples of a concept while working on a given task" (p. 291). Students need exposure to example groups before the groups can enter their personal example space. If tasks contain examples that the students do not have access to, then they will not be able to work in the task context.

Table 9

*Types of Definitions in Textbooks*

---

Formal	An isomorphism $\phi$ from a group $G$ to a group $G^-$ is a one-to-one mapping (or function) from $G$ onto $G^-$ that preserves the group operation. That is, $\phi(ab) = \phi(a)\phi(b)$ for all $a, b$ in $G$ . If there is an isomorphism from $G$ onto $G^-$ , we say that $G$ and $G^-$ are isomorphic and write $G \approx G^-$ (Gallian, 2009, p. 123).
Informal	At first glance, the groups don't seem the same. But we claim they are "essentially the same" except for labels on the elements (Hungerford, 2012, p. 214).
Translation	Every element of $G$ is paired with a unique element in $H$ (its new label.) In other words, there is a function $f: G \rightarrow H$ that assigns to each $r \in G$ its new label $f(r)$ in $H$ (Hungerford, 2012, p. 215).  An isomorphism is said to "preserve the operation," since condition 2 of Definition 3.25 requires that the result be the same whether the group operation is performed before or after the mapping." (Gilbert & Gilbert, 2008, p.177).

---

Similarly, the types of representations found across texts served as both what representations students should understand and what representations students have a fair opportunity to engage with. Students may not have been exposed to all representations, and so textbooks provide a way to determine what

representations are likely to be used frequently and consistently across different classes. Further, the frequencies of usage may reflect the level of value that each representation has within the field.

The differing definitions also have a huge impact on meaningfulness of questions. It is likely that students with differing definitions may interact differently with tasks. If an assessment was catered towards one definition over another, two differing classes could not be fairly compared.

The expected student activity analysis provided a detailed look into what type of applications existed related to the various important topics. These activities ranged in nature with procedural, conceptual and proof-based activity. The activities served less to address purpose, and more towards content validity-establishing what is valued in the domain.

By analyzing textbooks, I was able to explore both content and access issues related to the domain. Any assessment should aim to reach the goals of being valid for the entire population and be representative of what is actually valued in the domain. As no assessment can cover all aspects of a targeted domain, a textbook analysis can provide insight into what is most valued.

Textbook analyses do have some limitations. Textbooks represent only the *intended curriculum* (Travers & Westbury, 1989). They do not provide any evidence of what is actually occurring in classrooms – the enacted curriculum. Several institutions do not make use of textbooks at all. However, as noted by Zhu and Fan (2006):



...textbooks are a key component of the intended curriculum, they also, to a certain degree, reflect the educational philosophy and pedagogical values of the textbook developers and the decision makers of textbook selection, and have substantial influence on teachers' teaching and students' learning (p. 610).

Textbooks serve both as a proxy for what occurs in the classroom and as an artifact of what is valued.

**Literature review.** The last leg of the domain analysis triangle was a literature review. Literature on student conceptions underlies the formation of many concept inventories (Lindell, et al., 2007; Hestenes, et al., 1992; Carlson, et al., 2010). I contribute a methodology for investigating literature systematically.

*In the GCI.* Because the body of literature on student understanding in abstract algebra is relatively small, this could be done systematically and comprehensively. Literature was identified using the following steps:

1. A full text search of relevant journals
2. Title search of conference proceedings
3. Broad subject search in relevant database

For the case of the GCI, step one was accomplished by doing a full text search for the concepts selected from the Delphi study. For vague terms such as "group" the phrase "abstract algebra" was added. I searched each of the 35 journals identified by Special Interest Group of the Mathematical Association of America on Research in Undergraduate Mathematics Education as being research-based and relevant to undergraduate mathematics. Further, conferences that are research-based and attend to undergraduate mathematics education were explored including: *International Group for the Psychology of Mathematics Education*, *Research in Undergraduate Mathematics Education*, *Delta Conference on*

*Teaching and Learning of Undergraduate Mathematics and Statistics, Congress of European Research in Mathematics Education and International Congress on Mathematical Education.* Finally, I searched the terms “Group Theory” and “Abstract Algebra” in the Education Resources Information Center (ERIC) database to assure no articles were overlooked in the earlier process.

For each concept, I compiled a set of analytic notes that included any conceptions found in the literature and corresponding tasks found in publications. Additionally, I addressed overarching themes about how students understand concepts in group theory.

*Summary of literature on student conceptions.* Much of the work related to student understanding in group theory was done in relation to the *Action, Process, Object, and Schema* (APOS) framework (Dubinsky & McDonald, 2002). The framework serves to deconstruct mathematical topics in terms of the action, process, object, and schema conceptions. An action is “repeatable physical or mental manipulation that transforms objects in some way” (Dubinsky, Dautermann, Leron, & Zazkis, 1994, p. 270). Whereas an action relies on running discretely through each step, the action can be *interiorized* into a *process* where the actions are thought of as a whole rather than a series of steps. This process can then be *encapsulated* such that it becomes an object that can be transformed in some way such as coordinating with other processes. Objects and processes can be coordinated, *thematized*, to become a schema. A schema consists of a set of related objects and processes that compose a concept. A well-developed understanding of a complex concept such as a group requires a schema conception

where the various associated objects and processes can be unpacked in relevant situations. Abstract algebra concepts decomposed in terms of this framework include: binary operation, groups, subgroups, cosets, quotient groups and normality (Asiala, Dubinsky, Matthews, Morics, & Oktac, 1997; Brown, Devries, Dubinsky, & Thomas, 1997; Dubinsky et al., 1994). For example, Asiala et al. presented a genetic decomposition (set of mental constructions) of group. The group schema consists of coordinating three schemas: set, binary operation, and axioms. The axiom schema consists of the process of checking the axioms and four objects: the axioms themselves. To check axioms, the objects of the axioms must be de-encapsulated to processes to be utilized with the binary operation and set at hand. The four axiom processes must then be coordinated with the requirement that they are all satisfied in order for the set and operation to be a group. Groups can then be treated as an object to apply processes or actions such as determining if two groups are isomorphic or building the direct product of two groups. While experts might be expected to have well-developed schema regarding different concepts, Asiala et al. found that undergraduate students frequently struggle to coordinate and develop all the necessary mental constructions.

Students struggle with topics in abstract algebra for a number of reasons including their complexity, abstraction, and formality. The complexity of the group schema can be seen above with the number of aspects that need to be coordinated. In Leron, Hazzan, & Zazkis' (1995) exploration of isomorphism, they found that isomorphism contained many complexities leading to student

struggle. Understanding isomorphism involves leveraging both the group and function schemas. Further, isomorphism involves the map itself, the isomorphism, as well as the property groups can share of being isomorphic. There are also notable differences between showing two groups are isomorphic and showing they are not isomorphic and making sense of the general idea (that groups are the same) and how that comes to fruition in specific cases. The schemas and relationships amongst concepts are often quite complex in this setting. This is particularly problematic when paired with new levels of abstraction and rigor.

Hazzan (1999) discussed several issues surrounding abstraction level. These played out in one of three ways. First, the process-object duality can be difficult to manage in a meaningful way. The researchers found students would often rely on the calculation or process for finding cosets rather than treating cosets as objects themselves. Second, the unfamiliarity of objects can lead to inappropriately borrowing from known objects. They present an example of a student determining that  $\mathbb{Z}_3$  was not a group because the element 2 had no inverse. The student had mistakenly looked for  $1/2$ , the inverse of 2, in the more familiar group: reals under multiplication. Additionally, they presented complexity reduction where, for example, a student may use a specific case in place of a general one.

Connected to abstraction is the baggage that comes with the formal definition of concepts. Rather than describing existing objects, a definition stipulates an object. This means that quantifiers take on an enhanced meaning. For example, when providing an analysis of binary operations Novotná, Stehlíková,

and Hoch (2006) differentiated between levels of understanding where a student could correctly address the quantifier involved in the identity element definition. An element  $e$  is an identity for a given operation in a set if it fulfills the property  $ex=xe=x$  for *all*  $x$  in the set. An identity element cannot be different for differing elements of the set. Furthermore, the differences between intuitive understanding and formal definitions may cause issues in developing correct conceptions around topics. Lajoie and Mura (2000) found that many students utilized an intuitive definition of cyclic groups (repeatedly operating an element with itself) rather than the formal definition ( $\{x^n \mid n \in \mathbb{Z}\}$ ) to arrive at the conclusion that  $\mathbb{Z}$  is not cyclic. In formal mathematics courses, the differences between formality and intuition, as well as the dependence on quantifiers, provide a high level of difficulty associated with concepts.

*Discussion.* Using literature is an essential component of developing construct validity. If we wish to interpret students' cognition, research on their cognition should be leveraged. One area of domain analysis is determining which aspects of the domain are associated with potential difficulties. The various theories of cognition provide important information about students' cognition and the types of questions that may target important aspects of understanding. The foci of these studies also reflect what is valued in the field.

The biggest issue that can be caused by the usage of literature is failing to triangulate with other sources. This is especially the case in group theory where the literature base is not particularly extensive. Several of the topics identified as fundamental from the Delphi study had little to no associated literature on student

conceptions. Furthermore, most studies are dependent on local samples and consequently the curriculum associated with their group theory course. For example, many tasks for studies use examples that are not universally treated in textbooks. Blindly leveraging these studies may prevent many students from having access to questions. Literature should serve as a guiding tool, but not a limiting one.

**Conclusion.** The domain analysis unearthed information about what is valued in the field in terms of both topics and surrounding examples, representations, and activities. Additionally, the analysis provided information about known difficulties in the domain, previously studied tasks, and theories of student cognition. The results of the process serve to inform the question creation and selection through a list of starting tasks, known conceptions to target, and through leveraging what is representative and to what representations, examples, and definitions students likely have access.

### Question Creation and Selection

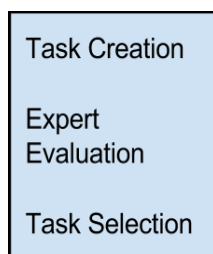


Figure 18. Task creation and selection.

**Open-ended question creation.** I created or adapted 42 open-ended questions for evaluation based on the domain analysis. In this section, I explain how the domain analysis informed this process.

*Topics.* The topics identified in the Delphi study provided the starting ground for task creation. All of the topics identified by the panel existed within the introductory chapters of all the textbooks, so each topic was used in the question-creation process. Every open-ended question was created to directly tie to one of the eighteen topics. For each topic between one and three questions were created depending on the amount of literature provided and complexity of the topic.

*Representations.* The analysis of representations in textbooks was leveraged in the question-creation process. Only representations found across textbooks and sections were used. Because verbal descriptions and symbolic names were dominant in all curricula, these were the primary representations used for the GCI. In the GCI, all symbolic names were accompanied by verbal descriptions to mitigate for any unfamiliarity and make the test more accessible to a general population. The table and a list of elements and operations were also a representation type found consistently through textbooks. Several GCI questions made use of a list of elements (particularly for subgroups which often lack either a symbolic or verbal description that would be universal) and one question uses the table representation for a group. If students are confronted with unfamiliar representations, they will not have access to the question and it will no longer be meaningful. However, representations are also an important aspect of concept understanding, so understanding which are valued was essential to determining which representations were used to use in questions.

*Examples.* Many GCI questions about groups involve reasoning in specific contexts about specific example groups. Both the Delphi study and textbook analysis informed the selection of examples for use in item design. During the Delphi study several example groups did not meet the threshold score for importance. The textbook analysis confirmed that these examples were inconsistently treated across textbooks. I did not include any unfamiliar example groups besides the modular addition group because of the low importance means and the treatment in textbooks.

*Activity.* The task activities were informed by both the literature and textbook analysis. The evaluation type of activity was chosen as one of the primary question types for the GCI. This was done for several validity reasons. First, evaluate-type questions did exist in all curricula indicating some degree of value. Second, these types of questions are frequently used in studies that explore student conceptions and have been shown to illustrate a multitude of conceptions that vary from novice to expert (*see* Dubinsky et al., 1994; Leron, et al., 1995; Weber & Alcock, 2004). Further, evaluation questions are not prompts for proofs. The goal of the GCI was not to assess formal proving ability but understanding of concepts. From a utility standpoint, evaluate-type questions allowed for more conceptions to be probed as both “yes” and “no” responses unearthed differing reasons.

Other activity types that were used for question development include the usage of a definition or usage of a theorem. Recognizing when a concept or theorem can be used is an essential part of understanding the concept. This falls in



the applying knowledge in appropriate contexts (Savinainen & Viiri, 2008) dimension of conceptual understanding. For example, suppose students are presented with a situation requiring conclusions about the image of a homomorphism. If someone has a strong understanding of the Fundamental Homomorphism Theorem, they would be able to recognize this context as related and apply the theorem. Understanding a concept is not limited to making sense of its definition but also involves being able to use the concept in conjunction with various situations.

*Targeting conceptions.* The tasks on the GCI were informed by the literature in one of two ways. I directly adapted tasks when available and consistent with examples and representations from the domain analysis. Otherwise, I created new tasks aimed at capturing student conceptions from the literature. A number of tasks came from the literature unaltered or slightly adapted such as the tasks in Paper 2. However, often tasks found in literature did not make good multiple-choice questions because they were either too complex or not accessible in terms of example groups or representations. Some papers contained relatively targeted misconceptions such as students failing to address both requirements for the definition of order of an element. Anderson et al. (1998) found that many mathematics majors defined order of an element,  $x$ , as an integer  $n$  such that  $x^n=e$  or 1 without attending to  $n$  being the smallest integer such that this equation holds. I developed a specific open-ended question aimed at this conception: “If  $a^6=e$ , what are the possible orders for  $a$ ?” This allowed for multiple responses that reflected student conceptions around the idea without the

multiple-choice options giving away the answer.

Other literature around student conceptions provided more complex theories and frameworks that served to inform task creation. For example, quotient groups and cosets provided a rich ground for exploring student conceptions. Asiala, et al. (1997) provided a genetic decomposition of coset and quotient groups and the necessary mental constructions for a well-developed quotient group schema. This schema involves coordinating the coset schema, the binary operation schema, and the group schema. Students must be able to conceive of cosets as objects in order to form a set of sets. Operating on the sets requires the usage of representative elements or coset multiplication. The process of forming a coset must be deencapsulated to robustly use representatives (verify that operating on any elements from respective sets produces the same result) or coset multiplication (where all combinations of elements must be multiplied). Siebert and Williams (2003) explored some of this complexity in terms of modular addition groups. They found students had three conceptions around the cosets in this quotient group. The cosets could be viewed as infinite sets, they could be viewed as a representative element and a set, or just as the representative element. Quotient group tasks were created to address the complexity of dealing with elements in quotient groups as both elements (with representative structure) and sets. The trajectory of the question targeting quotient groups can be found in the Question Refinement section.

Additionally, not all literature was specific to a certain concept, but rather presented theories that transcend given topics. For example, in Hazzan's (1999)

theory about reducing abstraction, she explained that students often reverted to familiar groups when reasoning about unfamiliar ones. This idea was incorporated into the concept inventory in a number of tasks where unfamiliar binary operations might be reasoned about as if they acted as familiar operations. For example, students may inappropriately borrow a known identity, such as zero, to consider the identity related to an unfamiliar operation such as the one discussed in the Minor Refinements section.

*The creation.* Underlying the creation of questions were the three components of internal validity: content, construct, and generalizability. The first set of open-ended questions were designed to (1) be accessible to a general introductory group theory student (based on representation and example type), (2) be related to an essential concept independent of proof, and (3) have the potential to connect to student conceptions in a multiple-choice format.

***Expert evaluation.*** Experts are often used to evaluate the questions for concept inventories. As instructors, they have insight into whether tasks are relevant or important, and what student conceptions they may target.

*In the GCI.* Each of the 42 open-ended questions were evaluated by at least two mathematicians with experience teaching the course and two mathematics educators who have published articles related to the teaching and learning of abstract algebra. The evaluators were prompted to address:

1. Is this task relevant to an introductory group theory course?
2. Does this task represent an important aspect of understanding [relevant concept]?

*If you said NO to either of the above, indicate why. (For example, the task might be too advanced, capture prerequisite knowledge, or be unrelated to the targeted concept.)*

3. What student conceptions might this task capture? (You could speak generally, or provide samples of how you think students might respond.)

The feedback from expert evaluations was analyzed using several dimensions:

- 1) Minor refinement and wording alterations.
- 2) Concerns about potential to become a multiple-choice question.
- 3) Concerns about relevance/importance.
- 4) Concerns about difficulty level.

*Discussion.* The expert evaluations served two important purposes: determining which open-ended questions to move into the next round and evaluating the content validity of the questions. Further, as experts in mathematics, they addressed issues with wording and mathematical language that could lead to non-construct related variance. As in any usage of experts, there are limitations because of subjectivity. This could be mitigated with a larger group of evaluators.

***Question selection.*** From the pool of open-ended questions, I selected a subset of 18 questions to pilot. These questions were selected to maximize validity.

*In the GCI.* Within each topic, a single task was selected that met the criteria of being relevant and important by all evaluators when possible. Of tasks that met the criteria, any that were previously studied and found in the literature were selected first. If no tasks existed from the literature, the task was selected where experts felt the student conceptions were most aligned with the underlying

concepts and did not contain construct-irrelevant difficulty. At this point, the questions selected were altered to address the feedback from the evaluators when necessary. Most alterations were minor wording changes at this stage, although one question was replaced based on a substantial suggestion from a reviewer. In this case, the reviewer-suggested question both aligned with tasks in the curriculum and was related to a concept (kernel) that has no existing literature about student conceptions.

*Discussion.* Any time questions are selected, there is a tension between depth, coverage and time constraints. For the GCI, I chose only one open-ended question per topic. The survey of 18 questions took roughly an hour for students to complete. It was not feasible to include more questions. The 18 questions were all evaluated as important and relevant to introductory group theory and the targeted concept. However, with only one question per topic, all topics are treated equally in this question set. Evaluators could consider whether all important aspects of concepts were considered and new questions may need to be developed to more appropriately address some of the more complex topics such as isomorphism.

### Question Refinement and Field Testing

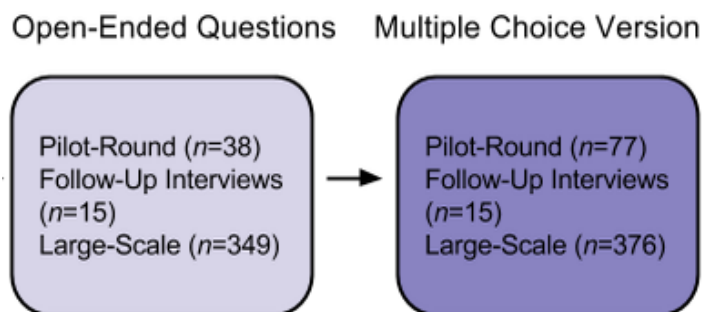


Figure 19. Open-ended (left) and multiple-choice (right) field testing.

For the open-ended round, students from 29 institutions representing five geographic regions of the United States (West, Midwest, Southeast, Northeast, and Mid-Atlantic) and varying levels of exclusivity were selected. Selectivity was based on acceptance percentage within categories of high (less than 25% accepted), mid-high (25-50% accepted), mid-low (50-75% accepted), and low (75% or greater accepted). For the closed-form version, I made an attempt to contact all instructors of abstract algebra courses in the United States during spring 2015. Any institution offering a mathematics major was part of the targeted population. Online schedules were used when available. If not, a member of the research team called the institution to find contact information of the current instructor. Institutions where group theory was addressed early in the term (quarter systems; group-first approach in semesters) were part of the pilot. Group theory classes that finished the material later in the term were part of the large-scale round. The sample represented all geographic regions and all levels of selectivity (see Table 10 for breakdown of selectivity at each round.)

The questions went through several stages of refinement during this time:

1. Open-Ended Round
  - a. Pilot ( $n=38$ )
  - b. Follow-up Interviews ( $n=15$ )
  - c. Large-Scale Field Testing ( $n=349$ )
2. Multiple-Choice Round
  - a. Pilot ( $n=77$ )
  - b. Follow-up Interviews ( $n=15$ )
  - c. Large-Scale Field Testing ( $n=376$ )

Table 10

<i>Selectivity of Sample Institutions</i>					
	Least Selective (>75% admitted)	Mid-Level Selective (50-75% admitted)	More Selective (25-50% admitted)	Most Selective (<25% admitted)	Not classified
Open- Ended Pilot	3 classes (23 students)	1 class (2 students)	1 class (13 students)	0 classes	0 classes
Open- Ended Full	13 classes (138 students)	12 classes (108 students)	4 classes (47 students)	1 class (57 students)	0 classes
Multiple Choice Pilot	2 classes (17 students)	0 classes	3 classes (26 students)	4 classes (44 students)	0 classes
Multiple Choice Full	13 classes (131 students)	10 classes (128 students)	6 classes (84 students)	1 class (14 students)	2 classes (19 students)
Total	28 classes (286 students)	22 classes (236 students)	13 classes (157 students)	6 classes (119 students)	2 classes (19 students)

This process was adapted from Carlson, et al. (2010)'s methodology for the creation of the PCA. All questions were field tested as open-ended questions. The most common student responses became the options for the closed-form multiple-choice questions. I conducted follow-up interviews to validate the questions. These interviews asked students to explain their answer choices and connect them to their understanding of the underlying concept. After both the open-ended and multiple-choice rounds, I conducted 15 interviews follow-up interviews respectively. Questions were refined at each stage of the instrument development to increase internal validity of the instrument. This took a number of

forms, ranging from direct adaptation to complete question deletion. Each of these paths is illustrated below with open-ended responses and interview excerpts as relevant.

Table 11.

*Open-Ended Responses for Cyclic Question*

Response Category	Sample Student Response
$\mathbb{Z}$ is cyclic because it can be generated by one element	Yes. To be cyclic means that a single element of $\mathbb{Z}$ can produce the entire set, which is $\mathbb{Z}$ under the following condition: $\langle a \rangle = a^i$ for some integer $i \in \mathbb{Z}$ . Letting $a = 1$ will produce all of $\mathbb{Z}$ .
$\mathbb{Z}$ is cyclic because it can be generated by a finite set of elements	Yes, we can start w 1 & -1 and generate all of $\mathbb{Z}$
$\mathbb{Z}$ is not cyclic because the elements do not cycle	No. If you take any number and keep applying it to itself (addition) you will never get back to that number while having gotten every other number since $\mathbb{Z}$ is infinite.
$\mathbb{Z}$ is not cyclic because no element will generate the whole set	No. Cyclic means generated by one element. If you start with a negative element, it can only generate a set of negatives. If you start with a positive element, it can only generate a positive set. So no element can generate $\mathbb{Z}$ .

**Direct adaptation.** The ideal situation occurred when an open-ended question remained unchanged and the student responses were easily categorized into multiple-choice options. The question about cyclic groups was adopted directly from the literature: “Is  $\mathbb{Z}$  a cyclic group?” (Lajoie & Mura, 2000). In the open-ended version, students were prompted to include their definition of cyclic



group to assure they shared some thinking beyond “Yes” and “No.” Each response was open-coded with every attempt to not be influenced by any pre-existing ideas of student conceptions. This decision was made to avoid any distortion of the data to fit expected responses. This coding process led to the following categories of responses found in Table 11.

As the responses categorized nicely, they became the options for the multiple-choice version. During field-testing the multiple choice version also contained the option “None of the above” to allow for relevant conceptions not uncovered in the first round. This was done to increase construct validity of the finalized version. The open-ended responses, combined with student interviews (following a think-aloud approach) and lack of students choosing “none of the above” served to provide a strong argument for construct validity of the question. A more detailed discussion of the question can be found in Paper 2.

7. Is  $\mathbb{Z}$ , the group of integers under addition, a cyclic group?
- a. No, because  $\mathbb{Z}$  is infinite and elements do not cycle.
  - b. No, because any element only generates part of the set (ex: 1 would only generate the positive integers.)
  - c. Yes, because  $\mathbb{Z}$  can be generated by a set of two elements.
  - d. Yes, because  $\mathbb{Z}$  can be generated by one element.

*Figure 20.* Multiple-choice version of cyclic group question.

**Minor refinements.** More common than directly adapting were minor refinements based on student responses. This was largely done to address both construct and content validity issues. The questions needed to be related to the concept and varying student responses should be tied to the intended content. Further, the responses should be able to differentiate different student conception.

This version can be seen with the adaptation of the identity question. The original question was adapted from Novotna, Stehlikova and Hoch (2006).

<p>Given <math>L</math> is the set of all positive rational numbers, consider the binary operation <math>*</math> defined:</p> $x * y = \frac{x}{2} + \frac{y}{2} + xy.$ <p>Determine if this operation has an identity. If so, identify the identity.</p>
--

*Figure 21.* Open-ended version of the identity question.

This question aims to probe whether students are conceiving of quantifiers correctly when addressing the identity element. If one were to solve  $x * e = x$ , they would arrive at an  $e$  that is dependent on the element  $x$ . This means that the “identity” would vary for each  $x$ . As an identity element must satisfy  $x * e = x$  for all elements, this binary operation would not have an identity over this set. Open-ended responses generally fell into the categories found in Table 12.

Two of these responses connect to the discussion from the literature that students may reduce abstraction level by inappropriately borrowing from known groups (or operations in this case.) Many students relied on the assumption that zero or one would be the identity or the only identity candidates. Zero and one are the identities for the familiar operations of addition and multiplication on the rational numbers. However, there is no reason the identity of an operation would need to be either. Consider the operation:

$$x * y = x + y + 1$$

The identity element would be negative one because:

$$a * (-1) = a + (-1) + 1 = a \text{ and } (-1) * a = (-1) + a + 1 = a.$$

Table 12

Open-Ended responses for identity question

Response Category	Sample Student Response
Identifying familiar identity zero or one	<i>the identity would be 0</i>
Tested familiar identities and rule them out	<p><u>No</u></p> <p>not 0 since <math>0 * y = \frac{y}{2}</math></p> <p>Not 1 since <math>1 * y = \frac{1}{2} + \frac{y}{2} + y</math></p> <p>No other possibilities</p> $\frac{1}{4} + \frac{1}{4} + \frac{1}{4} \left( \frac{1}{4} \right)$ $\frac{1}{8} + \frac{1}{8} + \frac{1}{16} = \frac{5}{16}$
Found identity dependent on element chosen	<p>what <math>x * e = x</math></p> $x * e = \frac{x}{2} + \frac{e}{2} + x * e = x$ $\frac{x}{2} + -x = -\frac{e}{2} - x * e$ $-\frac{1}{2}x = e(-\frac{1}{2} - x)$ $e = \frac{-\frac{1}{2}x}{-\frac{1}{2} - x}$ $e = \frac{\frac{1}{2}x}{\frac{1}{2} + x}$ <p><math>e = \frac{x}{1+2x}</math> ← positive rational number</p>
Identified that the identity cannot depend on element chosen	<p>5. Must a group of order 12 have a subgroup of order 6? Why or why not?</p> $x * e = \frac{x}{2} + \frac{e}{2} + x * e = x$ $e(x + \frac{1}{2}) = \frac{x}{2}$ $e = \frac{x}{2(x + \frac{1}{2})}$ <p>← depends on the value of <math>x</math> ⇒ no universal identity.</p>

The open-ended version successfully unearthed a number of issues with the identity element: both the issues surrounding quantifiers as well as issues around unfamiliar operations. At this point, one alteration was made before writing the closed-form version. Several students were concerned that domain was the reason that zero could not be the identity (see Figure 22). This was not meant to be probed by the question and could lead to misinterpretation of student

conceptions in the closed-form version. For construct validity purposes, the domain was altered to avoid this issue.

*No identity. 0 would be the identity. However, 0 is not a positive rational number.*

Figure 22. Sample student work illustrating the problematic domain.

Given  $L$  is the set of non-negative rational numbers, consider the binary operation  $*$  defined:

$$x * y = \frac{x}{2} + \frac{y}{2} + xy.$$

Determine if this operation has an identity in  $L$ .

- a. Yes, the identity would be zero.
- b. Yes, the identity would be  $-\frac{x}{1+2x}$ .
- c. No, as only 0 and 1 would be the candidates.
- d. No, the identity expression would depend on the element chosen.
- e. None of the above reasoning is valid.
- f. I don't know.

Figure 23. Multiple-choice question on identity.

Unlike in the first question discussed, “None of the above” was selected relatively frequently during the pilot highlighting that some conception was not being captured. Follow-up interviews and open-ended responses illustrated that this selection was not necessarily because of any issues with identity, but rather the complexity of the algebra involved. During an interview, one student explained that zero does not work because  $x$  times 0 would give you  $y/2$  and not  $x$ . She explained for an element to be the identity, “ $x$  times the identity is  $x$ .” She went on to eliminate the second option because, “You can try it with  $x$  over  $1+2x$

and that – yeah, that doesn’t work.” Later she elaborated, “Something like the second option is valid and sensible even though it has  $x$  in it.” She did not take issue with identity relying on an element, but rather made a computational error.

The content of the question was altered to eliminate this unintended difficulty. Ability to isolate a variable was not part of the intended content domain, causing issues with content validity and construct validity. (Students have the targeted conception but are not selecting that answer for a different reason.) The equation was simplified to reduce the computational nature of the question.

3. Given  $L$  is the set of non-negative rational numbers, consider the binary operation  $*$  defined:

$$x * y = 2x + 2y.$$

Determine if this operation has an identity in  $L$ .

- a. Yes, the identity would be zero.
- b. Yes, the identity would be  $-\frac{x}{2}$ .
- c. No, as only 0 and 1 would be the candidates.
- d. No, the identity expression would depend on the element chosen.
- e. None of the above reasoning is valid.

Figure 24. Revised multiple-choice question on identity.

Because “None of the above” was frequently selected in the pilot version, the option remained in the full-scale version, unlike the cyclic question where it was not needed. The alteration to the binary operation caused an additional issue with the set leaving “None of the above” as a popular selection. Because the potential identity element was  $-x/2$ , students were concerned that this element would not be in the domain of non-negative numbers. If any positive number were substituted for  $x$ , this expression would be negative. The finalized version of the

question altered the domain to rational numbers to prevent this issue from interfering with the intended conceptions being targeted.

**Replacement.** While most questions were either directly adapted or altered in minor ways through the process (16 of 18 questions), two had more substantial changes necessary. In one case, a question was completely replaced between the open-ended pilot and the eventual open-ended large-scale round. This was done for construct validity reasons. The open-ended pilot question was meant to target conceptions around quotient groups with the expectation that students may not have complete conceptions regarding the elements and operation in quotient groups.

Create an operation table for a two element quotient (factor) group formed from the group in the following table:

	e	a	b	c
e	e	a	b	c
a	a	e	c	b
b	b	c	e	a
c	c	b	a	e

Figure 25. Original quotient group question.

The quotient group would be isomorphic to  $\mathbb{Z}_2$ . However, this can be arrived at with no understanding of quotient groups as illustrated by the following student comment, “I didn’t really know what it was asking for because I didn’t know what this meant, but I found these two things – the two-element groups.”

	$e'$	$r$
$e'$	$e'$	$r$
$r$	$r$	$e'$

Figure 26. Student with correct conception of elements in quotient groups.

$e$	$a$
$a$	$e$

Figure 27. Student with incorrect conception of quotient groups.

This student found a two-element subgroup and presented it as the quotient group. This was entirely disconnected from their understanding of quotient group. As illustrated in Figure 26 and Figure 27, students were arriving at a similar (isomorphic) copy of the quotient group with a range from valid understanding of quotient groups to simply giving a subgroup with no connection to the topic. The quotient group created using the subgroup  $\{e, b\}$  and displayed using representative elements would be identical to the group provided by the student on the right. The student on the left explained  $e$  represented  $\{e, b\}$  and  $r$  was  $\{c, a\}$ . They were then able to operate the cosets to arrive at the table presented.

Additionally, because the quotient group would contain two elements, and each coset would contain two elements, a conflation between coset size and group order would not be unveiled with this question. In order to alter the question sufficiently, a larger group would be needed. The question was initially simplified in response due to evaluator concerns that the question was too time intensive if the student must check all subgroups for normality. Additionally, the question becomes tied to a procedure for finding a quotient group, which may mask a conceptual understanding. There are a number of ways the question could be

altered, such as changing the group from  $\mathbb{Z}_2 \times \mathbb{Z}_2$  to  $\mathbb{Z}_6$ , which would mean that the quotient group could not be made identical to a subgroup. Because of the potential to complicate the question, the question was replaced entirely.

The new version attempted to target the same student conceptions surrounding quotient groups in a different way where each response would be meaningful.

Let  $G = \mathbb{Z}_{12}$  (under addition) with subgroup  $H = \{[0], [4], [8]\}$ . Consider the quotient (factor) group  $\frac{G}{H}$ . What is the order of  $[2] + H$  in  $\frac{G}{H}$ ?  
 (Note: Depending on your textbook/instructor these elements might have different notations including:  $H = \{[0]_{12}, [4]_{12}, [8]_{12}\}$ ,  $\{\bar{0}, \bar{4}, \bar{8}\}$  or  $\{0, 4, 8\}$ .)

Figure 28. New open-ended quotient group question.

In this case, a student would need to attend to a particular coset as an element in the quotient group to find its order rather than only attending to the cardinality of the set. Further, the question aimed to pull out a third conception where only the representative element is used. If a student only attends to the 2, the representative and not the coset would determine then the order of the element. The new question tested well in the large-scale open-ended round. An exemplar from each case can be seen in Table 13. Based on open-ended surveys and interview data, four distinct conceptions existed for each numerical response:

- a. 2: A student is able to conceive of the elements in  $G/H$  as sets and elements.
- b. 3: A student only attends to the coset as a set, finding its cardinality.
- c. 4: A student finds the order of the group superficially using a memorized fact.
- d. 6: A student finds the order of the representative element in  $G$ , rather than consider  $2+H$ .

The replacement question was able to delineate between various student conceptions establishing construct validity and meaningfulness at the item level.



Further, each conception was connected to the targeted area and reflected understanding related to the domain of introductory group theory and the specific topic of quotient groups.

Table 13

*Open-Ended Responses for the Revised Quotient Group Question.*

Response Category	Sample Student Response
2	<p>Repeat adding 2 until an element of <math>H</math> is reached.</p> $2H = \{2, 8, 10\}$ $2+2H = \{4, 8, 0\} = H \Rightarrow  2H  = 2$
3	$ 2+H  = 3$ <p>Since <math>H</math> itself is a coset w/ order 3, <math>2+H</math> must have order 3.</p>
4	$\mathbb{Z}_{12} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$ $H = \{0, 4, 8\}$ $ \frac{G}{H}  = \frac{12}{3} = 4 \quad \text{the order} = 4.$
6	<p>since <math>([2]+H)([2]+H) = [2+2]+H</math>, then for</p> $([2]+H)^k = H, \quad k = 6 \text{ or } 0,$ $\underbrace{([2]+H)([2]+H) \dots ([2]+H)}_{6 \text{ terms}} = [0]+H = H.$

**Deletion.** A question was entirely eliminated and not replaced for one topic: order of a group. The question read, “Can an infinite group have a finite subgroup?” This decision to remove the question (and subsequently the topic) was because: 1) The difficulty level was found to be low by the Delphi study with an

average difficulty of 1.64 out of 10; 2) There was no evidence from the open-ended round that students struggled to appropriately deal with the concept of group order. The responses to the open-ended round were either correct or misconceptions were related to understanding of other topics. For example, many students indicated that an infinite group ( $\mathbb{Z}$ ) could have finite subgroups ( $\mathbb{Z}_n$ ), a misconception that mimics issues with the subgroup question (see Paper 2). The question failed to address the desired content area. Further, the concept of group order was reflected in a number of questions including the subgroup question and the Lagrange's Theorem question where students articulated notions of the order of a group while interpreting questions and making arguments. At no point during the interviews or in the open-ended surveys did students reflect anything but an accurate understanding of group order. This is not surprising as the concept itself is not highly complex (counting the number of elements in a set) and does not provide natural interference between an intuitive and formal understanding. Some of the complexity lies in the set concept, particularly that elements cannot repeat. However, this is a non-issue in group theory where sets are generally pre-packaged. Further, sets themselves were not identified as an important topic to explore further in this context based on the Delphi study. This question was markedly disconnected from its purpose and achieved no content validity. A concept that lacks a variety of incorrect conceptions does not meet the purpose of a test designed to probe understanding of various concepts and unearth student difficulties.

**Discussion.** While the domain analysis helped bolster content validity claims, students must engage in the tasks to make strong arguments of construct validity. Through analyzing student responses and conducting follow-up interviews, targeted conceptions can be connected to how students actually engaged in the tasks and thought about the underlying concepts.

Additionally, testing tasks with a diverse sample of students can bolster arguments about generalizability. Many concept inventories were developed internally where only students from a single university comprised the sample. Recent concerns about the reliability of the Calculus Concept Inventory (Inglis, 2015) reiterate the need to be transparent about the creation process. The test performed significantly differently in terms of both reliability and underlying factors when administered to a different population of students (Thomas et al., 2015). For the GCI, multiple institutions were used at each stage to maximize representativeness in order to bolster generalizability.

### **Psychometric Properties of the GCI**

A number of psychometric analyses were used to evaluate the structure of the test after the completion of the large-scale round of field testing.

**Reliability.** The most commonly used measure of reliability is Cronbach's Alpha. Cronbach's Alpha gives a measure of internal consistency by correlating items with the overall item score.

**In the GCI.** For the GCI, the Cronbach's Alpha was calculated giving a reliability estimate of .84 in the pilot and .65 for the large-scale round. The value of .84 indicated good internal consistency, while the .65 was acceptable. This may

reflect the low stakes of test takers. While many of the students received credit for completing the question set, there was no incentive for their performance. A high rate of guessing may account for a less-than-ideal reliability estimate. Notwithstanding, a value of .65 is still in the acceptable range and allows for some degree of cohesiveness.

**Discussion.** Cronbach's alpha is the most commonly used measure of reliability. However, Cronbach's alpha is sample dependent (Wallace & Bailey, 2010). In the case of the CCI, Epstein (2007) found the reliability estimate to be  $\alpha = .7$ . However, in a recent plenary, Inglis (2014) reported reliabilities of  $\alpha = .211$ ,  $.326$ , and  $.397$  with different samples of students. I purposefully selected a representative sample of students to try and contend with the limitations in homogeneous samples. While my reliability was lower with the more diverse sample, the sample was more reflective of the intended population.

The decision to only have one question per concept also limited internal consistency and likely impacted the alpha estimate. The more questions in a test, the higher the alpha reliability estimate in general (Cortina, 1993). In the case of the GCI, each question targeted one underlying concept. The questions would likely have higher correlations if multiple questions targeted the same topic. Increasing correlations between questions would cause positive effect on reliability in terms of an internal consistency estimate.

Furthermore, *because* the sample was representative, the classes of students varied tremendously. The questions were created to maximize access, but with the variation in courses, it is not possible to do so flawlessly. Students may

have differing exposure to various groups, theorems and problem types. This may also contribute to the lower reliability, as students may not access certain questions similarly. If a student has less exposure to certain topics or groups, guessing rates may increase.

Many of these limitations are logistical consequences. The GCI takes roughly 45 minutes to complete, and adding additional questions per topic would make this test quite long for its purpose. By raising the stakes of the test, guessing may be reduced. However, the test was not intended for assigning grades as the purpose is diagnostic in nature and the set of questions is intentionally not comprehensive. Only questions with the power to unearth common incomplete and incorrect conceptions were used.

**Criterion-related validity.** Another common way to address validity is to correlate the scores on a test with some related external measure.

***In the GCI.*** Students who completed the question set also self-reported their course grades. I converted the course grades to A=4, B=3, C=2, D=1 and F=0. As not all institutions used a +/- grading scheme, scores such as A- or A+ were recorded as simply a 4. The course grades were strongly correlated to their performance on the GCI,  $r=.43$ ,  $p<.001$ .

***Discussion.*** Criterion-related validity can provide evidence that an assessment is related to some other construct. But this evidence is external to the test. Lissitz and Samuelson (2007) argued that criterion validity is not a type of validity, but rather a measure for *utility*, an external factor. An instrument that measures how many days of sunshine in a month and an instrument that measures

the rates of robbery may be correlated, however, they certainly do not measure the same thing. The validity of a test should stem from the creation process, the questions themselves, and the connections between the question responses and how students think about a given topic.

**Principle component analysis.** Factor analytic methods, such as principle component analysis, provide a means for determining dimensionality of a test.

*In the GCI.* Principle Component Analysis was used to examine correlations that may exist between question performances. The analysis found the test was largely unidimensional with only one factor with an eigenvalue significantly above one. **Figure 29** illustrates this phenomenon with a clear elbow after one factor with the remaining factors accounting for minimal variation. There were generally low correlations between any items. This makes sense in terms of the test's purpose: to capture *different* facets of student understanding related to the underlying concept of groups. The items were created to target different conceptions, so although they are related through the lens of groups, they each have independent aims.

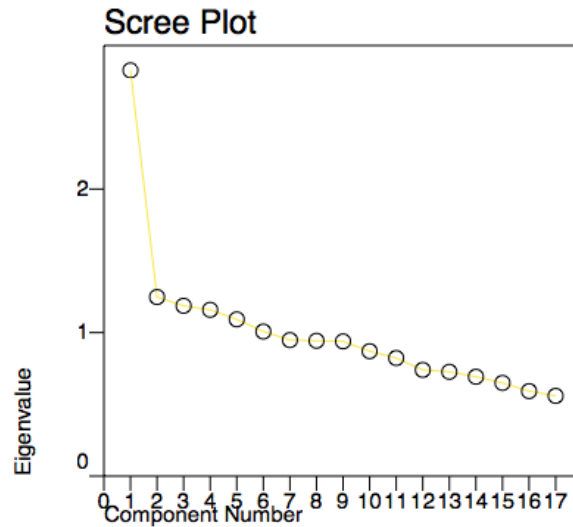


Figure 29. Scree plot illustrating the number of components in the GCI.

**Discussion.** Factor-analytic approaches condense the differing responses into either correct or incorrect, which does not fully capture the scope of a concept inventory. This type of analysis is limiting in a situation where questions have value in their different responses and not just at the level of right and wrong. A more targeted exploration informed by theory may provide a better analysis of the relation between questions. For example, a number of question distractors through the GCI were related to students reducing abstraction and reverting to familiar groups. These distractors may be related even if the questions in their entirety did not have high correlations.

## Conclusion

The methodology for the first rounds of GCI creation is intended to serve as a model for the initial question creation and refinement for concept inventories. The concept inventory was created in several stages: domain analysis, question

creation, and the field-testing of questions. At each stage, I aimed to incorporate the three factors of internal validity to maximize the meaningfulness of the GCI.

**Recommendations.** While psychometric analyses can provide important information about a test's reliability and validity, it is important to consider the context before applying and interpreting them. The nature of test can determine their relative importance. For concept inventories, their meaning is at the individual item option level. Traditional test analysis techniques condense questions to right and wrong to address correlations. This differentiates students who appear to have correct conceptions from those with incorrect, but does not account for the nuances of these sorts of tests. For this reason, attention to generalizability and validity should inform the entire creation process rather than relying on oversimplified psychometric analyses.

When creating assessments, the domain analysis is a critical step to establish content validity. The underlying assumption is that any measure is representative of the targeted domain. For measures aimed at specific course content, experts and textbooks can be leveraged to explore what is valued within a given domain. In the case of concept inventories, the purpose of the assessment is centered on conceptual understanding. A detailed textbook analysis can consider various aspects of concepts as valued by the field including representation types, examples, applications and definitions. Furthermore, this analysis provides a significant amount of background for creating questions that are accessible to the general population and not a particular subset of students in a given course. Additionally, this is the time to explore any literature about student conceptions.



The literature provides a source for tasks, theories on student conceptions, and further reflects the content that is valued in a field. Prior to the creation of any questions, the limits of the domain should be explored in order to bolster a question set's representativeness, generalizability, and potential to unearth student conceptions.

During the stages of piloting questions, starting with open-ended versions is essential to build multiple-choice questions that are authentic to student conceptions. Rather than stemming from the test-creator, the options should stem from genuine student responses. Furthermore, these responses need to be interpretable. Interviews provide the opportunity to probe a student's thinking and investigate interpretations of their answers. Answers need to clearly delineate between students with differing conceptions. If different conceptions can lead to the same answer, the question needs to be revised. The interviews also serve the purpose of providing warrant for the eventual diagnostic quality of the multiple-choice questions. Interviews should be conducted both during open-ended rounds and closed-form rounds. This is done to assure that students choose multiple-choice responses for the same conceptual reasons as the student responses to the open-ended questions. When field testing, the sample of students should represent the population that is being targeted. Limiting samples to a small number of institutions is likely to provide an incomplete view of student conceptions.

A strong assessment is meaningful. For a concept inventory, meaning needs to exist at both at the question level and at the test level. The test should represent the intended domain covering important and valued content.

Furthermore, the test should be connected to student conceptions. For each question, the responses should differentiate between students with differing conceptions and be clearly connected to alternate conceptions related to a topic. Additionally, these interpretations need to generalize beyond a given sample to the larger population; each questions must be accessible provide equal access and not be catered toward any particular subset of students.

By attending to generalizability (in terms of access and using representative samples), construct validity (through using literature on student conceptions, beginning with open-ended questions, and conducting follow-up interviews), and content validity (through using an expert panel and textbook analysis), I created the GCI questions with the purpose of maximizing all three factors of internal validity. The questions covered fundamental topics in the domain, the tasks were accessible to a representative sample of students, and all of the question options reflected genuine student conceptions around related topics.

Concept inventories continue to be a popular tool for a number of purposes including both diagnostic (unearthing student conceptions) and evaluative (assessing instructional efficacy). It is important to be transparent and systematic when creating an instrument intended for such broad usage. The creation of the GCI presents one model of this transparency with careful attention to developing both a useful and valid tool.

## **Paper 2: Scaling up and out: Reconsidering how students think about concepts in Group theory**

*In replication you learn a lot about what is still needed. That is not understood. In mathematics, there is no replication. When you have proved it, it's proved. But we are not mathematicians; we are a human science. And so when somebody has shown something, we have to try to do it again to figure out what the critical variables were that determined it and what might possibly affect the result. Because the result might be an artifact (Silver & Kilpatrick, 1994, p. 738).*

In recent years, numerous behavioral science fields have called for an increase in replication studies (Burman, Reed, & Alm, 2010; Makel & Plucker, 2014; Yong, 2012). The need to repeat findings is especially pertinent in the fields where small sample sizes dominate. Weber (2013) noted at his recent plenary address at the *Conference for Research in Undergraduate Mathematics Education* the dominance of small-scale and qualitative studies calling for quantitative and large-scale counterparts. Without scaling *up* (increasing numbers) or scaling *out* (using samples with differing characteristics), we are limited in terms of generalizability- leaving theories to stand without consideration to the role of time, place, and people involved. Theories risk being artifacts of their setting. Yet, a recent article in the *Educational Researcher* noted that only 0.13% of published articles in the major education journals are replication studies (Makel & Plucker, 2014). Replication is a vital part of the scientific method where theories are to be both created and scrutinized. I address this call by building on smaller qualitative studies related to student understanding in group theory, scaling up and out to utilize the same (or similar) questions with a larger and more diverse sample.

In this study, I report on large-scale survey results around three questions that have been previously explored in earlier studies related to student

understanding in group theory. The first question addresses whether  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ . This question was used to illustrate various student conceptions about both subgroups (Dubinsky, Dautermann, Leron, & Zazkis, 1994) and Lagrange's Theorem (Hazzan & Leron, 1996). A question about isomorphism, ("Are  $\mathbb{Z}$  and  $\mathbb{Q}$  isomorphic?") from Weber and Alcock (2004), was initially used to illustrate proof strategies. In this study, I use the question to both re-explore the original theoretical contribution about proof strategies, as well as identify alternate approaches and conceptions related to isomorphism found in the larger and more representative sample. The third question comes from Lajoie and Mura's (2000) exploration of cyclic groups where they found students struggling to mediate between their intuition of cycles and the formal definition of cyclic groups when determining if  $\mathbb{Z}$  is a cyclic group. Within each question, I discuss the generalizability of the original results and present alternate conceptions that emerged from the larger sample. I explore several issues including: (1) validating previous theories, (2) establishing how widespread various student conceptions are, and (3) exploring how new results may inform previous theory.

### **The Need for Replication**

Schoenfeld (2007) listed *replicability* as an important aspect of trustworthiness in mathematics education research. He equated replicability to *generalizability*. That is, "the expectation is that the lessons learned from a study will apply, in some way, to other situations" (p. 87). Every classroom and every student is different and so, replicability is not just about identically recreating a study, but conducting further studies to replicate the *lesson learned*.

Makel and Plucker (2014) differentiated types of replication into three categories adopted from Lykken's (1968) classic work. The categories are: *literal replications*, *operational replications*, and *constructive replications*. Literal replications are studies that have been replicated exactly including taking a sample from the original sample pool. Operational replications differ slightly as the requirement to pull from the same sample pool is lessened, but the methodology stays unchanged. Constructive replications aim to develop a new study that would either confirm or challenge the findings from another study. For example, if a case study develops a specific theory of how students understand groups, a constructive replication study might use a different institution-type, utilize an alternate methodology and create new questions that also aim to address the conception types associated with groups. The results would either confirm or refute the theory developed in the initial study. The group theory understanding replication studies presented herein are constructive in nature.

Makel and Plucker (2014) analyzed the top 100 education research journals (determined by impact factor) to identify the rates of replication studies. They found that 221 out of 164,589 articles (.13%) were in fact replication studies. A cursory look at the *Journal for Research in Mathematics Education (JRME)* and *Educational Studies in Mathematics (ESM)* reveals similar results. By searching the text for the term "replication" a number of studies were identified; however, most used replication in alternate contexts, did internal replication (that is reporting on a study and their replication of it), called for replication, or mentioned a different replication study in the literature background.

Of the 1,800 articles found in Google Scholar's database for JRME, only 14 were replication studies. Of the 2,160 articles found in Google Scholar's database for ESM, only two studies were replication studies.

As early as 1975, Eastman lamented a lack of replication studies in mathematics education noting that, "we are all aware that one empirical study does not by itself answer a general question about the teaching or learning of mathematics" (p. 67). Despite reported calls for replication studies, these studies remain rare in educational research. The research in this paper aims to both address the call for replications and illustrate the ways that replication can serve to inform theory.

### **Student Understanding in Group Theory**

The replication studies presented are situated in the context of student understanding of introductory group theory. Group Theory is a notoriously challenging course. As noted by Dubinsky, et al. (1994), "mathematics faculty and students generally consider it to be one of the most troublesome undergraduate subjects" (p. 268). Group theory is often the first time students have to reason about concepts from their formal definitions. Literature related to student understanding in group theory either highlights these difficulties, presents instructional innovations aimed to improve student understanding around concepts in the course, or a hybrid of the two (Weber & Larsen, 2008). The three questions explored in this replication study have been used in the past to illustrate student difficulties and provided the impetus for developing theory around student

conceptions in the subject area. I present background on each of these questions within the replication results section. The three questions are:

1. Is  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ ?
2. Is  $\mathbb{Z}$ , the set of integers under addition, a cyclic group?
3. Are  $\mathbb{Z}$ , the integers under addition, and  $\mathbb{Q}$ , the rationals under addition, isomorphic?

## Methods

This study is part of a larger project developing a validated assessment tool in introductory group theory. The results reported in this paper reflect a constructive replication. That is, tasks from prior studies were utilized to determine if their findings were generalizable. However, the tasks were not always identical. The previous studies that used the prompt about whether  $\mathbb{Z}_3$  was a subgroup of  $\mathbb{Z}_6$  varied in form from a direct question to evaluating a classmate that claims  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ . Furthermore, the studies were all open-ended whereas, the later rounds of this study included closed-form versions. Data collection took place in three rounds: large-scale open-ended, pilot closed-form and large-scale closed-form. During each stage, students from a variety of institution types participated.

## The Survey

The survey consisted of 18 questions related to topics determined to be essential by a panel of experts consisting of algebraist course instructors, non-algebraists course instructors, textbook authors and group theory mathematics education researchers. The questions used in this survey derive from a detailed textbook analysis and from existing literature related to the group concept. (For a

more detailed discussion of this process see Paper 1). Several of the 18 questions were adapted from previous studies. For the purpose of this paper, I present results from the aforementioned questions.

### **The Participants**

Instructors currently teaching introductory undergraduate abstract algebra courses were contacted across the United States. The Phase 1 sample was selected randomly within the classifications of region (West, Southeast, Northeast, Midwest, Mid Atlantic and New England) and selectivity (greater than 75% of applicants admitted; 50-75% admitted; 25-50% admitted, and less than 25% of applicants admitted.) The open-ended first round was limited to institutions with publically available course listings (about 90% of institutions provide that information). For the second and third round, an effort was made to contact all current introductory abstract algebra instructors at institutions that offer a mathematics major. When no course listing was available, a member of the research team called the institution to determine if a course was running and contacted the instructor. Geographic regions and selectiveness were leveraged to address how representative a given sample was. One of the major goals of replication is to address generalizability. This was particularly important in terms of selectiveness where the typical student at one of the most selective institutions has a significantly different background than the typical student at less selective institutions.

The open-ended round included 29 institutions (349 students). The closed form rounds (pilot and large-scale) included 8 institutions (87 students) and 32



institutions (376 students) respectively. The selectiveness breakdown can be found in Table 14. In the United States, there are many more institutions at the least selective and mid-level than in the categories of more and most selective.

Table 14

<i>Selectivity of Sample Institutions</i>					
	Least Selective (>75% admitted)	Mid-Level Selective (50-75% admitted)	More Selective (25-50% admitted)	Most Selective (<25% admitted)	Not classified
Open-ended Round	13 classes (138 students)	12 classes (108 students)	4 classes (47 students)	1 class (57 students)	0 classes
Multiple Choice Pilot	2 classes (17 students)	0 classes	3 classes (26 students)	4 classes (44 students)	0 classes
Multiple Choice Large-scale	13 classes (131 students)	10 classes (128 students)	6 classes (84 students)	1 class (14 students)	2 classes (19 students)
Total	28 classes (286 students)	22 classes (236 students)	13 classes (157 students)	6 classes (119 students)	2 classes (19 students)

## Interviews

During both the open-ended and closed-form rounds, interviews were conducted with 15 students (from 4 institutions for open-ended and 13 institutions for closed-form round) for a total of 30 interviews. The interviews served to strengthen the validity of interpretations of answers and allowed for deeper probing of conceptions. For each survey question, the student was asked to explain their thinking about the question by walking through their written response or explaining their multiple-choice option selection. If they did not address the meaning of a given concept, they were asked to explain their understanding of the relevant concept. The interview was semi-structured to allow for additional follow-up questions to better make sense of the students' thinking.

## **The Analysis**

The open-ended surveys were analyzed using a thematic analysis approach (Braunn & Clarke, 2006). An initial round of open-coding led to a development of themes within each question type. The themes were refined into a set of codes that were then applied to the entire set of questions. A subset of 270 student responses for each of the three questions in this paper, was then double-coded by a mathematics education graduate student for a reliability of 96% for the subgroup question, 94% for the cyclic group question, and 91% for the isomorphism question. This percentage was calculated by summing the number of codes agreed upon divided by the sum of the highest amount of codes on each item. The closed-form versions were explored using descriptive statistics. When appropriate, I compared the proportion of students with response types from the original studies to the replication rounds using a 2-sample proportion test in order to test the assumption that these proportions are the same in both samples. Student interviews were first analyzed based on the response conception corresponding to the open-ended coding categories. Each interview response was further analyzed in light of its relation to the theories from the original studies being replicated. If a student's response was not consistent with the original studies, I marked the response, then re-analyzed it to address what the response means in terms of the original theory.

## **Replication Results and Discussion**

In this section, I provide the results for each of the three replicated questions. Within each question, I first explain the relevant mathematics in order

to make sense of the prompts and the student responses. I then consider the original findings. After providing this background, I present the results from the large-scale replication studies. In each case, I present various ways that the new replication studies can inform the refinement of prior theories. This is done through a combination of analyzing both the surveys and subsequent follow-up interviews.

### **Replication: Is $\mathbb{Z}_3$ a subgroup of $\mathbb{Z}_6$ ?**

The groups  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  can be defined in one of two ways: as quotient groups or through clock arithmetic. Often the elements in  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  are treated as a subset of the integers with differing operations. So  $\mathbb{Z}_3$  would be the set  $\{0,1,2\}$  under addition modulo 3. For example,  $1+2$  equals 0 in this group, as 0 would be the remainder when  $1+2$  is divided by 3. Similarly,  $\mathbb{Z}_6$  would be  $\{0,1,2,3,4,5\}$  under addition modulo 6. Alternately,  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  can be thought of as quotient groups.  $\mathbb{Z}_6$  would be  $\mathbb{Z}/6\mathbb{Z}$  and  $\mathbb{Z}_3$  would be  $\mathbb{Z}/3\mathbb{Z}$ . The set  $\{0,1,2\}$  would be merely representative elements and better expressed as the set of cosets:  $\{0+3\mathbb{Z}, 1+3\mathbb{Z}, 2+3\mathbb{Z}\}$  where  $0+3\mathbb{Z} = \{\dots, -6, -3, 0, 3, 6, \dots\}$ ,  $1+3\mathbb{Z} = \{\dots, -5, -2, 1, 4, 7, \dots\}$  and  $2+3\mathbb{Z} = \{\dots, -4, -1, 2, 5, \dots\}$ . Similarly,  $\mathbb{Z}_6$  would consist of the set:  $\{0+6\mathbb{Z}, 1+6\mathbb{Z}, 2+6\mathbb{Z}, 3+6\mathbb{Z}, 4+6\mathbb{Z}, 5+6\mathbb{Z}\}$ .  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  are examples of a family of groups of the form  $\mathbb{Z}_n$ . They share similar structure built around binary operations that differ only in the modulus.

A subgroup is a subset of a group that forms a group itself under the same operation. Using the clock arithmetic interpretation,  $\mathbb{Z}_3$  would be a subset of  $\mathbb{Z}_6$ . However the operation would be different. In  $\mathbb{Z}_3$ ,  $1+2=0$ , but in  $\mathbb{Z}_6$ ,  $1+2=3$ . A

subgroup is a subset that forms a group under the *same* operation. So,  $\mathbb{Z}_3$  would not be a subgroup of  $\mathbb{Z}_6$ . Using the quotient group interpretation,  $\mathbb{Z}_3$  is not even a subset of  $\mathbb{Z}_6$ . Consider the element representative element 1 in each group. In  $\mathbb{Z}_3$  this is the element  $1+3\mathbb{Z} = \{\dots, -5, -2, 1, 4, 7, \dots\}$ , but in  $\mathbb{Z}_6$ , the respective element is  $1+6\mathbb{Z} = \{\dots, -11, -5, 1, 7, \dots\}$ .

However,  $\mathbb{Z}_6$  does have a subgroup that is isomorphic to  $\mathbb{Z}_3$ . The subgroup  $\{0, 2, 4\}$  is isomorphic to  $\mathbb{Z}_3$ . In fact, because  $\mathbb{Z}_6$  is cyclic, it is guaranteed to have subgroups of the form  $\mathbb{Z}_n$  for all  $n$  that divide 6. This is sometimes stated as part of the Fundamental Theorem of Cyclic Groups which states that: (1). Every subgroup of a cyclic group is cyclic; (2). The order of every subgroup is a divisor of the order of the cyclic group; and (3). There is exactly one subgroup of each order that divides the order of the cyclic group.

**Prior results.** The question of determining if  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  has been discussed numerous times in the mathematics education literature. The question initially served the purpose of theorizing how student conceive of subgroups and highlighting that that students note be coordinating both a set and operation. Notably, Dubinsky, Dautermann, Leron, and Zazkis (1994) presented exchanges with three students with varying levels of coordination ranging from identifying  $\mathbb{Z}_3$  as a subgroup of  $\mathbb{Z}_6$  with no attention to a differing operation to immediately stating this is not possible because of the differing operation. The researchers identified the coordination of operation and set as an essential aspect of understanding group and subsequently subgroup. Burn (1996) quickly challenged the use of this question noting that “there is something commendable

about suggesting that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  ..., since every cyclic group of order 6 has a cyclic subgroup of order 3” (p. 373). He was referencing the advanced conception of isomorphism where the set  $\{0,2,4\}$  could be thought of as the same as  $\mathbb{Z}_3$  and therefore  $\mathbb{Z}_3$  would in fact be a subgroup of  $\mathbb{Z}_6$ . Dubinsky, Dautermann, Leron, and Zazkis (1997) countered that this level of sophisticated thinking was not occurring as the students articulated that they were addressing the elements 0,1 and 2.

Leron and Hazzan (2006) and Hazzan and Leron (1996) addressed this idea when presenting their analysis of 113 computer science students responses to this question. They prompted, “A student wrote in an exam, ‘ $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ ’. In your opinion is this statement true, partially true, or false? Please explain your answer” (Leron & Hazzan, p. 199). They found that many of their students were using an invalid form of the converse of Lagrange’s Theorem to state that  $\mathbb{Z}_3$  was a subgroup of  $\mathbb{Z}_6$ . They noted:

[t]here is a sophisticated sense in which the statement “ $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ ” is partially true, namely, that  $\mathbb{Z}_3$  is isomorphic to the subgroup  $\{0, 2, 4\}$  of  $\mathbb{Z}_6$ . We would of course be thrilled to receive this answer, but none of our 113 subjects had chosen so to thrill us (Leron & Hazzan, p.119).

Leron and Hazzan (1996) concluded that their subjects were using superficial clues and consequently misapplying Lagrange’s Theorem.

Brenton and Edwards (2003) contended that the confusion with this question could be attributed to mistreating the groups  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$ . Often when using modular arithmetic, both  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  are treated as subgroups of  $\mathbb{Z}$  where only the operation differs. Brenton and Edwards suggest that  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  should be

treated as quotient groups. The simplification to elements of  $\mathbb{Z}$  creates an unnecessary obstacle. “ $\mathbb{Z}/3\mathbb{Z}$  is naturally a quotient group, not a subgroup of  $\mathbb{Z}$  and of  $\mathbb{Z}/6\mathbb{Z}$ . Students should then not mistake  $\mathbb{Z}_3$  for a subgroup of  $\mathbb{Z}_6$ , since as sets they do not share even one element in common” (p. 35). While an interesting conjecture, it is possible that students may attend only to representative elements and still come to the incorrect conclusion that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ . In fact, Seibert and Williams (2003) found that students struggled when dealing with  $\mathbb{Z}_n$  as a quotient group. They found that students did attend solely to representative elements even within the quotient group structure.

Table 15

*Previous Results: Is  $\mathbb{Z}_3$  a Subgroup of  $\mathbb{Z}_6$ ? (Dubinsky et al., 1994); A Student Wrote in an Exam, ‘ $\mathbb{Z}_3$  is a Subgroup of  $\mathbb{Z}_6$ ’. In Your Opinion is this Statement True, Partially True, or False? Please Explain Your Answer (Leron & Hazzan, 1996).*

	Dubinsky et al. (1994) n=3	Leron & Hazzan (1996) n=113
Yes	66.7%	64.6%
-Yes, by converse of Lagrange	-	17.7%
-Yes, Other		46.9%
No	33.3%	-

In both Dubinsky et al. (1994) and Leron and Hazzan (1996)’s studies, their sample was from a single institution. Therefore, it is impossible to generalize how widespread either of these issues are and perhaps, as suggested by Brenton and Edwards (2003), they are artifacts of the didactical treatment of the groups  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$ . Further, Leron, and Hazzan lamented the lack of advanced treatment of the isomorphic copy of  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  amongst their 113 subjects. However, it is possible, and was found in the replication study, that introductory students are

capable of making this advanced argument. Table 15 represents the reported results on the respective  $\mathbb{Z}_3 / \mathbb{Z}_6$  subgroup prompts from Dubinsky et al, (1994) and Leron and Hazzan (1996).

**Methods for the replication study.** As each of these studies is a constructive replication, there are alterations between the original study and the replication study. The original studies included two different questions, one that directly asked, “Is  $\mathbb{Z}_3$  a subgroup of  $\mathbb{Z}_6$ ?” and a second that had students evaluate a hypothetical student claim that  $\mathbb{Z}_3$  was a subgroup of  $\mathbb{Z}_6$ . In the replication studies, the prompt initially was the same as the former asking directly, “Is  $\mathbb{Z}_3$  a subgroup of  $\mathbb{Z}_6$ ?” This question was used in the open-ended round with 349 students. In the closed-form round the question was altered to “Is the set  $\{\bar{0}, \bar{1}, \bar{2}\}$  a subgroup of  $\mathbb{Z}_6$ ?” This decision was made to be able to clearly delineate student conceptions based on their selection from the multiple-choice question. The multiple-choice nature of later rounds was another departure from the original studies. This was done to allow for a clean collection of data with minimal interpretation.

**Results from the replication study.** The results from the open-ended round, and the two closed-form rounds can be found in

Table 16,  
Table 17, and Table 18 respectively. The remaining students not reported in the open-ended round provided responses that were not discernable as valid or invalid based on providing only an argument that 3 divides 6.

Table 16

<i>Open-Ended Round Results: Is <math>\mathbb{Z}_3</math> a subgroup of <math>\mathbb{Z}_6</math>?</i>	
	$n = 349$
Yes (valid reasoning)	7.4%
Yes (invalid reasoning)	58.5%
Because $3 6$	9.5%
Because $\mathbb{Z}_3$ is a subset of $\mathbb{Z}_6$ .	7.1%
Because $\mathbb{Z}_3$ is a subset of $\mathbb{Z}_6$ and meets requirements	30.7%
No (valid reasoning)	26.1%
No (invalid reasoning)	2.6%

Table 17

<i>Closed-Form Pilot Results: Is the Set <math>\{\bar{0}, \bar{1}, \bar{2}\}</math> a Subgroup of <math>\mathbb{Z}_6</math>?</i>	
	$n = 81$
Yes, because $\{\bar{0}, \bar{1}, \bar{2}\}$ is a subset of $\mathbb{Z}_6$ .	10.3%
Yes, because $\mathbb{Z}_3$ is a group itself contained in $\mathbb{Z}_6$ .	21.8%
Yes, because 3 divides 6.	6.4%
No, because the subset $\{\bar{0}, \bar{1}, \bar{2}\}$ is not closed.	60.3%
None of the above	0%
I don't know	1.3%

Table 18

<i>Closed-Form Large-Scale Results: Is the Set <math>\{\bar{0}, \bar{1}, \bar{2}\}</math> a Subgroup of <math>\mathbb{Z}_6</math>?</i>	
	$n = 376$
Yes, because $\{\bar{0}, \bar{1}, \bar{2}\}$ is a subset of $\mathbb{Z}_6$ .	13.8%
Yes, because $\mathbb{Z}_3$ is a group itself contained in $\mathbb{Z}_6$ .	36.7%
Yes, because 3 divides 6.	6.1%
No, because the subset $\{\bar{0}, \bar{1}, \bar{2}\}$ is not closed.	43.4%

All of the student conceptions found in the prior studies existed to some degree in the replication rounds. I conducted a 2-sample proportion test to compare the proportion of students in Leron and Hazzan's study who identified that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  for reasoning unrelated to Lagrange's Theorem to the



proportion of students who did this in the open-ended round of the replication study. The proportion of students saying that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  with invalid reasoning was lower, although not significantly, in this sample with  $p=.378$  compared to Leron and Hazzan's (1996)  $p=.469$ ,  $z=-1.72$ ,  $p=0.09$ . Comparing the number of students incorrectly applying the converse of Lagrange's theorem was not possible in this round as many student responses could be interpreted as valid or invalid depending on context. For example, a statement such as "Yes, because  $3|6$ " would be invalid if referring to Lagrange's Theorem, but valid if leveraging the fact that  $\mathbb{Z}_6$  is a cyclic group. (See the previous discussion on the relevant mathematics for a more detailed treatment of this idea). However, the open-ended round results illustrated that some undergraduate students are capable of identifying the isomorphic copy of  $\mathbb{Z}_3$  in  $\mathbb{Z}_6$  with 7.4% providing a valid argument of this nature.

Between the open-ended and closed-form rounds, the question was changed from, "Is  $\mathbb{Z}_3$  a subgroup of  $\mathbb{Z}_6$ ?" to, "Does the set  $\{\bar{0}, \bar{1}, \bar{2}\}$  form a subgroup in  $\mathbb{Z}_6$ ?" In this way, any argument based on 3 dividing 6 would in fact reflect an incorrect conception. In the closed-form rounds, only 5% and 6% of students, respectively, misapplied Lagrange's Theorem. Using a 2-sample proportion test assuming equal proportions, the combined proportion of students misapplying an argument about divisors in the closed-form rounds ( $p=0.061$ ) was significantly lower than the proportion of students doing so in Leron and Hazzan's study ( $p=0.177$ ),  $z=-3.97$ ,  $p<.001$ . However, this difference could

potentially be attributed to the different question structure and the usage of a closed-form version.

Dubinsky et al. (1994)'s findings were bolstered by the replication study as students consistently used incomplete conceptions of subgroups. Notably, many students just stated that  $\mathbb{Z}_3$  was a subset and therefore a group. (Although, this may represent an incomplete response, rather than lack of attention to operation. If a student knows that  $\mathbb{Z}_3$  is a group, they may just focus on checking that  $\mathbb{Z}_3$  is a subset of  $\mathbb{Z}_6$ .) Many students provided reasoning similar to the cases reported in the original study:  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  because  $\mathbb{Z}_3$  is both a subset and a group itself. Approximately a third of the sampled students provided this type of reasoning. Shifting from asking if  $\mathbb{Z}_3$  was a subgroup in  $\mathbb{Z}_6$ , to asking if  $\{\bar{0}, \bar{1}, \bar{2}\}$  was a subgroup in  $\mathbb{Z}_6$  did not appear to mitigate the issue with a slightly higher proportion of students responding in that manner. It is worth noting that the most common response in both the pilot closed-form round (62%) and the large closed-form round (44%) was the correct response. These students used the correct operation (inherited from  $\mathbb{Z}_6$ ) to explain that the subset was not closed. The discrepancy in these numbers is likely due to the overrepresentation of selective schools in the pilot while the full-scale round was more representative of national institutional breakdown in selectivity.

The follow-up interviews allowed for the probing of two issues related to this question: 1) If students treated the elements as cosets, did they avoid the issue? and 2) Are students aware that they are using different operations when treating the subset  $\{\bar{0}, \bar{1}, \bar{2}\}$  as  $\mathbb{Z}_3$ ? The following discussion highlights some of the

ways that theory could potentially be refined through the usage of replication studies. In the first case, including a more representative sample allowed for exploring differences amongst sub-portions of a population who had different experiences with modular addition groups. While Brenton and Edwards (2003) suggest that students may not struggle with the question if they treat the modular addition groups as quotient groups, the previous studies did not allow for any empirical consideration of this suggestion. In the second case, I present evidence that potentially challenges the original interpretations of student understanding of subgroups. The preliminary follow-up interviews included responses that indicated students might not be failing to coordinate operation and subset, but rather lack a robust understanding of binary operation itself.

*A pedagogical difference: The treatment of  $\mathbb{Z}_n$  as a quotient group.*

Although most students discussed  $\mathbb{Z}_3$  in terms of clock arithmetic, one of the interviewed students, Bob, treated the elements as cosets during throughout his survey response and follow-up interviews.

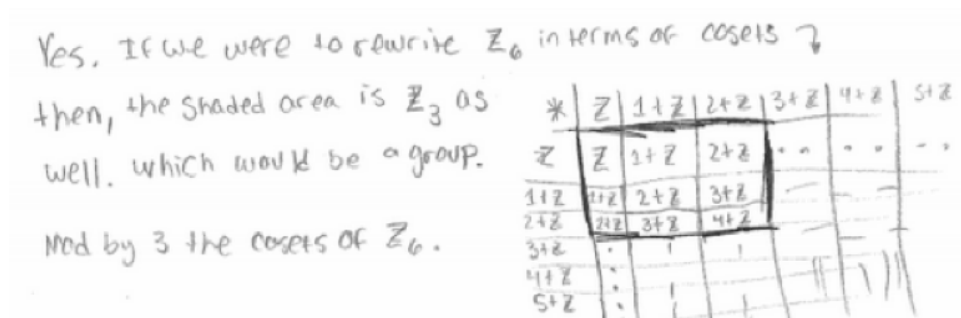


Figure 30. Bob's work on the quotient group question.

There are two pieces of Bob's method worth noting. First, as can be seen in Figure 30, he has written these cosets as  $1+\mathbb{Z}$  rather than  $1+6\mathbb{Z}$ . While he clearly sees the elements as sets, he likely has some conceptual limitations about

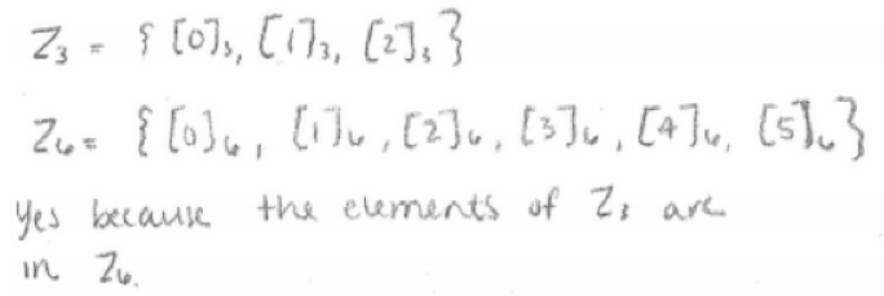
coset formation. He may be seeing  $+\mathbb{Z}$  as part of the ritual and using representative elements to treat this quotient group as if it were clock arithmetic. Furthermore, the fact that  $1+\mathbb{Z}$  looks the same in both  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  may be compounding the issue. When asked to discuss his written work on the quotient group question, Bob provided the following explanation:

So I did you know, I used the question group for  $\mathbb{Z}_6$ . Integers,  $1+\mathbb{Z}$  etc. This is  $\mathbb{Z}_3$  because you have three elements. And it follows all of the requirements. The only thing is here, I have  $3+\mathbb{Z}$  and  $4+\mathbb{Z}$  and you might say it's not closed but those can be rewritten because  $3+\mathbb{Z}$  would just be  $\mathbb{Z}$  and  $4+\mathbb{Z}$  would just be  $1+\mathbb{Z}$ .

Based on his discussion around the elements, Bob indicated he recognized that these are different sets. He noted that in  $\mathbb{Z}_6$ , there is  $4+\mathbb{Z}$ , but switches to a different equivalence class structure because in  $\mathbb{Z}_3$ , 4 would be in the  $1+\mathbb{Z}$  [sic] coset. Rather than changing just the operation, he appeared willing to restructure the cosets, adapting to the structure of  $\mathbb{Z}_3$ .

Bob's approach was not an anomaly when compared to the open-ended surveys. Several students included equivalence class notation with a subscript three and six on the elements, respectively (see *Figure 31*). These differences in label did not appear to cause the students' disequilibrium when identifying  $\mathbb{Z}_3$  as a subgroup of  $\mathbb{Z}_6$ . These cases indicate that treating the modular arithmetic groups as quotient groups may not be an easy pedagogical fix. Quotient groups are a challenging and complex topic because students must conceive of a group whose elements are sets and make sense of what that means for the resulting operation (Asiala, Dubinsky, Matthews, Morics & Oktac, 1997; Johnson & Larsen, 2012; Siebert & Williams, 2003). Learning and understanding the quotient group

structure may be more difficult than directly addressing the differing operation in modular arithmetic cases. These results provide an example of the power of replication studies to inform theory by providing information about more diverse samples. The representative samples included students who received a diverse pedagogical treatment of the modular arithmetic groups.



$$\mathbb{Z}_3 = \{[0]_3, [1]_3, [2]_3\}$$

$$\mathbb{Z}_6 = \{[0]_6, [1]_6, [2]_6, [3]_6, [4]_6, [5]_6\}$$
 Yes because the elements of  $\mathbb{Z}_3$  are in  $\mathbb{Z}_6$ .

Figure 31. Student response to subgroup question with differing element notation.

#### ***Challenging prior theory: effect of student conceptions of binary***

**operation.** In Dubinsky et al.'s (1994) study, they attributed students' failure to recognize that  $\mathbb{Z}_3$  was not a subgroup of  $\mathbb{Z}_6$  as a result of failing to coordinate the operation with the subset. While students might not be attending to subgroups as subsets with operations *inherited*, there may be additional factors at play. The structure of  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$  are quite similar beyond just their elements. In fact, it is likely that the modular arithmetic operation was defined for  $\mathbb{Z}_n$  generally, which may mask the differing operations in  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$ . The only difference in their operation is that of what modulus is being used. Students may not realize that the similar structures are different with different binary operations. Students' conceptions of binary operation might be the cause for failing to coordinate the operations in the desired ways. A binary operation differs if there are any elements that when operated with each other produce a different result (see the

previous section for discussion on the different operation in  $\mathbb{Z}_3$  and  $\mathbb{Z}_6$ ). The operations do seem similar in the generalized sense- they are versions of “addition.”

A number of the follow-up student interviews illustrated this issue. In both the open-ended round and the closed-form round, students addressed that the operation was “addition” without attending to the differences between addition modulo 3 and modulo 6. Elizabeth explained that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  because, “It has to be a subset of  $\mathbb{Z}_6$ , which it satisfies. You already have the group’s operation you inherit from it. It must satisfy closure. And as long as it’s non-empty.” As can be seen from explanation, she explicitly mentions that the group’s operation is inherited.

During the closed-form round, a student, Georgia, explained the definition for subgroups as, “We define it that, if  $H$  is a subset of  $G$  and  $H$  is a group using the same operation as  $G$ , then  $H$  is a subgroup.” She continued to the question explaining, “Then  $\mathbb{Z} \bmod 3$ , if we assume is under modular addition still, it’s the same operation as our group,  $\mathbb{Z} \bmod 6$ .” Several other students explicitly mentioned the operation in a similar manner. Students may see three different aspects of modular groups: the set, the operation, and the modulus. If a subset inheriting an operation is to mean that the containing group and the subset have the same operation, students need to be aware of what it means to have the same operation. This replication study raised questions as to the original interpretation of student responses. Follow-up studies that use sets with more obvious differing operations could serve to further determine whether the issue was a lack of

coordinating set and operation or was due to incomplete conceptions of binary operation.

Overall, the replications studies revealed that all of the conceptions identified by Dubinsky, et al. (1994) and Leron and Hazzan (1996) existed amongst the larger sample. However, many students engaged with the questions differently including a number of students providing a sophisticated isomorphism argument. Furthermore, the follow-up interviews revealed that students may have conceptual limitations around their understanding binary operation rather than a simple failure to coordinate set and operation. The replication studies served to validate previous theories, explore different facets of student responses, and raise challenges to prior interpretations.

**Replication: Is  $\mathbb{Z}$ , the integers under addition, cyclic?**

A group  $G$  is cyclic if there exists an element,  $x$  such that  $\{x^n \mid n \in \mathbb{Z}\} = G$ . That element is referred to as a generator and it is said that  $x$  generates all of  $G$ . In finite groups, this would be equivalent to the elements that are generated by repeatedly operating  $x$  with itself. For example, consider  $\mathbb{Z}_6$  (defined with clock arithmetic).  $\mathbb{Z}_6$  has generator 1. All of  $\mathbb{Z}_6$  can be generated by 1 as follows:

$$\begin{aligned}1 &= 1 \\1+1 &= 2 \\1+1+1 &= 3 \\1+1+1+1 &= 4 \\1+1+1+1+1 &= 5 \\1+1+1+1+1+1 &= 6 = 0\end{aligned}$$

However in an infinite group, such as  $\mathbb{Z}$ , this is not the case as repeatedly adding a positive element only arrives at positive numbers. However,  $\mathbb{Z}$  can be generated

by a single element, 1 (or -1) because 1 can be raised to positive or negative powers. The definition of generating has to be modified to account for additive notation. In  $\mathbb{Z}$ ,  $5^n$  would actually be  $5+5+\dots+5$ . By definition, an element raised to a negative one power is its inverse. So,  $1^{-1} = -1$  and  $1^{-3}$  would represent  $-1+-1+-1$  or  $3(-1)=-3$ . In this way, 1 can generate all of the elements.

**Prior results.** The question of whether  $\mathbb{Z}$  under addition is cyclic, stems from a study conducted by Lajoie and Mura (2000). They identified a potential conflict between students' formal and informal understanding of cyclic groups. They found students neglecting formal definitions in favor of intuition. Of 28 student responses, they found 18 students thought  $\mathbb{Z}$  was not cyclic. Through student interviews and subsequent surveys, Lajoie and Mura identified some informal conceptions of cyclic that students leveraged to address the prompt about  $\mathbb{Z}$ . This included believing that cyclic groups must "cycle." That is, if one begins with a generating element and then operates it with itself, they should eventually return to the initial element. Many students concluded that all cyclic groups are finite. The researchers suggested that students should be redirected to the formal definition to counter this semantically caused issue. A breakdown of their results can be found in Table 19 and Table 20.

Table 19

*Results From Lajoie & Mura's (2000): For each of The Following Groups, Say Whether or Not It Is Cyclic and Justify Your Answer. (a) The Set  $\mathbb{Z}$  under Addition.*

	$n = 29$
Yes	34.5%
No	62.1%
No Response	3.4%



Table 20

*Results from Lajoie & Mura's (2000): The Following Remarks were Made by Students in an Algebra Course. For each One, Say Whether You Agree or Disagree and Explain Why.*

	$n = 29$
Agreed: In a cyclic group, you start with an element, go round all the elements, then return to the first one.	51.7%
Agreed: In a cyclic group, there is an element that yields all the other elements when it is repeatedly combined with itself.	51.7%
Agreed: All cyclic groups are finite.	58.6%

While this issue existed amongst students in this specific class, there is evidence that this may not be universal. In Weber and Alcock's (2004) study, they interviewed four undergraduate students who did know that  $\mathbb{Z}$  was cyclic. Each of these studies was limited to specific institutions, so the replication studies served to address the frequency of this issue amongst a more representative sample.

**Methods for the replication study.** In the replication study, the question was more focused specifically on  $\mathbb{Z}$  asking, "Is  $\mathbb{Z}$ , the set of integers under addition, a cyclic group?" In the original study, students were asked to address  $\mathbb{Z}$  along with several other groups. They were also asked to agree and disagree with statements about cyclic groups independently of this question. These facets were condensed in the replication study. This question was largely unaltered between the open-ended rounds and the follow-up rounds. However, in the open-ended round, students were further prompted to explain, "Why or why not?" and "address what it means to be cyclic in your answer." This was done to ensure more than a "yes" and "no" response.

**Results from the replication study.** Table 21, Table 22, and Table 23

present the result from the three rounds in the replication study.

Table 21

*Open-ended Round Results: Is  $\mathbb{Z}$ , the Set of Integers under Addition, a Cyclic Group?*

	<i>n</i> =349
Yes	75.1%
No	19.8%
Infinite groups cannot be cyclic	4.9%
Does not cycle	19.8%
No element will generate whole set	7.7%
Not answered	5.1%

Table 22

*Closed-form Pilot Round: Is  $\mathbb{Z}$ , the Set of Integers under Addition, a Cyclic Group*

	<i>n</i> =81
Yes, because $\mathbb{Z}$ can be generated by two elements (1 and -1).	8.6%
Yes, because $\mathbb{Z}$ can be generated by one element (1).	60.5%
No, because $\mathbb{Z}$ is infinite and elements do not cycle.	17.3%
No, because any element only generates part of the set (ex: 1 would only generate the positive integers.)	13.6%
None of the above	0%
I don't know	0%

Table 23

*Closed-form Large-Scale Round: Is  $\mathbb{Z}$ , the Set of Integers under Addition, a Cyclic Group*

	<i>n</i> =376
Yes, because $\mathbb{Z}$ can be generated by a set of two elements.	12.0%
Yes, because $\mathbb{Z}$ can be generated by one element.	55.9%
No, because $\mathbb{Z}$ is infinite and elements do not cycle.	18.4%
No, because any element only generates part of the set (ex: 1 would only generate the positive integers.)	14.8%

In all three rounds of this study, a sizeable majority of students felt that  $\mathbb{Z}$  was a cyclic group. This is not surprising as the set of integers under addition is a standard example of cyclic groups found across textbooks (Gallian, 2009; Hungerford, 2012; Gilbert & Gilbert, 2008; Fraleigh, 2002). I conducted a 2-sample proportion test, assuming equal proportions, to determine if there was a significant difference between Lajoie and Mura's students and the students in my replication study. The proportion of students identifying  $\mathbb{Z}$  as cyclic in Lajoie and Mura's (2000) study ( $p=.345$ ) was significantly lower than the proportion of students in the replication studies ( $p=.717$ ),  $z=-4.32$ ,  $p<.001$ . In both the open-ended round and closed-form round, students chose "no" for the same reasons identified by Lajoie and Mura, namely that cyclic groups cannot be infinite as they fail to cycle or that no element can generate all of the integers using a repeated operation approach.

Lajoie and Mura (2000) framed their discussion primarily around the disconnect between everyday language and formal mathematical definitions. They found their students using the idea of repeatedly operating on an element and "cycling" back to a starting point as their primary conception of cyclic groups. This led to many students incorrectly determining that  $\mathbb{Z}$  was not cyclic. However, in this larger study, I found that the majority of students did feel that  $\mathbb{Z}$  was cyclic. Through follow-up interviews, I found that many students still relied on a similar intuitive understanding of cyclic groups, but found ways to adapt their image to incorporate the infinite case,  $\mathbb{Z}$ . By interviewing students who felt  $\mathbb{Z}$  was cyclic, I

was able to address whether the intuitive understanding had the limiting implications suggested in the original study.

***Reconciling the intuitive definition of cyclic with infinite groups.*** Of the thirty students interviewed, twenty of the students explained their concept of generating as repeatedly operating an element with itself. While this was unsurprising for students that determined  $\mathbb{Z}$  was not cyclic, this intuitive idea was quite prevalent with “Yes” responses with twelve of those twenty students responding that  $\mathbb{Z}$  was cyclic. In ten of those twelve cases, the students used the fact that  $\mathbb{Z}$  can be built by using 1 and -1. While just operating 1 with itself produces the positive integers, operating -1 with itself will produce all of the negative integers. Furthermore, if you add 1 to -1, you arrive at 0. So, 1 and -1 could generate the whole set by just operating combinations of those two elements. When asked why -1 could also be used if 1 is the generator, three different explanations emerged:  $\mathbb{Z}$  is a group (and so 1 has an inverse), the set generated by 1 must be a group, and a return to the formal definition of cyclic.

In the first case type, the student used the fact that  $\mathbb{Z}$  is a group to explain why the inverse of 1 would be in the set generated by 1. For example, Stan explained that all the elements in  $\mathbb{Z}$  can be generated by adding 1s and -1s. He elaborated that, “when you have one, you automatically have negative one.” When prompted to explain why the negative one was automatic, he continued, “[w]ell,  $\mathbb{Z}$  is a group right? And based on the definition a group, there exists inverses.” This return to  $\mathbb{Z}$ ’s group properties was not uncommon when students were pushed to explain why they could use -1 to build  $\mathbb{Z}$ . However, this is not a

valid approach. At this point the student is dealing with two separate sets: the group itself ( $\mathbb{Z}$ ) and the set of elements generated by 1. The student is borrowing from the properties of  $\mathbb{Z}$  to make an argument about the second set. While mathematically they are the same set, this argument for why is circular. These students are arguing that the set generated by 1 is  $\mathbb{Z}$  by only addressing properties of  $\mathbb{Z}$  appeared to be disconnected from the set they should be discussing.

In contrast, another student, Chad, debated whether -1 could be used in conjunction with 1 to build the integers. He first worried that, “the inverse of one won’t be in there because you will never go forward enough.” This, again, mimics the concern that 1 will never cycle back. However, he decides, “[i]t would be really dumb for us not to include the inverse in the group generated by, otherwise it wouldn’t be a group.” He had familiarity with generating sets and their relationship with groups and continued, “[w]hat we are generating better be a group, so we better include inverses. It’s just -it should be this way, it doesn’t mean it necessarily is.” While Chad remained unsure of -1’s place in the set generated by 1, he immediately recognized the utility of including it. Unlike the case above, Chad was not using  $\mathbb{Z}$  to argue about this set, but rather he was struggling to reconcile his intuitive understanding of generating and his knowledge that a set generated by an element *should* be a group. It may be worth noting that historically, cyclic groups were described in the finite case and relied on products of elements (Kleiner, 2007). Products are defined in terms of taking combination of elements and operating on them. This repeated operation works for finite groups as illustrated in the background section. It is likely this utility

that led to products including negative powers, which are defined as powers of the inverse. This was a convention decision and not a direct consequence of the meaning of powers. Chad's thinking could be quite powerful for connecting to the necessity of the formal definition relying on powers of elements. If we want all elements to generate groups, and not just those with finite orders, it is necessary to define power in such a way that it is not just repeatedly operating on an element with itself.

In a third case, Isaac was able to connect  $-1$  to the formal definition of cyclic. He explained, "it's generated in the way you take powers of one or in this case, you take one plus one plus one and inverses and you get the whole set back." When asked why inverses are also used, he explained:

Because, generating means - generating by definitions is all possible products, if you take a set and use it to generate a group, then you take all possible products. You want to make sure the group axiom of closure holds. You take  $n$  powers of all elements in and  $n$  can include negative powers and inverses.

Like the other students, Isaac discussed repeated addition of ones, but connected this to the idea of powers. His intuitive process approach to generating the set of integers did not seem to interfere with his ability to leverage a more formal definition of generating. In fact, he leveraged this formal definition to validly explain the use of both one and its inverse to build the set.

While the proportion of students recognizing that  $\mathbb{Z}$  is cyclic was significantly higher in the replication studies than the original, the intuitive underpinnings of an iterative process to build the group were still present in the replication study. In a number of the interviews, students acknowledged they

knew  $\mathbb{Z}$  was cyclic because they had established the fact in class. To make sense of this, many of the students used their informal understanding to allow for sums of 1s and -1s to build the group. Their intuitive understandings might not be a hindrance, but rather have the potential for building a better understanding of powers of elements in groups. This informal understanding of generating can be reconciled with the formal definition through consideration to what it means to take an element to a negative power within a group. These preliminary results challenge the implications from the original study by showing that intuitive understanding might be leveraged rather than abandoned for a more formal understanding.

**Replication: Are  $\mathbb{Z}$ , the integers under addition, and  $\mathbb{Q}$ , the rationals under addition, isomorphic?**

Two groups are isomorphic if they are essentially the same. Formally, this means that there exists a bijective homomorphism from one group to the other. A bijective map is one that is 1-1 and onto. A map is a homomorphism if it preserves the operation- that is, if we have a map  $f$  from group  $G$  to  $H$ , for any  $a$  and  $b$  from  $G$ ,  $f(a)f(b)=f(ab)$ . Informally, the isomorphic groups are equivalent and can be thought of just renamings of one another. All elements can be matched up and the operation functions in the same manner in both groups. Any structural property differing between two groups would assure that they are not isomorphic.  $\mathbb{Z}$  and  $\mathbb{Q}$  are not isomorphic for a number of reasons including that  $\mathbb{Z}$  is cyclic and  $\mathbb{Q}$  is not. Another property that differentiates them is solutions to the equation  $x+x=b$ . In  $\mathbb{Q}$ , this equation will have a solution for any  $b$ , namely  $b/2$ . However,

in  $\mathbb{Z}$ , not all numbers can be divided in half and still remain in the set. For example,  $x+x=3$  would have no solution in  $\mathbb{Z}$ . See Figure 32 for a formal proof that this property is structural. Alternately, one could show that two are not isomorphic if one could argue that no maps can be created that meet the requirement of both being a bijection (1-1 and onto) and operation-preserving.

*Proof.* Suppose  $G$  and  $H$  are isomorphic groups. Let  $x+x=b$  have a solution for all  $b \in G$ . Because  $G$  and  $H$  are isomorphic, there exists an isomorphic map  $\phi$  from  $G$  to  $H$ . Let  $h \in H$ . Because  $\phi$  is onto, there exists a  $g \in G$  such that  $\phi(g) = h$ . By assumption, there exists an  $x_1 \in G$  such that  $x_1 + x_1 = g$ . Because  $G$  is a group,  $x_1 + x_1$  is in  $G$ . Its image is then in  $H$ . Using the fact that  $\phi$  is structure-preserving we can conclude:  $\phi(x_1 + x_1) = \phi(x_1) + \phi(x_1)$ . Thus,  $\phi(x_1) + \phi(x_1) = \phi(g) = h$  and so  $\phi(x_1)$  is the solution to  $x + x = h$ . As  $h$  was arbitrarily chosen,  $x + x = h$  has a solution for all  $h$  in  $H$ . □

Figure 32. Proof that having a solution for  $x+x=b$  for all elements  $b$  in a group is a structural property.

$\mathbb{Z}$  and  $\mathbb{Q}$  do have some commonalities that could lead to unfruitful explorations. They are both abelian, both have the same cardinality (countably infinite), and they are defined under the same operation (when the operation in  $\mathbb{Q}$  is restricted to  $\mathbb{Z}$ ). A bijection can be built between  $\mathbb{Z}$  and  $\mathbb{Q}$  because they have the same cardinality. A bijective map is illustrated in Figure 33 and Figure 34. Notice this map is not a homomorphism because the operation is not preserved (i.e.  $f(2)+f(3) = 2/1 + 1/2 = 5/2$ , but  $f(2+3) = f(5) = 4/1$ .)



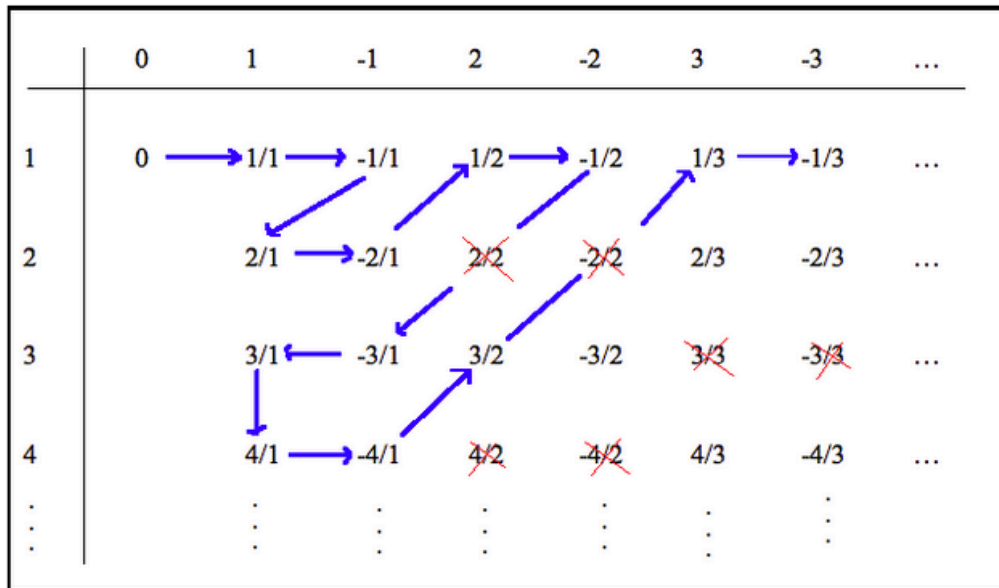


Figure 33. An ordering of the rationals to show they are countably infinite.

Bijjective Map  $f: \mathbb{Z} \rightarrow \mathbb{Q}$

$\mathbb{Z}$	0	1	-1	2	-2	3	-3	4	-4	5	...
$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$	
$\mathbb{Q}$	0	$1/1$	$-1/1$	$2/1$	$-2/1$	$1/2$	$-1/2$	$-3/1$	$3/1$	$4/1$	...

Figure 34. A bijjective map from  $\mathbb{Z}$  to  $\mathbb{Q}$ .

**Prior results.** This question finds its origin in two landmark papers related to student proving processes, namely Weber (2001) and Weber and Alcock (2004). Weber initially used this question when arguing that undergraduates often lacked the *strategic* knowledge needed to produce proofs. He provided data on eight tasks with four undergraduate and four doctoral students' responses. He found that even when the undergraduate students had the basic proof skills, and prerequisite knowledge, they frequently lacked the strategic knowledge to know when to use appropriate facts and strategies. When evaluating whether  $\mathbb{Z}$  and  $\mathbb{Q}$

are isomorphic, he found two of the undergraduate students had all the prerequisite knowledge, but could not produce a proof, while the other two lacked the necessary prerequisite knowledge. It is worth noting that the two students who had knowledge of the needed facts could produce a proof when reminded of them. (Factual knowledge was previously evaluated using a multiple-choice test.) In contrast, the four doctoral students correctly addressed the prompts with valid proofs.

Weber and Alcock (2004) revisited this prompt when introducing their framework on syntactic and semantic proof production. They found the four undergraduates in their study all relied on recreating a diagonalization argument (see *Figure 33*) to prove  $\mathbb{Z}$  and  $\mathbb{Q}$  are isomorphic. They noted that rather than attending to the operation-preserving aspect of an isomorphism, the undergraduate students relied exclusively on constructing a bijection between the two groups. In contrast, all of the doctoral students were easily able to show the groups were not isomorphic, referencing  $\mathbb{Z}$  being cyclic and  $\mathbb{Q}$  not being cyclic. This was a fact that all four undergraduates were aware of, but did not make use of when attending to this prompt. Weber and Alcock used the differing approaches to illustrate the difference between syntactic and semantic proof production. Where the undergraduates attempted to stay in one formal representation system, a syntactic approach, the doctoral students explored the groups  $\mathbb{Z}$  and  $\mathbb{Q}$  informally. This allowed them to use their conception of isomorphism as structural sameness to explore properties and potential issues in their respective structures. Table 24 summarizes Weber and Alcock's (2004) expert-novice study.

Table 24

<i>Weber &amp; Alcock's (2004) Results: Prove or Disprove: <math>\mathbb{Q}</math> is Isomorphic to <math>\mathbb{Z}</math></i>		
	Undergraduates n=4	Graduates n=4
Syntactic Proof Production (Building a bijective map)	100%	0%
Semantic Proof Production ( $\mathbb{Z}$ is cyclic, $\mathbb{Q}$ is not)	0%	100%

Weber and Alcock (2004) likened these approaches to several parallels in theory including proofs that explain and proofs that convince (Hanna, 1990), and instrumental and relational understanding (Skemp, 1976). They claim semantically produced proofs have both the power to convince and explain whereas syntactically produced proofs might just convince. By staying entirely in the formal system, students can create valid proofs through a series of logical manipulations that may or may not be fully connected to their conceptual understanding of the topics at hand. Similarly, they claim semantic proof productions involve relational understanding (the how and the why), whereas syntactic proofs can be produced with only an instrumental understanding (the how). Their case studies illustrated this phenomenon where the undergraduates had no informal understanding of isomorphism and the graduate students viewed isomorphism as “structurally the same.”

These expert-novice case studies illustrated a clear divide between undergraduate and doctoral cases. Furthermore, the case studies successfully identified two potential issues: syntactic vs. semantic proof production and a lack of strategic knowledge. By scaling up, I explore how generalizable these results are and whether other issues may contribute to incorrectly or correctly evaluating

or arguing about the prompt. These results also allowed me to re-explore the connections between (a) syntactic and semantic proof production and (b) the parallel theories of relational and instrumental understanding and proofs that convince and proofs that explain.

Beyond the studies connected to this question, there is some other work that deals with how students determine if two groups are isomorphic. Leron, Hazzan, and Zazkis (1995) explained some of the many complexities involved in understanding isomorphism. For example, they found that students often relied on *order type* to determine if two groups are isomorphic. The order type refers to the orders of the various elements in a group. So the order type of  $\mathbb{Z}_4$  would be 1, 2, 4, 4. This approach would only be valid to show two groups with different order type are not isomorphic. Leron, et al. attributed this confusion to one of three causes: neglecting all other properties for the simple and comfortable one, the fact that order type is sufficient for many small finite groups, or a confusion between a statement and its converse. Regardless of the reason, these students adapted a process that does not capture a complete view of isomorphism and only works in select cases.

Small-scale studies with convenience samples provided the needed impetus to develop theories that have been shown to be quite fruitful in explaining student thinking related to proving in advanced mathematics. However, little can be said in terms of how widespread particular issues are and how dependent they are on their given context.

**Replication methods.** In the replication studies, students were prompted to answer, “Is  $\mathbb{Z}$ , the integers under addition, and  $\mathbb{Q}$ , the rationals under addition, isomorphic?” In the case of the previous studies, the participants were interviewed with extended attention to this prompt. In the replication study, most students were surveyed only with the 30 students participating in follow-up interviews.

Between the open-ended round and closed-form round, one frequently presented student response was not included in the closed-form version. In the closed-form round,  $\mathbb{Z}$  being cyclic and  $\mathbb{Q}$  not being cyclic was not an option. This was done to mitigate for the fact that many students were unaware that  $\mathbb{Z}$  is cyclic. Instead, a less common structural property was presented: “ $x+x=b$  has a solution for all  $b$  in  $\mathbb{Q}$ , but does not for all  $b$  in  $\mathbb{Z}$ .” This is a property that can be explored semantically, but is less likely to be on a list of properties memorized from a textbook. In the closed-form rounds, I could explore the slightly altered version. Removing the response that was used by the experts in Weber and Alcock’s (2004) study raised a number of questions about the original theory.

### **Replication results.**

Table 25, Table 26, and Table 27 contain the results from the replication rounds. In these rounds, students were to determine whether  $\mathbb{Q}$  and  $\mathbb{Z}$  are isomorphic.

When this prompt was given to students across the country, a variety of responses appeared. While Weber and Alcock (2004) found that the four students they talked to worked syntactically by attempting to build a map between groups, this was not always the case with the larger sample. Under 30% of students attempted to build a bijection or stated that a bijection existed as part of their

reasoning in the replication rounds. Students who made an argument about the sizes of the groups or building a bijective map had one of two conceptions: a genuine belief equal-sized groups guarantees an isomorphism, or a belief that looking for the map is the correct next step once you know the size. The first is an issue of content where students have not developed a strong understanding of isomorphism and the necessity of checking that a map is a homomorphism. In the second category, the students likely lacked the strategic knowledge (as discussed in Weber, 2001), inappropriately relying on a syntactic approach rather than exploring the groups. However, because most of this data comes from surveys, it is not always possible to differentiate between the two cases.

Through replicating the study, I identified new ways students approached this prompt. Many students determined that  $\mathbb{Z}$  and  $\mathbb{Q}$  were not isomorphic because  $\mathbb{Z}$  had fewer elements. This is an issue that may be more related to cardinality than isomorphism. However, a bijective map is intimately tied to understanding of cardinality. Two groups have equal cardinality, or equal size, if a bijection can be created between them. The apparent size difference can likely be attributed to  $\mathbb{Z}$  being a proper subset of  $\mathbb{Q}$ . If a group is finite, a proper subgroup could not be of the same size. However, this not true in the infinite case. A simpler example would include  $\mathbb{Z}$  and  $2\mathbb{Z}$ . A bijection can be formed from  $\mathbb{Z}$  to  $2\mathbb{Z}$  where the map takes an element  $x$  and maps it to the element  $2x$ . This map is a bijection and a homomorphism, so  $\mathbb{Z}$  and  $2\mathbb{Z}$  are isomorphic. When using one group being a proper subgroup of another to state that groups are not isomorphic, they are

relying on an apparent difference that is, in fact, not a difference at all.

Table 25

*Open-Ended Survey Results: Are  $\mathbb{Z}$ , the Integers under Addition, and  $\mathbb{Q}$ , the Rationals under Addition, Isomorphic?*

	$n=349$
No, $\mathbb{Z}$ is cyclic, $\mathbb{Q}$ is not	17.8%
No, Other valid	1.3%
No, $\mathbb{Z}$ is smaller than $\mathbb{Q}$	21.8%
Yes, Bijection exists between them (same size)	22.2%
Yes, Other	11.9%

In the open-ended round, 17.8% of students determined that  $\mathbb{Z}$  and  $\mathbb{Q}$  are not isomorphic because  $\mathbb{Z}$  is cyclic and  $\mathbb{Q}$  is not. Follow-up interviews indicated students were taking a semantic approach that none of the undergraduates had used in Weber and Alcock's (2004) study. A number of students articulated notions of "structural sameness." The fact that many students were able to take this approach suggests that perhaps the gap between novice and expert is not as wide as indicated by the strict dichotomy in Weber and Alcock's study.

Table 26

*Closed-Form Pilot Results Are  $\mathbb{Z}$ , the Integers under Addition, and  $\mathbb{Q}$ , the Rationals under Addition, Isomorphic?*

	$n=87$
Yes, because there exists a bijective map between them.	27.6%
Yes, because they both have infinite order, have operation addition and are abelian.	5.8%
No, because $\mathbb{Z}$ is a proper subset of $\mathbb{Q}$ and so no bijection can exist between them.	18.4%
No, because $x+x=b$ has a solution for all $b \in \mathbb{Q}$ but $x+x=b$ does not have a solution for all $b \in \mathbb{Z}$	32.2%
None of the above	14.9%
I don't know	0%

Table 27

*Closed-Form Large Round Results: Are  $\mathbb{Z}$ , the Integers under Addition, and  $\mathbb{Q}$ , the Rationals under Addition, Isomorphic?*

	$n=376$
Yes, because there exists a bijective map between them.	21.0%
Yes, because they both have infinite order, have operation addition and are abelian.	12.0%
No, because $\mathbb{Z}$ is a proper subset of $\mathbb{Q}$ and so no bijection can exist between them.	31.6%
No, because $x+x=b$ has a solution for all $b \in \mathbb{Q}$ but $x+x=b$ does not have a solution for all $b \in \mathbb{Z}$	12.0%
No, but the above responses did not consider the necessary property.	23.4%

In the closed-form pilot round, 32.2% of students picked the correct response with 14.9% picking “none of the above.” The selection of the correct answer dropped noticeably in the larger round with 12.0% of students picking the correct response. This is likely an artifact of the differing characteristics of the sample. In the pilot round, highly selective institutions were significantly overrepresented with over half of the students coming from institutions that accept less than 25% of the applicants (see Table 14). The large-scale round was more representative of the typical breakdown of institutions in the United States. This difference illustrates the importance of exploring the impact of a given sample’s characteristics when making claims.

**Testing the relationship of the original study’s theory and other theories.** Weber and Alcock (2004) used their cases to illustrate the hugely impactful framework of semantic and syntactic proof production. By staying entirely in the formal system, the undergraduate students were not able to recognize that  $\mathbb{Q}$  and  $\mathbb{Z}$  were not isomorphic. Their graduate counterparts explored



the properties of the groups semantically and were able to quickly discover they were not isomorphic. Weber and Alcock further associated semantic proof production with proofs that explain and syntactic with those that convince. They also connected this theory to the parallel between instrumental and relational understanding.

In the replication study, preliminary results reflected that students may work semantically and successfully without necessarily having the robust understandings attributed to successful semantic proof production. I used interview excerpts to illustrate several potential areas of nuance in semantic approaches. I outline three cases: (1) students who appeared to have proceduralized their semantic approach; (2) students who worked semantically building on intuition in some instances but not others; and (3) students who appeared to have a similar rich semantic approach to the experts from the original study. In each case, students engaged with the task semantically, but had significant differences illustrating potential routes of theory refinement and exploration of the original theory's relationship to other theories.

Of the three interview students who stated that  $\mathbb{Z}$  and  $\mathbb{Q}$  were not isomorphic because one was cyclic and other was not, two attributed this approach to using a list provided by the textbook rather than any informal understanding of isomorphism. Tony explained that he thinks his answer is right:

[B]ut I'm not sure how I made that logic leap. I might have just read that in the book and noted it in my mind. Or at least noted in this text we were using. "Well, to prove something is isomorphic or not is actually kind of difficult, but here are some things you can look out for." One might have a single generator and I knew the integers did.

He was working semantically in that he was able to use an instantiation of  $\mathbb{Z}$  and  $\mathbb{Q}$  and determine that one had a single generator and the other did not. While this may on the surface look similar to the rich semantic approaches described by Weber and Alcock (2004), this student did not appear to have an understanding of why, but rather a proceduralized semantic approach. Tony was unable to connect his list of properties to his conception of isomorphism. This is similar to Leron, Hazzan, and Zazkis' (1996) study where they found students relied on a list of properties including order type and commutativity to determine if groups were isomorphic. Isomorphism is a complex topic, and so students may rely on procedural or syntactic approaches that may not address the underlying *why* behind evaluating if groups are isomorphic.

During follow-up interviews for the closed-round version, this same idea manifested. In this case, a student was unable to recognize the structural property that was not on their textbook list. He noted:

It didn't follow from anything I had done in class. Sorry, I'm referring to [the fourth choice]. It didn't follow from anything I studied in class. Like if it has a solution for  $b$  and this didn't have a solution for  $b$ , that wasn't something we studied for isomorphism.

When confronted with a property not available in class, he did not see the relevance to isomorphism. The proceduralized list checking may be limiting in terms of flexibly considering atypical structural properties.

During the closed-round interviews, one student, Molly, was able to connect cyclic groups to her informal understanding of isomorphism. Her explanation was a departure from the above as she articulated informal notions of

isomorphism that were cohesive with the cyclic property she addressed. Molly explained, “I had said ‘No’ because the integers are cyclic and the rationals aren’t cyclic because they don’t have a generator or a finite set of generators, I would think.” She also provided a standard formal definition for isomorphism. When asked how that related to a group being cyclic, Molly explained, “We’ve been thinking about isomorphism as the definition of the structure of a group. And a non-cyclic group cannot be isomorphic to a cyclic group.” She relied on an informal understanding, that of structural-sameness, to explain. She also explained that cyclic was a property used in class. However, when prompted to look at the property in the fourth option (that of solutions to  $x+x=b$ ), Molly did not think it was relevant declaring, “it just seemed like a random something.” Molly could link her understanding of isomorphism to a known property, but at this point did not flexibly address an unfamiliar property. This raises questions about whether a semantic approach guarantees a deep relational understanding. Additional research into this question could further explore the relationship between semantic approaches and a flexible relational understanding.

In two of the interviews, students did think flexibly about groups as structurally similar and recognized the new property presented in fourth option as structural. One of the students, Fionna, explained that when two groups are isomorphic:

There’s a bijection that preserves the operation of the group. Intuitively, it’s the same structure just described in a different way. So any property [referring to  $x+x=b$ ’s solutions] that holds in one structure would have to hold in the other one as well.

Fionna used what Weber and Alcock (2004) might have identified as similar to their expert approaches. She was able to attend to a new structural property (of  $x+x=b$ 's solutions) and determine that it would mean that the structure of  $\mathbb{Q}$  and  $\mathbb{Z}$  could not be the same.

By providing students with different response options, the replication studies unveiled some potential nuances between the original theory connections and how students engaged with the prompts. Students appeared to be able to prove semantically without necessarily having a deep relational understanding of isomorphism. There were also cases of students able to address familiar properties and connect them to their conceptual understanding, but did not successfully address the unfamiliar properties. Furthermore, in the replication study, several students used the expert approach that was not used by any undergraduates in the original study. The apparent differences in semantic approaches can provide impetus for additional studies that may elaborate on the previous framework.

Both the large-scale round results and the follow-up interviews illustrated that there were many different ways students may approach this prompt. The large-scale frequencies reflect the prevalence of certain response types amongst students at the introductory group theory level. The follow-up interviews provided starting grounds for exploring the generalizability of Weber and Alcock's (2004) work and highlighted some nuances within the theory. Notably, the relationship between semantic proving and rich relational understanding may not be as clear-cut as suggested by the original study.

## Conclusion

Pairing large-scale studies with interviews allowed for both holistic pictures and more detailed probing into the thinking of individual students. Rather than interviewing a few students, the large-scale results were used to select students with a range of conceptions to interview. The interviews served the dual role to bolster validity of interpreting student survey responses, and allow for deeper theoretical explorations. Replications allow for broader sampling of a population in order to evaluate the generalizability of any theory. For each of these cases, the replication studies unearthed differences in frequencies of various conceptions and provided additional insight into how a theory might be refined or elaborated. In any field, it is essential that we test the generalizability and repeatability of results in the field. Small-scale studies and studies using convenience samples can be very powerful in exploring student conceptions; however, they should not be the end point in an exploration.

The replication studies were valuable in a number of ways. First, replication served to validate theories by reproducing a number of the original results in all three questions. The validation was further bolstered because these samples were representative of the general population. The replication studies also served to document frequencies of different student conceptions within the greater population. Within the larger samples, a number of new conceptions emerged such as students struggling with the cardinality of  $\mathbb{Q}$  and  $\mathbb{Z}$ .

The replication studies also served to inform theory and raise new questions about generalizability and original interpretations. This took a number

of forms including exploring different pedagogical treatment of modular arithmetic groups, and challenging whether students truly are failing to coordinate operation and sets in the case of  $\mathbb{Z}_3$  being a subgroup of  $\mathbb{Z}_6$ . The replication studies also allowed for informing theory through unearthing some unconsidered nuances such as re-exploring how semantic and syntactic proof production aligns with other theories. Finally, replication studies can be used to re-examine implications from prior studies. This occurred when re-exploring the prompt, “Is  $\mathbb{Z}$ , the set of integers under addition, a cyclic group?” Students were able to appropriately deal with the infinite case while still maintaining their intuitive ideas of generating.

While replication studies are largely non-existent in education, their value is considerable. In the social sciences, much of theory and results stems from subjective interpretations greatly informed by the time, place, and participants in a given study. The results from these replications highlight just how important it is to re-examine prior results to consider their generalizability amongst the population of interest.

### **Paper 3: A tool for unearthing student conceptions in group theory**

The Group Concept Inventory (GCI) is a tool created to provide instructors, mathematics departments, and researchers a way to assess student understandings in a both a quick and meaningful way. The GCI incorporates knowledge of genuine student thinking about group theory concepts stemming from prior educational studies as well as validation studies conducted during its creation.

While there is a growing community of mathematics educational researchers studying student learning and teaching within the undergraduate contexts, their research largely does not filter down into the hands of the courses' instructors. This likely reflects a major difference in research methodology and goals. Mathematics education research has more commonalities to psychology than mathematics research. There is a divide between the mathematics and mathematics education communities that should be addressed in order to improve instruction within our mathematics courses.

This gap is being bridged in a number of ways in the sciences. One model of bringing education research to instructors in in the form of *concept inventories* and other similarly styled measures. These tools can be utilized for diagnostic purposes so that lessons learned from researching student thinking become immediately applicable in the classroom. In general, the instruments consist of a set of multiple-choice question set that is developed to target conceptual understanding over procedural fluency. Starting with Hestenes, Wells, and Swackhamer's (1992) *Force Concept Inventory (FCI)*, many assessment

instruments have been built within various STEM subject areas that incorporate genuine student conceptions about major ideas in various introductory courses.

The FCI is credited as an impetus for major changes in physics education as the measure unearthed a variety of incorrect intuitive understandings that could be hidden when students have strong procedural knowledge. Concept inventories can be used in classrooms to quickly gather information about how students are conceiving of topics. Concept inventories are not as widespread in mathematics courses with their existence limited to the Calculus Concept Inventory (CCI) (Epstein, 2006) and the Precalculus Concept Assessment (PCA) (Carlson, Oehrtman, & Engelke, 2010). No such instruments have been introduced for proof-based courses such as abstract algebra. The GCI aims to fill that gap in introductory group theory by providing a tool that instructors and departments can use to evaluate and make sense of student thinking.

In the last twenty years, there has been a small, but important, body of literature addressing some of the conceptions students have in introductory abstract algebra courses. For example, I suggest asking your students if  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ . While the question seems trivial (no, they either have different elements or a different operation depending on how  $\mathbb{Z}_n$  is defined in your course), students will frequently attend to different aspects than experienced mathematicians might assume.  $\mathbb{Z}_3$  is a group in itself. Its elements *look* like they belong in  $\mathbb{Z}_6$ . Dubinsky, Dautermann, Leron and Zazkis (1994) illustrated this phenomenon as part of larger study aiming to provide decompositions of how students think about fundamental concepts in group theory. This is a powerful



question that may reflect that students are not coordinating their operation and set in the ideal way. In fact, the question also unveiled that many students were misapplying Lagrange's Theorem to arrive at the same conclusion (Hazzan & Leron, 1996). In my recent work, I found that 37.8% of a sample of 384 students stated that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  with invalid reasoning.<sup>4</sup> Questions of this nature can provide powerful insight into student thinking around fundamental group theory concepts.

The GCI tool serves two purposes:

- (1) usage by mathematics departments and researchers to evaluate programs and instructional innovations;
- (2) usage by instructors to arrive at diagnostic information about their students' understandings

Abstract algebra is a notoriously challenging course. “[M]athematics faculty and students generally consider it to be one of the most troublesome undergraduate subjects” (Dubinsky, et al., 1994, p. 268). For many students, this course is the first time concepts are to be reasoned about formally based on their properties (Hazzan, 1999). Research on student understanding in group theory unanimously illustrates that students struggle with the underlying concepts in introductory group theory.

### **Why Should We Care about Student Understanding of Concepts in Proof-Focused Courses?**

The aim of the GCI is to assess conceptual understanding independent of proof skills. While strong proof skills are essential in advanced courses, well-

---

<sup>4</sup> Note: Some students with an advanced conception of isomorphism correctly identified  $\{0,2,4\}$  as a copy of  $\mathbb{Z}_3$  in  $\mathbb{Z}_6$ . Those students are not included in this percentage.

developed understanding of concepts is also vital for continued success in mathematics. Proof is often the focus of assessment in advanced mathematics, ignoring the role of strong conceptual understanding. Attending to concepts is important for several reasons:

1. Understanding concepts in group theory is not trivial.
2. Understanding concepts is an essential part of advanced mathematics.
3. Understanding concepts is a necessary for a high degree of success in proving.

The small body of literature related to abstract algebra documents the difficulty of complete conceptual understanding of the various topics. Students must reason about objects that are brought into existence through definition rather than describing known objects. Dubinsky (1997) and his colleagues established that students frequently lacked advanced conceptions of topics in group theory ranging from groups to normality and quotient groups. Hazzan (1999) found that students struggled with abstraction level when dealing with objects in algebra defined solely by their properties. Students might substitute information about familiar groups such as the real numbers, to reason about unfamiliar groups such as modular groups. For example, students may determine that 2 does not have an inverse in  $\mathbb{Z}_3$  because they checked for the element  $1/2$ . Leron, Hazzan and Zazkis (1995) documented the struggles students have when coordinating the complexity of isomorphism, such as differentiating between the idea of isomorphic groups and an isomorphism map, or differentiating between how to show groups *are* isomorphic versus how to show groups *are not* isomorphic. The complexity,

formality, and abstract nature of the concepts in group theory provide a significant hurdle for students to overcome.

Concepts play a vital role both in the formal proving process and independent from it. While formal proof may dominate assessment at the advanced level, it is not the only important activity. Hanna (1999) warned, “competence in mathematics might readily be misperceived as synonymous with the ability to create the form, a rigorous proof” (p. 60). In fact, as mathematicians, a lot of the work done is related to bigger conceptual understanding rather than just rote formal proofs. Weber and Mejia-Ramos (2011) recently confirmed that when reading proofs, mathematicians rarely verified they formal deductive steps. Rather, they were tested examples, determined the big ideas and evaluated methods for their own use. These activities all require an understanding of concepts, instantiations of them (such as tables, diagrams, and examples), and relationships between ideas.

Even if the focus of a course is entirely on producing formal proof, conceptual understanding can be a mitigating and necessary component in the creation of valid proofs (Tall & Vinner 1981; Moore, 1994; Hart, 1994). Through a series of interviews with both students and mathematicians, Weber and Alcock (2004) found that undergraduates often relied on purely syntactic (symbol-pushing) proving strategies. As a result, they were limited with their approaches to proofs and what statements they could prove. In contrast, advanced graduate students would often utilize semantic strategies, that is, leaving the formal system and reasoning with instantiations. When prompted to assess whether groups ( $\mathbb{Q}$

and  $\mathbb{Z}$ ) were isomorphic, undergraduates were limited to trying to create a map and so failed to show groups were not isomorphic. Their graduate counterparts were able to reason about properties and easily determine the two groups were not isomorphic. Weber and Alcock identified several requirements for successful proving related to concept understanding including instantiating rich and accurate reflections of “the object and concepts they represent” (p. 229). Further students need to be able to connect the formal definitions with instantiations. Having the formal definition of group and isomorphism was not a sufficient condition for students to be able to describe either informally or utilize intuition to explore conjectures.

Moore (1994) explained several ways that understanding of concepts becomes important when proving. He gathered data from observing a transition to proof course and interviewing students from the class. He found the students had a multitude of concept-based errors. These include not being able to generate examples, lacking an intuitive understanding, not being able to use concept images to create formal proofs, and not knowing the appropriate definitions.

Concepts play an important role in advanced mathematics. As objects are now defined solely on properties, students often struggle to achieve complete understanding of concepts. Mathematical activities, such as using examples to make sense of formal statements or test conjectures, often rely on underlying conceptual understand. Having syntactic strategies alone has been shown to be insufficient in many cases. The GCI can provide complementary information

about student understanding of concepts that may explain their success or lack of success when producing and evaluating proofs beyond just logic skills.

### **How was the GCI Created?**

The tool was created using a rigorous methodology to assure both validity and reliability. Further, each multiple-choice option reflects an alternate conception about the topic from genuine student responses. At this point, I must caution that this tool was not created for students to be given a grade, but rather a tool for the class level and to provide quick formative feedback for mathematics instructors.

The first step in creating such an instrument was to identify the fundamental topics in an introductory group theory. This was done through a Delphi Study (Dalkey & Helmer, 1963), a protocol for expert consensus, amongst a panel of group theorists, non-group theorists mathematicians, textbook authors and mathematics educators who have published pedagogy articles related to group theory. The final list of topics included: *Associative Property, Inverses, Identity, Binary Operation, Group, Isomorphism, Homomorphism, Lagrange's Theorem, The First Isomorphism Theorem, Cyclic Groups, Abelian Groups, Cosets, Quotient Groups, Order of an Element, Order of a Group and Modular Arithmetic*. These topics represent a significant subset of the topics recommended by the Committee on Undergraduate Programs in Mathematics (Isaacs, Bahls, Judson, Pollastek, & White, 2015). It is impossible for any assessment to cover

the entirety of a domain, and so the GCI was created such that each of these concepts were directly reflected in one question.

Using this list of concepts, textbooks were analyzed as a tool to explore curriculum across the country. While the textbooks do not perfectly reflect what happens in classes, they do give insight into general trends across courses. Within introductory abstract algebra classes, the most common textbooks included: Gallian (2009), Fraleigh (2002), Gilbert and Gilbert (2008), and Hungerford (2012). These books were identified through a random survey of 294 institutions with mathematics majors. The textbook analysis gave insight into what types of examples students uniformly had access to, what formal and informal definitions were provided and what type of exercises and activities were typical and valued. Using the results of this analysis and leveraging what currently exists in literature about student understanding in group theory, a set of 42 open-ended questions were created. At least two mathematicians who had experience teaching the course and two mathematics educators who had published articles related to group theory pedagogy evaluated every task. A subset of 18 tasks was then given to 383 students who were completing the group theory component of an introductory algebra course. Each of the student responses were analyzed to identify the most common conceptions, correct and incorrect, related to each question. These responses became the options for a multiple-choice version. Methods from classical test theory were utilized to refine the questions to maximize reliability. At each stage, 15 students were interviewed to bolster validity that the responses were being chosen for the reasons hypothesized. Further, the tool was correlated

with grades (self-reported) to achieve criterion-related validity (See Paper 1).

Figure 35 includes an overview of the GCI creation process.

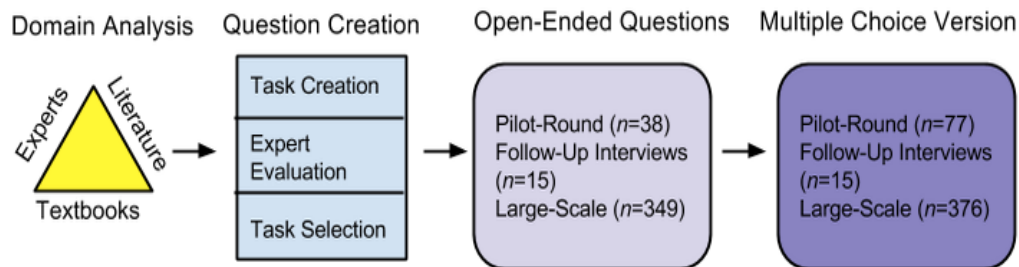


Figure 35. Overview of GCI creation stages.

### How Can the GCI be Utilized?

The GCI can be a powerful tool for both individual instructors and departments. At the class level, the tool can be used to evaluate the impact of instructional innovations. There are a number of innovative curriculums that exist in abstract algebra geared towards helping students develop a better understanding of group theory topics (Larsen, Johnson, & Weber, 2013; Dubinsky, 1997).

Departments can use such tools to collect data over time to evaluate a class's impact on students' conceptual understanding. In the *Mathematical Association of America's* study of successful undergraduate mathematics programs, they recommended, "A mathematics department that is trying to take stock of its undergraduate program is well advised to undertake a systematic program assessment" (Tucker, 1995, p. 18). One way this can be supported is through the usage of concept inventories. In fact, an institution recently identified as a model for a highly successful calculus program, was notable for utilizing the CCI (Epstein, 2006) to evaluate student learning as pedagogical changes were being made (Larsen, Glover, & Melhuish, 2015). The CCI supported their aims to

focus on data about student learning and allowed them to leverage that information and provide evidence that their pedagogical decisions improved student understanding of calculus concepts.

As recommended in CUPM's guidelines for assessment, assessment of student learning can provide powerful feedback on student learning and provide data for a department, but, "[a]ssessment should be used to improve teaching and learning for all students, not to filter students out of educational opportunities" (Anderson, Marion, & Maki, 2006, p. 232). A concept inventory directly ties to these goals as the questions are not developed to identify right and wrong answers, but unearth incorrect and incomplete conceptions that might otherwise be hidden.

This diagnostic usage is even more powerful for individual instructors. There has been an overarching call for formative assessment in classrooms especially in undergraduate mathematics and science classes where lecture is the dominant form of pedagogy. Instructors can quickly gauge incorrect conceptions that might exist in their classes by incorporating questions that are fast and provide immediate feedback. With the rise of *clicker* technology, there is a need for multiple-choice questions that will provide genuine information about students' thinking (Caldwell, 2007). Caldwell elaborates that strong questions can offer a powerful formative assessment, providing opportunities to capture student understandings of certain material, reveal potential misunderstandings, inform decisions around future lectures, and provide students with a way to "assess their



own level of understanding” (p. 11). By incorporating such questions, students have the opportunity to reassess their own levels of understanding.

With its widespread usage, the FCI provides an image of the type of impact concept inventories and their questions can have on pedagogy. Concept inventories have the power to evaluate students in the moment, but also more generally influence instruction of the course. Instructors have used the FCI itself to purposefully teacher content with the intention of confronting incorrect conceptions and helping develop correct intuition around concepts (Savinainen & Scott, 2002b; Savinainen & Viiri, 2008). Similarly, the GCI is an artifact of known student conceptions and can shape the instruction of a course to confront and address potential issues before they arise. (See Paper 1 for a discussion on pedagogical power of building on student intuition around cyclic groups.)

### **Sample Items**

The following sample items illustrate the utility of the concept inventory. Every question in the GCI began as open-ended questions. Through extensive surveying and interviewing students, genuine student responses became the options for the multiple-choice version.

Returning to our prior example, the question of whether  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$  eventually became the multiple choice question seen in Figure 36. The question was altered to ask specifically about the subset  $\{0,1,2\}$  in order to assure that students are considering the same set. (Asking about  $\mathbb{Z}_3$  allowed for the freedom to address the isomorphic copy  $\{0,2,4\}$ .)

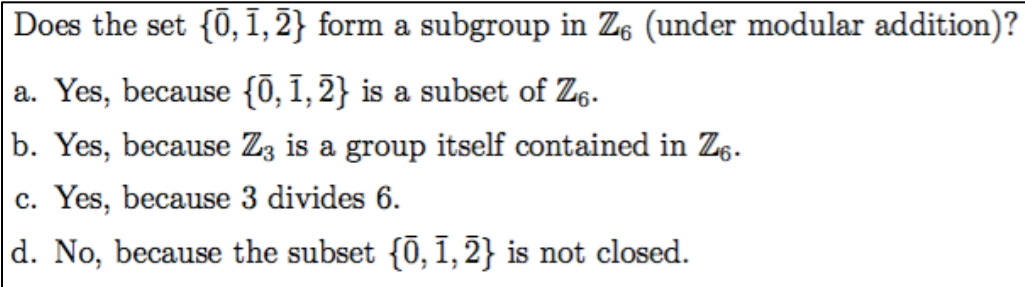


Figure 36. Question on Subgroups

Each of these responses reflects a different conception related to subgroups. The first response attends only to the subset aspect of the set without attention to the operation. The second response alters the binary operation from addition modulo 6 to addition modulo 3. This may reflect that students are not properly attending to operation and what it means to restrict an operation to a subset. The third option likely reflects a surface level understanding and misapplication of Lagrange's Theorem. Finally, a student is likely to pick the forth response if they are appropriately coordinating the operation and set attributes of subgroup.

This question has been discussed in a number of studies. If you are interested in learning more about how students may conceive of this question, Dubinsky, et al. (1994), provide theory about the mental constructions needed to make sense of subgroups with illustrations from students answering this prompt. Hazzan and Leron (1996) provide a discussion of students' misapplying Lagrange's Theorem to this problem. Further, discussion about results from both the open-ended and closed-form GCI versions can be found in Paper 1.

While many questions in the inventory have been explicitly studied previously, other items were newly developed with the intention of probing

known incorrect or incomplete understandings of a concept. The question focusing on the quotient group was developed in this manner. (See Figure 37.)

Let  $G = \mathbb{Z}_{12}$  (under addition) with subgroup  $H = \{\bar{0}, \bar{4}, \bar{8}\}$ . Consider the quotient (factor) group  $\frac{G}{H}$ . What is the order of  $\bar{2} + H$  in  $\frac{G}{H}$ ?

- The element has order 2.
- The element has order 3.
- The element has order 4.
- The element has order 6.

Figure 37. Question of Quotient Groups

This questions aims to address known issues of coordinating elements in quotient groups. These elements are particularly challenging as cosets are simultaneously sets themselves and elements of quotient group. Further, cosets are often explored using representative elements, which may obscure their set structure. To successfully answer this question, a student must make sense of the coset  $2+H$  as both an element and a set. The first response captures this coordination. They are able to recognize that  $2+H$  represents an element with an order in the group  $G/H$ . Further, they can recognize that  $(2+H)^2=H$ . This involves making sense of the identity element as the set  $\{0,4,8\}$  in the quotient group. If a student selects order 3, they are likely attending to the set part of coset. Despite the intentional use of term “order of element,” the student is attending to the cardinality of the coset without attending to its role in the quotient group. If the student selects 4, then the student is likely identifying the number of elements in  $G/H$ . Students selecting this response are likely applying a memorized fact in a superficial manner. Rather than determining the order of the element, they are providing the index. The final response likely reflects attention only to

representative elements. Instead of seeing 2 as a member of a coset and the identity as  $H$ , they are seeing 2 as an element, and not a representative, finding the  $n$  such that  $2^n = 0$  in the original group structure  $G$ .

The complexities of understanding quotient group elements have been explored in several studies. If you are interested in learning more about why this issue may be occurring, several studies address student conceptions of quotient groups including Asiala, Dubinsky, Mathews, Morics, and Oktac (1997) who have outlined the mental constructions students may need to build to understand quotient groups and illustrated students at various points in this process. Further, Larsen and Lockwood (2013) recently provided a local instructional theory and task sequence that allows for students to reinvent quotient groups from a more intuitive starting space: odds and evens.

### **Companion Website**

In order to make the GCI of the most use, the instrument has a companion website that includes explanations of the various questions, links to mathematics education articles that may further illustrate some of the conceptions captured, and links to alternate curriculums that may help students better come to understand these concepts. (See Figure 38 for a sample website page.) If you have any resources you would like to add or if you are interested in your class using the inventory, please contact the author at the provided email address.



meant to provide grades, but rather aid instructors and departments in addressing their students' learning. The nature of the questions makes them quick and easy to be use in a diagnostic capacity. Furthermore, the GCI provides a validated tool for departments to track to student learning and test the effectiveness of instructional innovations within introductory group theory courses. It can provide complementary information to proof production and has the power to highlight underlying conceptual issues that may be interfering as students attempt to grow in the formal mathematical world.

## **Conclusion**

The GCI meets a major need in the field of research on student understanding in group theory: a validated instrument that can be easily administered to large samples of students. This allows for evaluating instructional innovations and exploring the general state of understanding amongst the population of group theory students.

In Paper 1, I introduced an assessment design methodology as a new model for incorporating literature, curriculum and experts in order to best create questions in an area without an extensive literature base on student conceptions. By leveraging multiple sources, using student responses from open-ended versions for the multiple-choice distractors, and pairing field-testing with student interviews, I attended to various forms of validity. The most important aspects of a strong assessment include content and construct validity. The items should be relevant to the domain being tested. The items should be important to the domain being tested. Furthermore, the items should accurately capture student conceptions related to domain. Attending to both of these makes the results for the GCI meaningful, both at the item level and at the test level.

Through both the open-ended and closed-form rounds, the creation process also allowed for continued exploration of student thinking related to introductory group theory. In Paper 2, I explored how my large-scale results compared to smaller studies done on three questions that were replicated from earlier studies and became part of the GCI. Larger samples unearthed a number of new conceptions such as building infinite cyclic groups from both an element and

its inverse. Further, they validated many of the claims about student thinking in the smaller studies such as the prevalence of thinking that  $\mathbb{Z}_3$  is a subgroup of  $\mathbb{Z}_6$ . The larger samples provided a better sense for the frequencies of different conceptions. Paired with student interviews, I analyzed how these new responses could inform the original theory. For example, I found that semantic and syntactic proof production may not have as clear of a relationship between relational and instrumental understanding as found in the original study. These results provide additional information about how students are thinking about concepts in these courses.

By writing a complementary practitioner paper, I am attempting to bridge the gap between researchers and instructors of the course. We have learned a lot about how students conceive of various group theory topics, but often these theories remain in journals that are typically only read by fellow mathematics education researchers. By writing an article for a commonly read mathematics journal, mathematicians can connect to some of the research on student conceptions.

The GCI is not just a research tool, but also ideally serves a dual purpose for instructors to better connect to the conceptions their students have about group theory topics. By carefully linking student conceptions to the multiple-choice items, and using expert instructors to evaluate the importance and relevance of items, the inventory is both meaningful and useful.



## Directions For Future Research

In its current state, the GCI consists of a set of seventeen questions that have gone through a careful creation and refinement process. However, the GCI should be further refined through both continued field-testing and leveraging experts in the field.

The reliability estimates are adequate, but ideally could be raised. Some ways to increase reliability include exploring individual items that are not correlated strongly with the overall test score. For example, the isomorphism question is currently correlated with the item-corrected score at .051. The item-correct score means the correlation between student's item score (1 or 0) and the average score of the remaining 16 questions. This is a measure of item discrimination. While positive, the correlation is quite low. One potential explanation is that the question is too complex. While all of the distractors are related to isomorphism, they may inappropriately condense a number of different issues that may emerge. (See Paper 2.) Rather than a single question, this question may benefit from being split into two questions. One question can target issues with the cardinality of infinite groups. One question can target a flexible understanding of structural properties.

*Isomorphism Question A:* Does an isomorphism exist between  $\mathbb{Z}$ , the integers under addition, and  $4\mathbb{Z}$ , the multiples of 4, under addition?

Hypothetical answers:

- No, because  $4\mathbb{Z}$  is a proper subgroup of  $\mathbb{Z}$ , and therefore no bijection can exist between them.
- No, because the map  $f(x)=(4x+2)$  from  $\mathbb{Z}$  to  $4\mathbb{Z}$  is a bijection, but does not preserve the operation.
- Yes, because  $\mathbb{Z}$  and  $4\mathbb{Z}$  have the same cardinality.

- Yes, because there exists a map between  $\mathbb{Z}$  and  $4\mathbb{Z}$  that is bijective and preserves the operation.

*Isomorphism Question B:* Consider the following student argument about  $\mathbb{Q}$ , the rationals under addition, and  $\mathbb{Z}$ , the integers under addition:

“ $\mathbb{Q}$  and  $\mathbb{Z}$  are not isomorphic because the equation  $x+x=b$  has a solution for  $x$  for any  $b$  you pick in  $\mathbb{Q}$ , but that’s not true in  $\mathbb{Z}$ .” Do you agree or disagree with this reasoning?

Hypothetical answers:

- I disagree because this equation is not relevant to whether  $\mathbb{Q}$  and  $\mathbb{Z}$  are isomorphic.
- I disagree because  $\mathbb{Q}$  and  $\mathbb{Z}$  are isomorphic.
- I agree because this argument means  $\mathbb{Q}$  and  $\mathbb{Z}$  have different structures.

The complexity of isomorphism is well-documented (Leron, Hazzan, & Zazkis, 1996). By splitting this question, both the intuitive and formal ideas surrounding isomorphism can be targeted without any exclusion to one another.

There are also three questions where “none of the above” has not been eliminated as answer. The number of students selecting the choice remains above 5%. These questions may need some refinements to their wording such as changing the set used in the identity question. (See Paper 1 for an elaboration of this idea.) The next round will likely be a hybrid of multiple choice and open-ended to allow for the testing of new questions (such as the one above) and to allow students to explicate their reasoning for selecting “none of the above.” In this way, the option can be changed to best reflect student reasoning.

Besides question refinement, more complex psychometric techniques may be better equipped to handling this data set. Lower reliability may reflect too much guessing. This is going to be a natural issue with a low consequence test. If possible, the question set should be used to some degree of consequence such as

extra credit based on actual correctness, or as a small class quiz with some value. Other than changing the testing consequences, using item response models that allow for guessing may be more appropriate. However, this would require additional data and powerful software. While a viable option in the future, the limitation on numbers may be insurmountable. Generally, only mathematics majors take introductory group theory and the average class size at each round has hovered around 11 students. The number of classes needed to arrive at a dataset large enough for that complex of item modeling may not be feasible.

The next stage in GCI development would also include a return to experts. While an argument has been established that the current items are important, they should also be reevaluated in terms of representing the domain. In the model of the CAOS test, experts could evaluate the items and determine if any fundamental concepts are not included. Potentially more items might be created to fill in gaps. Several concepts were just under criteria for importance including equivalence classes and permutations. By having a panel of experts (experienced group theory instructors) evaluate the question set's representativeness, new questions may be formulated.

The creation of a concept inventory is an iterative process that involves continued attention to the validity and reliability of the instrument. The current version of the GCI has undergone a strong validation process with careful attention to generalizability. However, over the next years, the GCI should continue to be field tested and evaluated in order to best meet the needs for a validated instrument.

## References

- Almeida, D. F. (1999). Visual aspects of understanding group theory. *International Journal of Mathematical Education in Science and Technology*, 30(2), 159–166.
- Andersen, J., Marion, W., & Maki, D. (2006). CUPM Guidelines for Assessment of Student Learning. *Supporting Assessment in Undergraduate Mathematics*, 229.
- Anderson, J., Austin, K., Barnard, T., & Jagger, J. (1998). Do third-year mathematics undergraduates know what they are supposed to know? *International Journal of Mathematical Education in Science and Technology*, 29(3), 401–420.
- Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Allyn & Bacon.
- Arnon, I., Cottrill, J., Dubinsky, E., Oktac, A., Fuentes, S. R., Trigueros, M., & Weller, K. (2013). *Apos Theory: A framework for research and curriculum development in mathematics education*. Springer London, Limited.
- Asiala, M., Dubinsky, E., Mathews, D. M., Morics, S., & Oktac, A. (1997). Development of students' understanding of cosets, normality, and quotient groups. *The Journal of Mathematical Behavior*, 16(3), 241–309.
- Asiala, M., Kleiman, J., Brown, A., & Mathews, D. (1998). The development of students' understanding of permutations and symmetries. *International Journal of Computers for Mathematical Learning*, 3(1), 13–43. doi:10.1023/A:1009738109343
- Biehler, R. (2005). Reconstruction of meaning as a didactical task: the concept of function as an example. In *Meaning in Mathematics education* (pp. 61–81). Springer US.
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101.
- Brenton, L., & Edwards, T. G. (2003). Sets of sets: a cognitive obstacle. *College Mathematics Journal*, 31–38.
- Brown, A., DeVries, D. J., Dubinsky, E., & Thomas, K. (1997). Learning binary operations, groups, and subgroups. *The Journal of Mathematical Behavior*, 16(3), 187–239.
- Burman, L. E., Reed, W. R., & Alm, J. (2010). A call for replication studies. *Public Finance Review*, 38(6), 787–793.
- Burn, B. (1996). What are the fundamental concepts of group theory? *Educational Studies in Mathematics*, 31(4), 371–377.
- Caldwell, J. E. (2007). Clickers in the large classroom: Current research and best-practice tips. *CBE-Life Sciences Education*, 6(1), 9–20.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, 78(1), 98.

- Carlson, M., Oehrtman, M., & Engelke, N. (2010). The precalculus concept assessment: A tool for assessing students' reasoning abilities and understandings. *Cognition and Instruction*, 28(2), 113-145.
- Cnop, I., & Grandsard, F. (1998). Teaching abstract algebra concepts using small group instruction. *International Journal of Mathematical Education in Science and Technology*, 29(6), 843-850.
- Cook, J. P. (2014). The emergence of algebraic structure: students come to understand units and zero-divisors. *International Journal of Mathematical Education in Science and Technology*, 0(0), 1-11.
- Dalkey, N., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management science*, 9(3), 458-467.
- Delmas, R., Garfield, J., Ooms, A., & Chance, B. (2007). Assessing students' conceptual understanding after a first course in statistics. *Statistics Education Research Journal*, 6(2), 28-58.
- Devellis, D. R. F. (2011). *Scale Development: Theory and Applications*. SAGE Publications.
- Docktor, J., Heller, K., Henderson, C., Sabella, M., & Hsu, L. (2008). Gender differences in both force concept inventory and introductory physics performance. In *AIP Conference Proceedings* (Vol. 1064, p. 15). Retrieved from [http://groups.physics.umn.edu/physed/Talks/Docktor\\_Heller\\_PERC08.pdf](http://groups.physics.umn.edu/physed/Talks/Docktor_Heller_PERC08.pdf)
- Dubinsky, E. (Ed.) (1997). An investigation of students' understanding of abstract algebra (binary operation, groups and subgroups) and The use of abstract structures to build other structures (through cosets, normality and quotient groups.) [Special Issue]. *The Journal of Mathematical Behavior*, 16(3).
- Dubinsky, E., & Yiparaki, O. (2000). On student understanding of AE and EA quantification. *CBMS Issues in Mathematics Education*, 8, 239-289.
- Dubinsky, E., Dautermann, J., Leron, U., & Zazkis, R. (1994). On learning fundamental concepts of group theory. *Educational Studies in Mathematics*, 27(3), 267-305.
- Dubinsky, E., Dautermann, J., Leron, U., & Zazkis, R. (1997). A Reaction to Burn's "What Are The Fundamental Concepts of Group Theory?". *Educational Studies in Mathematics*, 34(3), 249-253.
- Dubinsky, E., & McDonald, M. A. (2002). APOS: A constructivist theory of learning in undergraduate mathematics education research. In *The teaching and learning of mathematics at university level* (pp. 275-282). Springer Netherlands.
- Eastman, P. M. (1975). Replication Studies: Why so Few?. *Journal for Research in Mathematics Education*, 6(2), 67-68.
- Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, N.J.: Prentice Hall.
- Edwards, B. S., & Ward, M. B. (2004). Surprises from mathematics education research: Student (mis) use of mathematical definitions. *American Mathematical Monthly*, 411-424.

- Epstein, J. (2006). The calculus concept inventory. *National STEM Assessment, Washington, DC*, 60-67.
- Epstein, J. (2007, September). Development and validation of the Calculus Concept Inventory. In *Proceedings of the ninth international conference on mathematics education in a global community* (Vol. 9, pp. 165-170). Charlotte, NC.
- Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and psychological measurement*, 58(3), 357-381.
- Findell, B. (2002). The operation table as metaphor in learning abstract algebra. In *Proceedings of the twenty-fourth annual meeting of the North American chapter of the international group for the psychology of mathematics education Athens, Georgia* (pp. 233-245).
- Fraleigh, J. B. (2002). *A First Course in Abstract Algebra, 7th Edition* (7 edition.). Boston: Pearson.
- Fukawa-Connelly, T. (2012). Classroom sociomathematical norms for proof presentation in undergraduate in abstract algebra. *The Journal of Mathematical Behavior*, 31(3), 401-416.
- Fukawa-Connelly, T. P., & Newton, C. (2014). Analyzing the teaching of advanced mathematics courses via the enacted example space. *Educational Studies in Mathematics*, 87(3), 323-349.
- Gallian, J. (2009). *Contemporary Abstract Algebra* (8th edition.). Boston, MA: Cengage Learning.
- George, D., & Mallery, P. (2003). SPSS for Windows step by step: A simple guide and reference. 11.0 update (4<sup>th</sup> ed.). Boston: Allyn & Bacon.
- Gilbert, L., & Gilbert, J. (2008). *Elements of Modern Algebra*. Cengage Learning.
- Glaser, R., Chudowsky, N., & Pellegrino, J. W. (Eds.). (2001). *Knowing What Students Know:: The Science and Design of Educational Assessment*. National Academies Press.
- Goldin, G. A. (2002). Representation in mathematical learning and problem solving. *Handbook of International Research in Mathematics Education*, 197-218.
- Grassl, R., & Mingus, T. T. Y. (2007). Team teaching and cooperative groups in abstract algebra: nurturing a new generation of confident mathematics teachers. *International Journal of Mathematical Education in Science and Technology*, 38(5), 581-597.
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. New York, NY: Routledge.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309-333.
- Hambleton, R. K., & Jones, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.

- Hanna, G. (1990). Some pedagogical aspects of proof. *Interchange* 21.1 (1990): 6-13.
- Hanna, G. (1991). Mathematical proof. In *Advanced mathematical thinking* (pp. 54-61). Springer Netherlands.
- Harel, G., & Kaput, J. (1991). The role of conceptual entities and their symbols in building advanced mathematical concepts. *Advanced Mathematical Thinking*, 82-94.
- Hart, E. W. (1994). A conceptual analysis of the proof-writing performance of expert and novice students in elementary group theory. *MAA NOTES*, 49-62.
- Hazzan, O. (1999). Reducing abstraction level when learning abstract algebra concepts. *Educational Studies in Mathematics*, 40(1), 71-90.
- Hazzan, O. (2001). Reducing abstraction: The case of constructing an operation table for a group. *The Journal of Mathematical Behavior*, 20(2), 163-172.
- Hazzan, O., & Leron, U. (1996). Students' use and misuse of mathematical theorems: The case of Lagrange's theorem. *For the Learning of Mathematics*, 23-26.
- Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory. *The Physics Teacher*, 33(503), 507-511.
- Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory. *The Physics Teacher*, 33(8), 502-506.
- Hestenes, D., & Wells, M. (1992). A mechanics baseline test. *The Physics Teacher*, 30(3), 159-166.
- Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *The Physics Teacher*, 30, 141.
- Hiebert, J., & Lefevre, P. (1986). Conceptual and procedural knowledge for teaching on student achievement. *Hiebert, J. Ed. Conceptual and procedural knowledge: the case of mathematics*. Hillsdale, NJ: Erlbaum, 1-27.
- Hoellwarth, C., Moelter, M. J., & Knight, R. D. (2005). A direct comparison of conceptual learning and problem solving ability in traditional and studio style classrooms. *American Journal of Physics*, 73(5), 459-462.
- Hsu, C.-C., & Sandford, B. A. (2007). The Delphi technique: making sense of consensus. *Practical Assessment, Research & Evaluation*, 12(10), 1-8.
- Hungerford, T. (2012). *Abstract algebra: an introduction*. Cengage Learning.
- Inglis, M. (2015) Measuring conceptual understanding quickly and reliably. Presented at a plenary session of the *16th Conference on Research in Undergraduate Mathematics Education*. Pittsburgh, Pennsylvania.
- Isaacs, M., Bahls, P., Judson, T., Pollastek, H., & White, D. (2015). *Report from working group on Abstract Algebra*. Retrieved from [http://www2.kenyon.edu/Depts/Math/schumacherc/public\\_html/Professional/CUPM/2015Guide/Course%20Groups/abstractalgebra.pdf](http://www2.kenyon.edu/Depts/Math/schumacherc/public_html/Professional/CUPM/2015Guide/Course%20Groups/abstractalgebra.pdf)
- Jaworski, B. (2002). Sensitivity and challenge in university mathematics tutorial teaching. *Educational Studies in Mathematics*, 51(1-2), 71-94.

- Johnson, E. (2013). Implications of realistic mathematics education for analyzing student learning. Presented at the Conference on research in undergraduate mathematics education, Denver, CO. Retrieved from [http://pzacad.pitzer.edu/~dbachman/RUME\\_XVI\\_Linked\\_Schedule/rume16\\_submission\\_106.pdf](http://pzacad.pitzer.edu/~dbachman/RUME_XVI_Linked_Schedule/rume16_submission_106.pdf)
- Johnson, E., Caughman, J., Fredericks, J., & Gibson, L. (2013). Implementing inquiry-oriented curriculum: From the mathematicians' perspective. *The Journal of Mathematical Behavior*, 32(4), 743-760.
- Johnson, E. M. S., & Larsen, S. P. (2012). Teacher listening: The role of knowledge of content and students. *The Journal of Mathematical Behavior*, 31(1), 117-129.
- Kaput, J. J. (1998). Representations, inscriptions, descriptions and learning: A kaleidoscope of windows. *The Journal of Mathematical Behavior*, 17(2), 265-281.
- Kleiner, I. (2007). *A history of abstract algebra*. New York, NY: Springer Science & Business Media.
- Lajoie, C., & Mura, R. (2000). What's in a name? A learning difficulty in connection with cyclic groups. *For the learning of Mathematics*, 29-33.
- Larsen, S. (2009). Reinventing the concepts of group and isomorphism: The case of Jessica and Sandra. *The Journal of Mathematical Behavior*, 28(2-3), 119-137.
- Larsen, S. (2010). Struggling to disentangle the associative and commutative properties. *For the Learning of Mathematics*, 37-42.
- Larsen, S., & Lockwood, E. (2013). A local instructional theory for the guided reinvention of the quotient group concept. *The Journal of Mathematical Behavior*, 32(4), 726-742.
- Larsen, S., & Zandieh, M. (2008). Proofs and refutations in the undergraduate mathematics classroom. *Educational Studies in Mathematics*, 67(3), 205-216.
- Larsen, S., Glover, E. Melhuish, K. (in press). Beyond Good Teaching: The Benefits and Challenges of Implementing Ambitious Teaching. In D. Bressoud, V. Mesa, & C. Rasmussen (Eds.) *Insights and Recommendations from the MAA National Study of College Calculus*. Mathematical Association of America.
- Larsen, S., Johnson, E., & Bartlo, J. (2013). Designing and scaling up an innovation in abstract algebra. *The Journal of Mathematical Behavior*, 32(4), 693-711.
- Larsen, S., Johnson, E., & Weber, K. (Eds.). (2013). The teaching abstract algebra for understanding project: Designing and scaling up a curriculum innovation [Special Issue]. *The Journal of Mathematical Behavior*, 16(3).
- Leron, U., & Dubinsky, E. (1995). An abstract algebra story. *The American Mathematical Monthly*, 102(3), 227-242.
- Leron, U., & Hazzan, O. (2006). The rationality debate: Application of cognitive psychology to mathematics education. *Educational Studies in Mathematics*, 62(2), 105-126.



- Leron, U., Hazzan, O., & Zazkis, R. (1995). Learning group isomorphism: A crossroads of many concepts. *Educational Studies in Mathematics*, 29(2), 153-174.
- Lesh, R. (1981). Applied mathematical problem solving. *Educational Studies in Mathematics*, 12(2), 235–264.
- Lesh, R. A., Jr. (1976). The influence of two types of advanced organizers on an instructional unit about finite groups. *Journal for Research in Mathematics Education*, 7(2), 87–91.
- Li, Y. (2000). A comparison of problems that follow selected content presentations in American and Chinese mathematics textbooks. *Journal for Research in Mathematics Education*, 234–241.
- Lin, C.-J. (2008). Comparisons between classical test theory and item response theory in automated assembly of parallel test forms. *The Journal of Technology, Learning and Assessment*, 6(8).
- Lindell, R. S., Peak, E., & Foster, T. M. (2007, January). Are they all created equal? A comparison of different concept inventory development methodologies. In *2006 Physics Education Research Conference* (Vol. 883, No. 1, pp. 14-17). AIP Publishing.
- Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational researcher*, 36(8), 437-448.
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological bulletin*, 70(3p1), 151.
- Macdonald, P., & Paunonen, S. V. (2002). A Monte Carlo comparison of item and person statistics based on item response theory versus classical test theory. *Educational and Psychological Measurement*, 62(6), 921–943.
- Makel, M. C., & Plucker, J. A. (2014). Facts Are More Important Than Novelty Replication in the Education Sciences. *Educational Researcher*.
- Melhuish, K., & Larsen, S. (2015) *An investigation of students' strategic decisions when structuring proofs of if-then statements*. Manuscript in preparation.
- Mesa, V. (2004). Characterizing practices associated with functions in middle school textbooks: An empirical approach. *Educational Studies in Mathematics*, 56(2-3), 255–286.
- Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741.
- MisLevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary research and perspectives*, 1(1), 3-62.
- Moore, R. C. (1994). Making the transition to formal proof. *Educational Studies in mathematics*, 27(3), 249-266.
- Nardi, E. (2000). Mathematics undergraduates' responses to semantic abbreviations, "geometric" images and multi-level abstractions in group theory. *Educational Studies in Mathematics*, 43(2), 169–189.

- Novotná, J., Stehlíková, N., & Hoch, M. (2006, July). Structure sense for university algebra. In Proceedings of the 30th Conference of the International Group for the Psychology of Mathematics Education (Vol. 4, pp. 249-256).
- Olds, B. M., Streveler, R. A., Miller, R. L., & Nelson, M. A. (2004). Preliminary results from the development of a concept inventory in thermal and transport science. In *CD) Proceedings, 2004 American Society for Engineering Education Conference*. Retrieved from <http://www.thermalinventory.com/images/Papers/2004PrelResultsDevelopment.pdf>
- Otten, S., Gilbertson, N. J., Males, L. M., & Clark, D. L. (2014). The mathematical nature of reasoning-and-proving opportunities in geometry textbooks. *Mathematical Thinking and Learning*, 16(1), 51–79.
- Otten, S., Males, L. M., & Gilbertson, N. J. (2014). The introduction of proof in secondary geometry textbooks. *International Journal of Educational Research*, 64, 107-118.
- Pegg, J. (2003). Assessment in mathematics. *Mathematical cognition*, 227-259.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. National Academies Press.
- Raman, M. (2002). Coordinating informal and formal aspects of mathematics: Student behavior and textbook messages. *The Journal of Mathematical Behavior*, 21(2), 135–150.
- RUMEonline! - Periodic Publications in Mathematics Education. (2011). Retrieved August 28, 2014, from <http://sigmaa.maa.org/rume/journals.html>
- Savinainen, A., & Scott, P. (2002a). The Force Concept Inventory: a tool for monitoring student learning. *Physics Education*, 37(1), 45.
- Savinainen, A., & Scott, P. (2002b). Using the Force Concept Inventory to monitor student learning and to plan teaching. *Physics Education*, 37(1), 53.
- Savinainen, A., & Viiri, J. (2008). The force concept inventory as a measure of students conceptual coherence. *International Journal of Science and Mathematics Education*, 6(4), 719-740.
- Schoenfeld, A. (2007). Method. In F. K. Lester. (Ed.). (2007). *Second handbook of research on mathematics teaching and learning: A project of the National Council of Teachers of Mathematics* (Vol. 1, pp. 69-107). United States: Information Age Publisher.
- Selden, A., & Selden, J. (1987). Errors and misconceptions in college level theorem proving. In *Proceedings of the second international seminar on misconceptions and educational strategies in science and mathematics* (Vol. 3, pp. 457–470).
- Selden, J., & Selden, A. (1995). Unpacking the logic of mathematical statements. *Educational Studies in Mathematics*, 29(2), 123–151.

- Sfard, A., & Linchevski, L. (1994). The gains and the pitfalls of reification: The case of algebra. *Educational Studies in Mathematics*, 26(2/3), 191–228.
- Siebert, D. & Williams, S. R. (2003). Students' Understanding of  $\mathbf{Z}_n$ . *International Group for the Psychology of Mathematics Education*, 4, 167-173.
- Simpson, A., & Stehlíková, N. (2006). Apprehending mathematical structure: A case study of coming to understand a commutative ring. *Educational Studies in Mathematics*, 61(3), 347–371. doi:10.1007/s10649-006-1300-y
- Sinclair, N., Watson, A., Zazkis, R., & Mason, J. (2011). The structuring of personal example spaces. *The Journal of Mathematical Behavior*, 30(4), 291–303.
- Skemp, R. R. (1976). Relational understanding and instrumental understanding. *Mathematics teaching*, 77, 20-26.
- Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The genetics concept assessment: a new concept inventory for gauging student understanding of genetics. *CBE-Life Sciences Education*, 7(4), 422–430.
- Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education*, 36(5), 404–411. doi:10.2307/30034943
- Stone, A., Allen, K., Rhoads, T. R., Murphy, T. J., Shehab, R. L., & Saha, C. (2003). The statistics concept inventory: a pilot study. In *Frontiers in Education, 2003. FIE 2003 33rd Annual* (Vol. 1, pp. T3D–1–6 Vol.1). doi:10.1109/FIE.2003.1263336
- Streveler, R. A., Miller, R. L., Santiago-Román, A. I., Nelson, M. A., Geist, M. R., & Olds, B. M. (2011). Rigorous methodology for concept inventory development: Using the “assessment triangle” to develop and test the Thermal and Transport Science Concept Inventory (TTCI). *International Journal of Engineering Education*, 27(5), 968.
- Streveler, R. A., Olds, B. M., Miller, R. L., & Nelson, M. A. (2003, June). Using a Delphi study to identify the most difficult concepts for students to master in thermal and transport science. In *Proceedings of the Annual Conference of the American Society for Engineering Education*.
- Suominen, A. (2014). Perceptions in Abstract Algebra: Identifying major concepts and concept connections within Abstract Algebra . In *17th Conference on Research in Undergraduate Mathematics Education*, Denver, CO.
- Tall, D. (1999). Reflections on APOS theory in elementary and advanced mathematical thinking. In *PME CONFERENCE* (Vol. 1, pp. 1–111). Retrieved from <http://homepages.warwick.ac.uk.proxy.lib.pdx.edu/staff/David.Tall/pdfs/dot1999c-apos-in-amt-pme.pdf>
- Tall, D. (2004, July). Thinking through three worlds of mathematics. In *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 281-288).
- Tall, D. O. (2007). Developing a theory of mathematical growth. *ZDM*, 39(1-2), 145–154.

- Tall, D., & Vinner, S. (1981). Concept image and concept definition in mathematics with particular reference to limits and continuity. *Educational Studies in Mathematics*, 12(2), 151–169.
- Tallman, M. A., & Carlson, M. P. (2012). A characterization of Calculus I final exams in US colleges and universities. In *The 15th annual conference on Research in Collegiate Mathematics Education*. Portland, OR.
- Thomas, M., Gleason, J., Bagley, S., Rice, L., Clements, N., & White, D. Psychometric analysis of the Calculus Concept Inventory. In *The 18th annual conference on Research in Collegiate Mathematics Education*. Pittsburgh, PA.
- Thompson, D. R., Senk, S. L., & Johnson, G. J. (2012). Opportunities to learn reasoning and proof in high school mathematics textbooks. *Journal for Research in Mathematics Education*, 43(3), 253–295.
- Toerner, G., & Arzarello, F. (2012). Grading mathematics education research journals. *Mathematical Society*, 52–54.
- Travers, K. J., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula*. Elmsford, NY: Pergamon Press.
- Trochim, W. M., & Donnelly, J. P. (2008). *Research methods knowledge base*. Atomic Dog/Cengage Learning Mason, OH. Retrieved from <http://anatomyfacts.com/Research/ResearchMethodsKnowledgeBase.pdf>
- Tryfos, P. (1998). *Methods for Business Analysis and Forecasting: Text and Cases*. Wiley.
- Tucker, A. (1996). Models that work: Case studies in effective undergraduate mathematics programs. *MAA Notes*, 38.
- Wage, K. E., Buck, J. R., Wright, C. H. G., & Welch, T. B. (2005). The signals and systems concept inventory. *IEEE Transactions on Education*, 48(3), 448–461.
- Wallace, C. S., & Bailey, J. M. (2010). Do concept inventories actually measure anything? *Astronomy Education Review*, 9(1).
- Watson, A., & Mason, J. (2006). *Mathematics as a constructive activity: Learners generating examples*. Routledge.
- Weber, K. (2001). Student difficulty in constructing proofs: The need for strategic knowledge. *Educational studies in mathematics*, 48(1), 101–119.
- Weber, K. (2009). How syntactic reasoners can develop understanding, evaluate conjectures, and generate counterexamples in advanced mathematics. *The Journal of Mathematical Behavior*, 28(2), 200–208.
- Weber, K. (2013). Benefits and limitations of complementing qualitative research with quantitative studies: Lessons learned from four studies on proof reading. Presented at a plenary session of the *16th Conference on Research in Undergraduate Mathematics Education*. Denver, Colorado.
- Weber, K., & Alcock, L. (2004). Semantic and syntactic proof productions. *Educational Studies in Mathematics*, 56(2), 209–234.
- Weber, K., & Larsen, S. (2008). Teaching and learning group theory. *Making the Connection: Research and Teaching in Undergraduate Mathematics Education*, (73), 139.

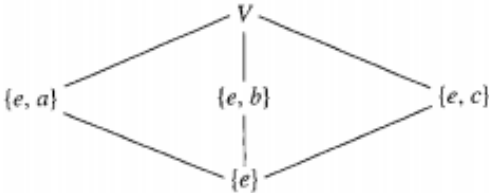
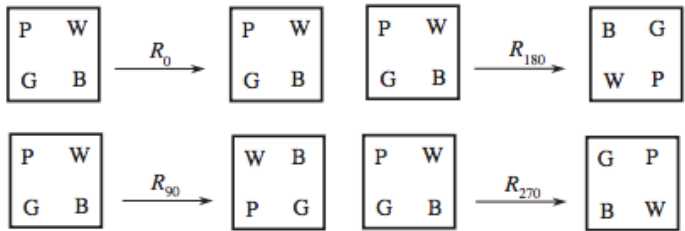
- Weber, K., & Mejia-Ramos, J. P. (2011). Why and how mathematicians read proofs: an exploratory study. *Educational Studies in Mathematics*, 76(3), 329–344.
- Wheeler, A., & Champion, J. (2013). Students' proofs of one-to-one and onto properties in introductory abstract algebra. *International Journal of Mathematical Education in Science and Technology*, 44(8), 1107–1116.
- Yong, E. (2012). In the wake of high profile controversies, psychologists are facing up to problems with replication. *Nature*, 483, 298–300.
- Zaslavsky, O., & Peled, I. (1996). Inhibiting factors in generating examples by mathematics teachers and student teachers: The case of binary operation. *Journal for Research in Mathematics Education*, 27(1), 67–78.  
doi:10.2307/749198
- Zazkis, R., & Leikin, R. (2008). Exemplifying definitions: a case of a square. *Educational Studies in Mathematics*, 69(2), 131–148.
- Zazkis, R., Dubinsky, E., & Dautermann, J. (1996). Coordinating visual and analytic strategies: A study of students' understanding of the group D4. *Journal for Research in Mathematics Education*, 27(4), 435–457.  
doi:10.2307/749876
- Zhu, Y., & Fan, L. (2006). Focus on the representation of problem types in intended curriculum: A comparison of selected mathematics textbooks from Mainland China and the United States. *International Journal of Science and Mathematics Education*, 4(4), 609–626.

# Appendix A

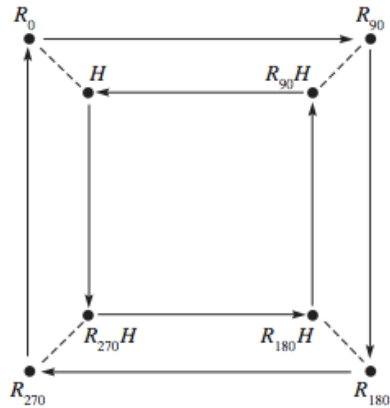
## Textbook Analysis Codes

Table A.1

### Representation Codes for Groups

Representation	Code	Example																									
Group Name	GN	$\mathbb{Z}$ or $\langle \mathbb{Z}, + \rangle$ (integers under addition)																									
Set Builder	GS	$\{\pi^n \mid n \in \mathbb{Z}\}$ under multiplication																									
Subgroup Diagram	GD																										
Generators (Group Presentation)	GP	(Klein-4) $\langle 1 \rangle$																									
Cayley Table	GT	(Integers under addition) <table data-bbox="755 907 1133 1104"><tr><th>mod 10</th><th>1</th><th>3</th><th>7</th><th>9</th></tr><tr><td>1</td><td>1</td><td>3</td><td>7</td><td>9</td></tr><tr><td>3</td><td>3</td><td>9</td><td>1</td><td>7</td></tr><tr><td>7</td><td>7</td><td>1</td><td>9</td><td>3</td></tr><tr><td>9</td><td>9</td><td>7</td><td>3</td><td>1</td></tr></table>	mod 10	1	3	7	9	1	1	3	7	9	3	3	9	1	7	7	7	1	9	3	9	9	7	3	1
mod 10	1	3	7	9																							
1	1	3	7	9																							
3	3	9	1	7																							
7	7	1	9	3																							
9	9	7	3	1																							
List of Elements	GE	(U(10)) $\{1, 3, 7, 9\}$ under multiplication modulo 10																									
Verbal Description	GV	(U(10)) All $n \times n$ diagonal matrices under matrix multiplication.																									
Geometric Image	GG																										
Left Regular	GL	(Rotations of a Square) (I) (R, R2, I)(FR2, FR, F) (I, R2, R)(F, FR, FR2) (I, F)(FR2, R2)(FR, R) (FR, I)(F, R2)(FR2, R) (F, R)(FR, R2)(R, I) (Dihedral group of order 6)																									

Cayley Digraph      GC



(Dihedral group of order 8)

Matrix      GM

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}$$

Permutation      GP

(Klein-4)

$$T_1 = \begin{bmatrix} 1 & 5 & 7 & 11 \\ 1 & 5 & 7 & 11 \end{bmatrix}, \quad T_5 = \begin{bmatrix} 1 & 5 & 7 & 11 \\ 5 & 1 & 11 & 7 \end{bmatrix},$$

$$T_7 = \begin{bmatrix} 1 & 5 & 7 & 11 \\ 7 & 11 & 1 & 5 \end{bmatrix}, \quad T_{11} = \begin{bmatrix} 1 & 5 & 7 & 11 \\ 11 & 7 & 5 & 1 \end{bmatrix}.$$

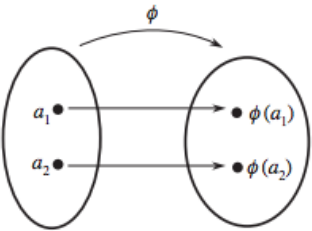
(U(10))

Isomorphisms/ Homomorphisms / Mappings

Table A.2

Representation Codes for Maps

Representation	Code	Example
Function Diagram	FD	



Symbolic defined function      FS

$$\theta(x + y) = 2^{xy}$$

Function defined element-wise	FR	$  \begin{array}{lcl}  x & \xrightarrow{\phi_{R_{90}}} & R_{90} x R_{90}^{-1} \\  R_0 & \rightarrow & R_{90} R_0 R_{90}^{-1} = R_0 \\  R_{90} & \rightarrow & R_{90} R_{90} R_{90}^{-1} = R_{90} \\  R_{180} & \rightarrow & R_{90} R_{180} R_{90}^{-1} = R_{180} \\  R_{270} & \rightarrow & R_{90} R_{270} R_{90}^{-1} = R_{270} \\  H & \rightarrow & R_{90} H R_{90}^{-1} = V \\  V & \rightarrow & R_{90} V R_{90}^{-1} = H \\  D & \rightarrow & R_{90} D R_{90}^{-1} = D' \\  D' & \rightarrow & R_{90} D' R_{90}^{-1} = D  \end{array}  $
Function defined on generating set	FG	$\phi$ is an automorphism of $D_4$ such that $\phi(R_{90})=R_{270}$ and $\phi(V)=V$
Function defined verbally	FV	For any $f$ in $R[x]$ , let $f'$ denote the derivative of $f$ .

Table A.3

*Expected Student Activity Codes*

Code	Expected Activity	Description
RC	Change Representation	Changing between representations such as going from a description of a group to a table
DC	Determine/Calculate	Computational questions such as finding order, number of subgroups, etc.
EV	Evaluate	Determine if a statement is true or false or if something is an instance of concept or not. Generally true or false or yes or no questions.
RE	Recall	Correcting or recalling a definition or theorem without usage
CO	Conjecture	Determine the response to an open-ended question requiring a mathematical statement
EX	Provide Example	Provide an example that meets some criteria. This might get double-coded as PC if the exercise addresses that this example is a counterexample to some statement.
PS	Prove Definition Satisfied	Show that a given example meets the definition of a concept or satisfies the definition of having some property.
PC	Prove by Counterexample	Show a statement is false via example.
PT	Proof Direct Consequence of Theorem	Show a direct consequence of one of the identified theorems.
PD	Proof Direct Consequence of Definition	(This is not saying something satisfies a definition, but rather is some conclusion you can get to by just using the definition of a topic. For example, $\phi(x^3)=\phi(x)^3$ is a direct consequence of the definition of homomorphism.)
PA	Proof Advanced	Proofs that rely heavily on advanced proof techniques, putting things together in nontraditional manners or pulling together many concepts in a way that is not immediately obvious from theorems and definitions alone
PN	Proof near concept in other manner	Proving something that seems still intimately tied to a concept other than the ways above. (This is sort of a back up code to capture exercises that still seem essential)
PP	Show not isomorphic	Use a structural property to show two groups are not isomorphic.



---

Table A.4

*Example Codes*

---

Code	Description
X	Example
NX	Non-example
EMD	Example motivating a definition
EFD	Example of concept following definition
EP	Example illustrating a specific property a concept does or does not have
EC	Example illustrating how to calculate or determine something
ET	Example illustrating a proving technique
EMT	Example motivating a theorem
EIT	Example illustrating a theorem
EUT	Example using a theorem
EIN	Example illustrating a notation

---

## Appendix B

### Sample Analytic Report from Textbook Analysis

#### Isomorphism

##### Formal Definition

Let  $G$  be a group with respect to  $*$ , and let  $G'$  be a group with respect to  $\bullet$ . A mapping  $\phi: G \rightarrow G'$  is an isomorphism from  $G$  to  $G'$  if

1.  $\phi$  is a one-to-one correspondence from  $G$  to  $G'$ , and
2.  $\phi(x*y) = \phi(x) \bullet \phi(y)$  for all  $x$  and  $y$  in  $G$ .

Formal definitions are pretty consistent.

##### Informal Description

Informal descriptions exist in all textbooks of some form where it is only “names” or “labels” that are different. Three of the texts used operation tables to illustrate the sameness.

This definition can be visualized as shown in Figure 6.1. The pairs of dashed arrows represent the group operations.

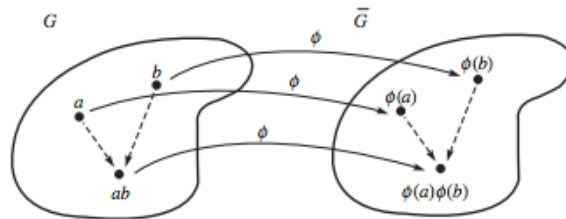


Figure 6.1

Table A.5

#### Representations of Maps Across the Textbook

	Function Diagram	Function Described Verbally	Function Described by Symbolic Rule	Function Defined Element- Wise	Function Defined on Generators
Fraleigh					
Examples	0%	5.26%	78.95%	10.53%	5.26%
Hungerford					
Examples	0%	0%	92.86%	7.14%	0%
G & G					
Examples	0%	0%	81.82%	9.09%	9.09%
Gallian					
Examples	12.5%	0%	87.5%	0%	0%

<b>Examples</b>					
<b>Average</b>	<b>3.13%</b>	<b>1.32%</b>	<b>85.28%</b>	<b>6.69%</b>	<b>3.59%</b>
Frleigh					
Exercises	0%	4.05%	79.73%	4.05%	12.16%
Hungerford					
Exercises	0%	4.35%	95.65%	0%	0%
G & G					
Exercises	0%	0%	100%	0%	0%
Gallian					
Exercises	0%	6.45%	77.42%	3.23%	12.9%
<b>Exercises</b>					
<b>Average</b>	<b>0%</b>	<b>3.7%</b>	<b>88.2%</b>	<b>1.8%</b>	<b>6.3%</b>

### Examples

Table A.6

#### *Example Motivating Definition*

Author	Examples
Frleigh	French Numbers, Three Element tables
Gallian	German numbers D <sub>4</sub> (described geometrically vs. permutations) cyclic group isomorphic to modular groups U(43), U(49) <a>
Hungerford	Roman numerals {1,i,-i,-1} and U5
G & G	cyclic group order 4 D <sub>3</sub> , S <sub>3</sub>

Table A.7

#### *Example Following Definition*

Author	Examples
Frleigh	R,+ and R <sup>+</sup> ,* Z and 2Z
Gallian	3 element group and image under isomorphism R,+ and R <sup>+</sup> ,* infinite cyclic groups and Z U(10), Z <sub>4</sub> and U(5) SL(2,R) to itself with conjugation map R+ C+ C, *, C, * with  1

Hungerford	$U_8$ and $Z_2 \times Z_2$ $2Z$ and $Z$ $R, +$ and $R^+, *$ identity map automorphisms inner automorphisms
G & G	$\{1, i, -1, -i\}$ and $Z_4$ 3x3 permutation matrices and $S_3$

Table A.7

*Non-Examples*

Author	Examples
Fraleigh	$Q$ and $R$ $Z$ and $Z^+$ $Q$ and $Z$ $C$ and $R$ $M_2(R)$ and $R$
Gallian	$R^+ : x \rightarrow x^3$ $U(10)$ and $U(12)$ $Q^*$ and $Q^+$ $R^*$ and $C^*$
Hungerford	$Z_5$ and $Z_{10}$ $S_3$ and $S_6$ $Z_4$ and $Z_2 \times Z_2$
G & G	none

**Exercises**

Table A.8

*Exercise Types*

	Show Map is Iso	Show Groups are Iso	Show Map is not Iso	Show groups not Iso	Evaluate if groups are iso	Find Iso between groups	Properties preserved	Find Iso group	Eval if map is iso
Fraleigh	<b>X</b>	<b>X</b>	<b>X</b>			<b>X</b>	<b>X</b>		<b>X</b>
Gallian	x	<b>X</b>	<b>X</b>	<b>X</b>		x	x	x	
Hungerford	<b>X</b>	<b>X</b>	x	<b>X</b>	x				
G & G	<b>X</b>	<b>X</b>			<b>X</b>	<b>X</b>	<b>X</b>		

*Note:* x=single exercise, **X**=multiple exercises

## Appendix C

### Sample Literature Review Report

Table A.10

#### *Literature Report on Order of an Element*

APA citation	Student Conception/Results	Task
Hazzan, O., & Leron, U. (1996). <i>Students' use and misuse of mathematical theorems: The case of Lagrange's theorem. For the Learning of Mathematics</i> , 23-26.	<ul style="list-style-type: none"> <li>• Applying Lagrange's Theorem to find elements of a certain order.</li> <li>• Believing the <math>a^7=e</math> implies <math>a^8=e</math>.</li> </ul>	True or false? Please justify your answer. "In S7 there is no element of order 8."
Brown, A., DeVries, D. J., Dubinsky, E., & Thomas, K. (1997). <i>Learning binary operations, groups, and subgroups. The Journal of Mathematical Behavior</i> , 16(3), 187-239.	<ul style="list-style-type: none"> <li>• Randomly trying all combinations of elements</li> <li>• No Progress</li> <li>• Leveraging the LCM relationship</li> </ul>	Find an element of order 6 in commutative group with element of order 2 and 3.
Anderson, J., Austin, K., Barnard, T., & Jagger, J. (1998). <i>Do third year mathematics undergraduates know what they are supposed to know? International Journal of Mathematical Education in Science and Technology</i> , 29(3), 401-420.	<ul style="list-style-type: none"> <li>• Few could do so</li> <li>• Only consider <math>x^n=e</math> without the minimum requirement</li> </ul>	Define the order of an element.
Nardi, E. (2000). <i>Mathematics Undergraduates' Responses to Semantic Abbreviations, 'Geometric' Images and Multi-Level Abstractions in Group Theory. Educational Studies in Mathematics</i> , 43(2), 169-189.	<ul style="list-style-type: none"> <li>• Not understanding how elements and groups have order</li> <li>• Not seeing the connection between order of element and set generated by element</li> <li>• Order of an element is both static (the number of elements) and process (how to generate these elements)</li> </ul>	Working through a proof

## Appendix D

### Expert Evaluation Summary

*Task 1:* Let  $a, b \in G$  where  $G$  is a group. Consider the equation:

$$(ab)^n = a^n b^n.$$

Will this equation always hold? Why or why not?

Table A.10

<i>Expert Evaluation for Abelian Groups Task 1</i>				
Evaluator	Important	Relevant	What might students do?	Other Notes
1	Y	Y	Informally think of ability to “collect terms” Formally, a recursive/inductive argument	
2	Y	Y	Borrowing from familiar groups	Easy
3	Y	Y	Leverage definition	Consider switching from $n$ to concrete number Connects to difference between Abelian and non-Abelian groups
4	Y	Y	Borrowing from familiar groups	
5	Y	Y	Borrowing from familiar groups	Consider switching from $n$ to concrete number

*Task 2:* Let  $G$  be a non-Abelian group. What can we say about its subgroups?

Table A.11

<i>Expert Evaluation for Abelian Groups Task 2</i>				
Evaluator	Important	Relevant	What might students do?	Other Notes
1	N	N		Too vague
2	Y	Y	Might not know where to start	Too vague
3	N	N	Consider things not related to abelian	Too vague
4	Y	Y	Students may assume non-abelian is inherited	
5	Y	N		Too vague About non-abelian not abelian

## Appendix E

### Item Analysis

Table A.12

<i>Item Analysis Statistics</i>					
Item Number	Topic	Item Difficulty	Corrected Item Correlation	Cronbach's Alpha if Deleted	Question Evaluation
1	Abelian Groups	.49171	.332	.620	Optimal
2	Group	.1883	.068	.653	Consider Revising
3	Identity	.2679	.272	.629	Difficult
4	Isomorphism	.1220	.051	.652	Consider Revising
5	Homomorphism	.3634	.119	.650	Difficult
6	Lagrange's Theorem	.5385	.257	.631	Optimal
7	Cyclic Groups	.5597	.152	.646	Mid
8	Kernel	.5066	.389	.612	Optimal
9	Associativity	.3263	.347	.619	Difficult
10	First Isomorphism Theorem	.1989	.194	.639	Consider Revising
11	Subgroup	.4350	.391	.612	Optimal
12	Cosets	.3077	.305	.625	Difficult
13	Quotient Groups	.2228	.077	.653	Consider Revising
14	Inverse	.7029	.201	.639	Optimal
15	Order of an Element	.4642	.361	.616	Optimal
16	Normal Groups	.3660	.368	.615	Difficult
17	Binary Operation	.5464	.258	.631	Optimal



## Appendix F

### Principal Component Analysis - % of Variance

Table A.13

#### *PCA Loadings*

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.793	16.432	16.432	2.793	16.432	16.432	1.931	11.357	11.357
2	1.222	7.186	23.618	1.222	7.186	23.618	1.565	9.205	20.562
3	1.191	7.009	30.627	1.191	7.009	30.627	1.443	8.488	29.049
4	1.163	6.844	37.470	1.163	6.844	37.470	1.220	7.179	36.228
5	1.103	6.487	43.957	1.103	6.487	43.957	1.196	7.035	43.263
6	1.016	5.976	49.933	1.016	5.976	49.933	1.134	6.670	49.933
7	.951	5.597	55.530						
8	.949	5.585	61.115						
9	.941	5.535	66.650						
10	.874	5.140	71.790						
11	.828	4.871	76.662						
12	.742	4.365	81.027						
13	.728	4.280	85.307						
14	.695	4.091	89.397						
15	.648	3.815	93.212						
16	.595	3.499	96.711						
17	.559	3.289	100.000						

## Appendix G

### Distractor Frequencies

Table A.14

*Percentage of Students Selecting Each Answer*

Item Number	Topic	Percentage of students selecting each response
1	Abelian Groups	a. 15% b. 14% <b>c. 59%</b> d. 12%
2	Group	<b>a. 19%</b> b. 38% c. 18% d. 10% e. 14%
3	Identity	a. 30% b. 16% c. 13% <b>d. 27%</b> e. 14%
4	Isomorphism	a. 21% b. 12% c. 30% <b>d. 13%</b> e. 24%
5	Homomorphism	a. 44% <b>b. 37%</b> c. 12% d. 7%
6	Lagrange's Theorem	a. 8% b. 38% <b>c. 54%</b>
7	Cyclic Groups	a. 13% <b>b. 56%</b> c. 17% d. 14%
8	Kernel	a. 28% b. 10% <b>c. 51%</b> d. 11%

9	Associativity	a. 28% b. 22% c. 17% <b>d. 32%</b>
10	First Isomorphism Theorem	<b>a. 20%</b> b. 30% c. 28% d. 23%
11	Subgroup	a. 14% b. 36% c. 7% <b>d. 44%</b>
12	Cosets	a. 10% b. 44% <b>c. 31%</b> d. 16%
13	Quotient Groups	<b>a. 22%</b> b. 35% c. 29% d. 15%
14	Inverse	<b>a. 70%</b> b. 14% c. 9% d. 8%
15	Order of an Element	a. 36% <b>b. 45%</b> c. 12% d. 6%
16	Normal Groups	a. 43% <b>b. 37%</b> c. 14% d. 6%
17	Binary Operation	a. 10% b. 20% c. 17% <b>d. 54%</b>

---

*Note: Bolded response represents correct answer*