A Method for Assessing and Describing the Informal Inferential Reasoning of Middle School Students

Joshua Michael Goss

Western Michigan University, joshuamgoss@aol.com

Follow this and additional works at: https://scholarworks.wmich.edu/dissertations

Part of the Curriculum and Instruction Commons, Junior High, Intermediate, Middle School Education and Teaching Commons, and the Science and Mathematics Education Commons

Recommended Citation

https://scholarworks.wmich.edu/dissertations/278

This Dissertation-Open Access is brought to you for free and open access by the Graduate College at ScholarWorks at WMU. It has been accepted for inclusion in Dissertations by an authorized administrator of ScholarWorks at WMU. For more information, please contact maira.bundza@wmich.edu.
A METHOD FOR ASSESSING AND DESCRIBING THE INFORMAL INFERENTIAL REASONING OF MIDDLE SCHOOL STUDENTS

by

Joshua Michael Goss

A dissertation submitted to the Graduate College
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
Mathematics
Western Michigan University
June 2014

Doctoral Committee:

Christine Browning, Ph.D., Chair
Steven W. Ziebarth, Ph.D.
Dwayne Channell, Ph.D.
J. Michael Shaughnessy, Ph.D.
A METHOD FOR ASSESSING AND DESCRIBING THE INFORMAL INFERENTIAL REASONING OF MIDDLE SCHOOL STUDENTS

Joshua Michael Goss, Ph.D.
Western Michigan University, 2014

Informal Inferential Reasoning (IIR) has emerged in the last decade in the study of statistics education. Developing students’ IIR ability is seen as a way of preparing students for the important topic of Formal Statistical Inference (FSI); however, research is still needed in order to investigate how students transition between informal and formal statistical reasoning. A primary difficulty is that we do not have a way of assessing and describing students’ IIR ability levels. In order to address this, an Assessment of Informal Inferential Reasoning (AIIR) was developed, along with a Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs & Collis, 1982) adaptation focusing on describing students’ IIR abilities. Results of the research show that for the age range investigated, the AIIR and SOLO taxonomy adaptation reliably identified and classified students’ IIR abilities.
Copyright by
Joshua Michael Goss
2014
ACKNOWLEDGMENTS

I want to begin by thanking the members of my dissertation committee for their support. Dr. J. Michael Shaughnessy served as my outside reader, and was invaluable in helping focus and guide the direction of the research. Dr. Dwayne Channell was third reader, and his encouragement, support, and at times commiseration was an enormous help in encouraging me to be bold in letting my stance on issues show through in my writing. I owe an enormous amount of thanks to Dr. Steven Ziebarth, who acted as second reader; his insight and keen observations were invaluable at every step of the research. As chair of my dissertation committee, Dr. Christine Browning has been everything I could have asked for in a mentor; she kept me moving forward when I felt stuck, coached me through the many unexpected challenges that came up during the research, and helped me feel like a part of the mathematics education research community in general.

I would certainly not be here were it not for the support of my parents. My educational journey has not taken the straightest or easiest route, and their constant support has made an enormous difference. I’d also like to thank my siblings Dan and Beth and their families for encouraging me along the way, and my son Drake for helping me find an all new perspective on why this work is so important.
Acknowledgments—Continued

Finally, for her encouragement, for her steadfast belief in my abilities, and for the countless ways that she stood by and supported me, I would like to thank my wife, Feng-Chiu.

Joshua Michael Goss
# TABLE OF CONTENTS

**ACKNOWLEDGMENTS** .......................................................................................................................... ii

**LIST OF TABLES** ................................................................................................................................. vii

**LIST OF FIGURES** ................................................................................................................................. viii

Chapter 1: Introduction ............................................................................................................................. 1

  The Case for Statistics Education ........................................................................................................... 2

  The Importance of Studying Informal and Formal Inferential Reasoning ........................................... 3

  Purpose of Current Study ......................................................................................................................... 7

Chapter 2: Review of the Literature ......................................................................................................... 9

  Link Between Formal and Informal Statistical Reasoning ...................................................................... 9

  AIIR Construction .................................................................................................................................. 13

  Structure of Observed Learning Outcomes (SOLO) .............................................................................. 21

  Explanations in IIR .................................................................................................................................. 23

  Summary ................................................................................................................................................... 25

Chapter 3: Methodology .......................................................................................................................... 27

  Frameworks ............................................................................................................................................. 27

  Pilot Study Results .................................................................................................................................. 30

  Research Design ...................................................................................................................................... 32

  Summary ................................................................................................................................................... 39
Table of Contents—Continued

Chapter 4: Unit and Task Design and SOLO Modification .......................................................... 40
  Design of the Teaching Unit and Additional AIIR Items ...................................................... 40
  SOLO Taxonomy Adaptation ................................................................................................. 45

Chapter 5: AIIR Modification and Co-Measurability Analysis .............................................. 53
  Initial Co-Measurability Check ............................................................................................. 53
  Modifications ......................................................................................................................... 68
  Final Co-Measurability Check .............................................................................................. 70
  Problem-Type Analysis ........................................................................................................ 81
  Summary ............................................................................................................................... 91

Chapter 6: Conclusions ............................................................................................................. 93
  Question 1: SOLO Taxonomy Results .................................................................................. 93
  Question 2: AIIR Results ..................................................................................................... 95
  Other Results ....................................................................................................................... 98
  Limitations .......................................................................................................................... 101
  Implications and Open Questions ....................................................................................... 103
  Closing Remarks ................................................................................................................ 111

References .................................................................................................................................. 113

Appendix A: Initial AIIR Items ................................................................................................. 120

Appendix B: AIIR Item Bank (Initial Co-Measurability Check) ............................................ 122
Table of Contents—Continued

Appendix C: Additional Initial Co-Measurability Analyses ......................................................... 134

Appendix D: Additional Final Co-Measurability Analyses ............................................................. 140

Appendix E: Final Item Bank ........................................................................................................ 165

Appendix F: Human Subjects Institutional Review Board Approval ............................................ 174
**LIST OF TABLES**

1. Assessment framework from Zieffler et al. (2008) ................................................................. 13
2. SOLO adaptation cycle definitions developed during the pilot study .............................. 29
3. Research question/sample/methodology brief overview .................................................. 38
4. Unit outline .................................................................................................................................. 42
5. Contexts for new AIIR items ......................................................................................................... 45
6. Final SOLO cycle indicators ......................................................................................................... 49
7. SOLO taxonomy examples and codes for EPG question type .............................................. 50
8. SOLO taxonomy examples and codes for CTS question type .............................................. 51
9. SOLO taxonomy examples and codes for CBM question type .............................................. 52
10. Initial co-measurability check AIIR versions and their items .............................................. 54
11. Assessed and interview SOLO level results for initial co-measurability check............. 59
12. Final co-measurability check AIIR versions ........................................................................... 69
13. EPG final co-measurability check results by item combinations ......................................... 83
14. CTS item combination results organized by student pseudonyms ........................................ 84
15. CBM item combination results organized by student pseudonyms .................................... 90
LIST OF FIGURES

1. SOLO taxonomy visualization from Groth (2012, p. 31)................................. 23
2. Sample organization...................................................................................... 37
3. Example of the CTS alternative task type...................................................... 44
4. Graph form CBM3 part A............................................................................. 47
5. Example of the doubling strategy Betsy’s pre-AIIR ........................................ 56
6. Erica’s AIIR (left) and interview (right) growing a sample task responses .......... 63
7. Frank’s ‘growing a sample’ response, with new points as boxes......................... 66
8. Reupert’s population graph for EPG4............................................................. 75
9. Johny’s CTS4 response................................................................................. 81
10. Sofia’s CTS4 response.................................................................................. 85
11. Johny’s CTS2 response............................................................................... 88
12. Item CBM1 from the final item bank............................................................ 110
Chapter 1: Introduction

Statistics education has experienced a rise to prominence in recent years for the K-12 grades, with new and challenging topics for students, such as statistical inference, making their way into the middle school curriculum. A concern for middle school teachers is how to teach the challenging topics when research on how students best learn statistics, in particular statistical inference, is sparse. What has recently emerged when considering how to prepare students for their study of formal statistical inference (FSI) is research is on promoting informal inferential reasoning (IIR).

Zieffler, Garfield, Delmas, and Reading (2008) define IIR as “the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples” (p. 44). In this research Zeiffler et al.’s (2008) definition was used, supplemented by Makar and Rubin’s (2009) three main characteristics of appropriate use of IIR: generalizations beyond data, using data as evidence, and using probabilistic language in expressing uncertainty about generalizations.

The research on IIR suggests a need for understanding how students transition from informal statistical inference (ISI) tasks (that make use of their IIR) to their study and use of FSI. Studying this transition of reasoning requires a way of assessing and describing students’ IIR. This dissertation study built on previous pilot study work in the design and development of an assessment tool, an Assessment of Informal Inferential Reasoning (AIIR), to assess students IIR. Further
validation and refinement of a Structure of Observed Learning Outcomes (SOLO) taxonomy (Biggs & Collis, 1982) was also needed that describes and interprets the progression of student learning based upon the AIIR responses.

**The Case for Statistics Education**

As a consequence of the rapidly expanding availability of information (Makar & Rubin, 2009) statistical literacy has become an integral part of becoming an informed citizen, making informed personal choices, being a professional in the modern workplace, or working in the sciences (Franklin et al., 2005; Zieffler et al., 2008). In response to this increased importance of being statistically literate, there has been an increased emphasis on statistics education both in the US (National Council of Teachers of Mathematics, 1989, 2000; National Governors Association Center for Best Practices [NGACBP] & Council of Chief State School Officers [CCSSO], 2010), and internationally (Leavy, 2010; Makar & Rubin, 2009; Shaughnessy, 1992, 2007). This increased societal emphasis has resulted in a greater research emphasis, that has expanded our understanding of what statistics is important, when it is developmentally appropriate to teach and learn statistical topics, and what are some of the instructional concerns.

Makar and Rubin (2009) cite the prevalence and pervasiveness of technology in the educational system as encouraging a shift from teaching statistics as a system of tools and procedures towards a more “holistic, process-oriented approach that go(es) beyond data analysis techniques” (p. 83). In the US, this is consistent with the National Council of Teachers of Mathematics (NCTM) push for statistics in lower
grades in its *Principles and Standards for School Mathematics* (PSSM) (2000), and is also supported by the American Statistical Association's (ASA) 2007 endorsement of the *Guidelines for Assessment and Instruction in Statistics Education* (GAISE) report (Franklin et al., 2005). Consistent with PSSM (NCTM, 2000), the GAISE (2005) report identifies four goals that should pervade pre-kindergarten through grade 12, enabling students to:

- formulate questions that can be addressed with data and then collect, organize, and display relevant data to answer them;
- select and use appropriate statistical methods to analyze data;
- develop and evaluate inferences and predictions that are based on data; and
- understand and apply basic concepts of probability (Franklin et al., 2005).

A third important U.S. document is the *Common Core State Standards for Mathematics* (CCSSM) (NGACBP & CCSSO, 2010) that outlines a minimum core of proficiencies a student should achieve and the latest grade by which they should achieve them. With respect to statistics, CCSSM outlines the progression of topics and directly informs the forthcoming national assessments that will be used to assess learning of the proficiencies. Since high stakes testing has historically been a key influence on curriculum, this document can lend a great deal to understanding the anticipated classroom level focus.

**The Importance of Studying Informal and Formal Inferential Reasoning**

The GAISE (Franklin et al., 2005) report shows the recommendations that are directly supported by the ASA and are compatible with recommendations from
NCTM (2000). A primary contribution of the GAISE document is in the framework for classifying student statistical thought, organized by process components (I-IV), and progressing hierarchically through three levels (A, B, and C). The process components are:

I   Formulate Questions;
II  Collect Data;
III Analyze Data; and
IV Interpret Results.

The A-through-C levels provide a continuum from novice to proficient for the above process components. The intention of the GAISE document is to provide recommendations appropriate for pre-K through grade 12 with level A tasks directed towards elementary students, level B for middle school students, and level C for high school students. While these recommendations for grade-band appropriate work are the ideals put forward by Franklin et al. (2005) with the support of the ASA, it is by no means a given that students in a certain grade band will be operating at a given level. This is particularly apparent when considering that appropriate use of IIR is a level A objective that was largely beyond the ability of middle school students in the pilot version of this study (who should have been at level B). This was also demonstrated in a study done by Jacob (2013) where some level A types of thinking needed to be intentionally developed in an Advanced Placement (AP) statistics course.
In the GAISE document (Franklin et al., 2005), IIR begins in the elementary grade band, showing both support for the idea of IIR as foundational to the learning of statistics, and the importance for beginning informal statistical inference early in a student’s education. For example, in level A, GAISE advocates for tasks to include ideas about data and context, and notably includes the following two statements:

Students also should learn how to use basic statistical tools to analyze the data and **make informal inferences** in answering the posed questions.

Finally, students should develop **basic ideas of probability in order to support their later use of probability in drawing inferences at levels B and C.** (Franklin et al., 2007, p. 23)(emphasis added)

Note that GAISE is recommending that informal statistical inference (ISI) be introduced at the lower grade levels, and that ISI is considered separate from FSI at later levels. This inclusion of informal inference at the lower grade levels is consistent with some research being done internationally that emphasizes fostering IIR through tasks emphasizing ISI utilizing authentic statistics practices as early as Kindergarten and continuing through the elementary grades (Makar & Rubin, 2009; Paparistodemou & Meletiou-Mavrotheris, 2008, 2010). The investigative cycle introduced by Wild and Pfannkuch (1999) is an example of an authentic statistical practice that was used in the unit design for this current study and will be further described in Chapter 3.

If the GAISE document (Franklin et al., 2005) can be seen as the recommendations from the American statistics community with support from the
international statistics education research community, then the CCSSM (NGABP & CCSSO, 2010) can be seen as the requirements adopted by the (majority of) U.S. states for implementation in the classroom. For example, the Measurement and Data strands for grades K-5, and the Statistics and Probability strands in subsequent grades, show that in grades K-5 the data component focuses on representing data and drawing direct comparisons; specifically at grade five, the goal is for students to represent fractions on a line plot and to use those plot points to solve arithmetical problems about the data. CCSSM intends that grade 6 introduces measures of center and spread, along with concepts of variability, and in grade 7, students are tasked with drawing “informal comparative inferences about two populations” (p. 46) and using “random samples to draw inferences about a population” (p. 46).

The CCSSM (NGABP & CCSSO, 2010) K-5 Measurement and Data strand focus is primarily on graph creation and secondarily on using graphs as a context to frame arithmetic problems. This directly conflicts with research that has shown that the all-too-common case of students being unable to use descriptive statistics appropriately (Pfannkuch, Budgett, & Parsonage, 2004) stems from this focus on “graphs...with little attention given to the analysis or reasons the graphs were constructed in the first place” (Friel, Curcio, & Bright, 2001, p.132). With construction of graphs receiving attention in CCSSM six years before any discussion concerning graph analysis beyond direct comparison, the creation of the graphs themselves becomes the objective of learning about graphs; this creation of graphs for their own sake is mediated in some curricular materials (Pearson Scott
Foresman TERC, 2012), but these curricula tend to be the exception rather than the rule.

This presents a significant disparity between what is recommended by the ASA and the research literature and what is expected in CCSSM. In thinking about these differences it is valuable to look at what research has been done related to the ideas of informal inference in order to see if there are recommendations that can be drawn on to resolve this suggested disparity. Since IIR is about leveraging intuitive and conceptual understandings about data beyond the more arithmetical comparisons of data described in elementary content standards, but short of FSI ideas developed in later grades, it positions IIR as a possible bridge between what is recommended and what is required, making understanding of IIR and its ramifications all the more important. The development of a way of assessing and describing student IIR ability is a foundational need to accomplishing research goals related to the development of IIR and the impact of students’ IIR ability on their learning of FSI.

**Purpose of Current Study**

Research in educational psychology on the development of informal ability in general suggests that a student’s informal ability in a topic does not necessarily improve as content knowledge increases (Zieffler et al., 2008); this disconnect is supported in IIR, specifically by Jacob’s (2013) work with high school students in an AP statistics course, where work with the students’ informal statistical understanding had to be directly addressed. Research in IIR has shown that it is
possible to intentionally develop this ability (Jacob, 2013; Langrall, Nisbet, Mooney, & Janssen, 2011; Makar, Bakker, & Ben-Zvi, 2011), but the extent to which it can be developed, including drawing comparisons between development achieved across different studies and meta-analyses and longitudinal work on IIR abilities at different grade levels, have been hampered by not having a validated means of assessing and describing student IIR abilities. It is also not known what affect students’ IIR has on their learning of FSI, owing partially to not having an instrument to measure or means to describe IIR ability. With this gap in our understanding in mind, this research was guided by the following questions:

1. What developmental descriptions of students’ IIR abilities are suggested by a SOLO taxonomy and how do they help inform the interpretation of the AIIR?

2. How effective are the items in the Assessment of Informal Inferential Reasoning (AIIR) item bank in eliciting and recording students’ IIR abilities, relative to extensive classroom observations and student interviews?
Chapter 2: Review of the Literature

The value of developing the AIIR and associated SOLO adaptation is demonstrable through a look at the type of fundamental open questions in the study of IIR that cannot be addressed without such tools. To that end, this chapter begins by establishing the need for understanding the link between IIR level and the learning of FSI, which is a question that requires a reliable measure of IIR in order to be addressed. Subsequent sections will present the research basis for the AIIR, and relevant research related to the classroom portion of this study.

Link Between Formal and Informal Statistical Reasoning

Informal Inferential Reasoning (IIR), the type of reasoning a student uses when making informal statistical inferences, is an emerging field of study in statistics education. While early work related to IIR such as research focusing on how students interpret graphs goes back several decades (Friel, Curcio, & Bright, 2001), most of the current research builds on work from 2005 where IIR is considered to be multi-faceted. In an attempt to define exactly what IIR is, Zieffler et al. (2008) offered, “the way in which students use their informal statistical knowledge to make arguments to support inferences about unknown populations based on observed samples” (p. 44). Makar and Rubin (2009) expanded this definition by outlining three main characteristics that should all be present when a student is making appropriate use of IIR: including generalizations beyond data, using data as evidence, and using probabilistic language in expressing uncertainty about generalizations.
IIR grew out of a concern that while inferential reasoning is a central objective of statistical reasoning (Paparistodemou & Meletiou-Mavrotheris, 2008; Rubin, Hammerman, & Konold, 2006), it is also one of the facets with which students have a great deal of difficulty (Falk & Greenbaum, 1995; Pratt, Johnston-Wilder, Ainley, & Mason, 2008; Rubin et al., 2006). Rubin et al. (2006) made one of the earliest cases for studying IIR as something beyond just having a graph sense: it allows for “delaying the problem of quantifying probabilities and instead (lays) the conceptual groundwork for inferential reasoning...” which is consistent with the idea that introducing conceptual underpinnings for a topic before the procedural elements is an important consideration in the learning process (Pesek & Kirshner, 2000). While the idea is well supported theoretically, there is a need for more empirical studies looking at the influence of IIR on the efficacy of learning formal statistical inference. The need is particularly pronounced in middle school where GAISE level B tasks (Franklin et al., 2005) are present in the CCSSM, without being supported at lower grades by a learning trajectory that incorporates level A tasks (including development of student IIR ability).

In addition to providing the conceptual underpinnings to formal inference, IIR encompasses many of the data-handling skills that are required in work settings (such as monitoring readouts in manufacturing processes) where practices of formal inference are not required (Bakker, Kent, Derry, Noss, & Hoyles, 2008). This suggests IIR is not only an important subject in its support of formal inference, but also important as a general reasoning tool in real world applications.
Since IIR is valuable both in its presumed influence on formal learning and its role in supporting reasoning about data in situations where formal methods are not needed, it is important to look at how mathematics education in the US currently supports the students’ development of IIR. As an extension of the work done in informal reasoning, Zieffler et al. (2008) note that IIR is unlikely to develop with “maturation, education, or life experience” (p. 44), and that it must therefore be intentionally developed. The CCSSM document overtly mentions developing informal notions of inference in only one place: 7th-grade students are to begin developing informal notions of inference related to comparing two populations (NGACBP & CCSSO, 2010, p. 47). CCSSM only requires arithmetic comparisons of data based on graphs prior to middle school and only deals with formal inference after middle school, resulting in the seventh-grade standard referencing informal inference as the only case where IIR is developed before FSI. This progression of statistical reasoning appears to be based on the notion that students incidentally develop their IIR abilities, which has been shown not to be the case (Zieffler et al., 2008).

Initial work in the more general field of informal reasoning found that increasing students’ content knowledge was insufficient to increase their informal reasoning ability in that content (Zieffler et al., 2008); however, Paparistodemou and Meletiou-Mavrotheris (2008) found that using inquiry methods with a conceptual focus during a four-week course of instruction did produce gains in third
grade students’ abilities to successfully complete ISI tasks when the development of IIR was the primary focus.

The importance of IIR, both for its own sake and for the anticipated benefits in learning FSI, along with knowing that it must be intentionally developed, strongly suggest that there is a need to look at the interplay between the development of IIR and the development of formal statistical reasoning. Barriers to doing this sort of work include the lack of an instrument to measure and a way to describe students’ IIR ability. Initial work to address these shortcomings was begun in a pilot study conducted by this researcher that established the feasibility of a tool to measure IIR, the Assessment for Informal Inferential Reasoning (AIIR). The pilot study also resulted in a preliminary adaptation of the Structure of Observed Learning Outcomes (SOLO) taxonomy designed specifically for describing students’ IIR abilities. A bank of validated items was developed for the AIIR for assessing students’ IIR and the SOLO descriptors were further tested and refined in this dissertation study. Supporting research for the AIIR and the SOLO taxonomy will be presented in the next section with the pilot study results reported in Chapter 3.
### Table 1
**Assessment framework from Zieffler et al. (2008)**

<table>
<thead>
<tr>
<th>Type of Task</th>
<th>IIR Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make judgments or predictions</td>
<td>Use or integrate prior knowledge</td>
</tr>
<tr>
<td>Estimate and draw a population graph</td>
<td>Predict characteristics of a population (shape, center, spread) that are represented in student-constructed graphs</td>
</tr>
<tr>
<td>Compare two samples of data</td>
<td>Judge whether there is a difference between two populations; based on similarities or differences in the samples of data</td>
</tr>
<tr>
<td>Judge between two competing models</td>
<td>Judge whether sample data provide more support for one model than the other</td>
</tr>
</tbody>
</table>

### AIIR Construction

Zieffler et al. (2008) developed a conceptual framework where they present three types of tasks that can be used in assessing or supporting students’ IIR and three components of IIR that can be used to analyze those tasks. The three question types and the recommendations from their analysis were used directly in the creation of the AIIR items. Zieffler et al.’s (2008) task type framework (see Table 1)
was used as a framework for the study, but is also presented here, as this framework and related literature are fundamental to the motivation and execution of this study. This framework and key examples from the literature are presented in the following subsections, with the first, second, and third task types corresponding to the first, second, and third AIIR tasks, respectively.

**Estimate and draw a population graph.** The first task type focuses on estimating and drawing a population graph based on a sample, a task modeled after Bakker’s (2004) idea of growing a sample (Zieffler et al., 2008). The growing-a-sample study had three key phases (Bakker, 2004). The first phase consisted of having eighth-grade students predict how a graph of students’ weights for a sample of size ten would look, and then giving them three sets of real data to compare their predictions and make some basic observations about the differences. During the second phase, students were instructed to construct predictions of samples of increasingly larger size (27 then 67). In the third phase students were asked to predict what the graph would be for the entire city with the intention that they would “stimulate students to use continuous shapes and dynamically relate samples to populations without making the distinction between samples and populations explicitly yet” (Bakker, 2004). This third phase attends to the facet of the IIR definition that discusses how students should be able to predict possible characteristics about populations based on samples.

While Bakker’s (2004) original task sequence was for a more cyclic instructional design, Zieffler, delMas, Garfield, and Gould (2007) developed a single
task based on that instructional design in order to assess students’ IIR at the college level. Similar to the Bakker instructional sequence, the task presented a sample of ten students’ test scores from a psychology class of population size 1000. Students were asked to fill in an additional 25 data points along with an explanation of their reasoning, with a follow-up question asking for an outline of what the overall scores looked like supported by an explanation. This more condensed task maintains the salient points of Bakker’s original instructional design while maintaining the potential for eliciting aspects of IIR outlined in the framework.

Zieffler et al. (2008) describe how the three components of IIR would be analyzed for this item type. Analysis would attend to the three components by looking at how well students made judgments or predictions concerning characteristics of the population graph, looking at their application of prior knowledge (such as using previously acquired statistical language; for example, use of the term skewed-ness), and requiring an explanation on the choices students made. In addition, attention would be paid to the extent to which students used probabilistic language in their explanations.

**Compare two or more samples to infer differences between populations.** The second task type deals with determining whether or not two samples provide sufficient evidence of real differences in their respective populations. The first sample problem type is from research by Watson and Moritz (1999). Their study entailed examining student reasoning about samples of unequal size in grades 3, 5, 6, 7, and 9. The students were presented with the pair of graphs
where there were equal sample sizes in order to ensure they understood how to read the graphs, then subsequently shown a set of four graphs where sample sizes were the same, and finally, shown a set of four graphs where sample sizes were varied. The task goal was primarily to see how the students would determine which of the four classrooms in the task context did better than the others. They found that proportional reasoning became a more prominent strategy as students became older (Watson & Moritz, 1999), though at the time they attributed it to maturation, and the analysis focused on problem solving rather than IIR (which would not become an area of study for another six years). Also notable in this study was that analyses of student responses were done using the SOLO taxonomy, which will be further detailed in a subsequent section of this study.

Zieffler et al. (2008) noted that a critical shortcoming in these questions as an assessment of IIR is that the students were not asked to speak about the possible differences in the larger populations from which the samples were drawn; they were only asked about differences between the students represented in the graphs. This shortcoming is also reflected in the CCSSM Data and Measurement strand in K-5 where students are only asked to draw arithmetic comparisons of sample data (NGABP & CCSSO, 2010).

Again, Zieffler et al. (2008) detail three important pieces that should be looked at in an analysis: making judgments about differences and similarities in populations based on samples, incorporating prior knowledge in the comparisons, and justifying their responses.
Judge between two competing models or statements. Zieffler et al. (2008) present the third task type in the form of two different types of problems. The first problem type focused on evaluating whether or not a statement is true based on sample data and the second required the students to determine whether or not aspects of a graph were due to population characteristics or due to chance.

The first question type is exemplified by a task called Schoolopoly, which first appeared in Stohl and Tarr (2002). On the final three days of a twelve-day instructional sequence focusing on probability utilizing the Probability Explorer software package, 6th-grade students were given the situation of evaluating the fairness of dice in preparation for ordering dice to accompany a board game that their school was hypothetically creating (Stohl & Tarr, 2002). Similar to tasks from the previous nine days, students were asked to run simulations on the computer that would generate both the data of how many of a given response occurred and a frequency plot for the results. Each student pair generated data from increasingly many trials in order to help them decide whether or not each result, when a single die was thrown, was equally likely.

Other examples of this question type are from Konold (2005) and Rubin et al. (2006) who adapted it from Konold. Since the only change that Rubin et al. made in implementation was the substitution of the company name and product type, the original task from Konold (2005) will be described here as it includes more details about intention and implementation. Students are given the context of a robbery at a yo-yo factory, where the suspect has an alibi for a few hours overnight. In order to
see if the suspect’s alibi is valid, students must determine when the break-in occurred by determining when the factory experienced a short power outage resulting in subsequent decreased average production as measured over two-minute intervals (Konold, 2005). Data are provided in a pre-made TinkerPlots (Konold & Miller, 2005) file. Students are first asked questions to clarify their understanding of the data representation and then tasked with finding a representation that helps them solve the problem, along with determining whether or not the suspect’s alibi would prohibit him from performing the robbery. The students must therefore decide whether or not the power outage occurred during the time provided in the suspect’s alibi or not. Like the previous problem, the students evaluate the validity of a statement by analyzing data that have some variance (yo-yo production varied across the two-minute time periods), and take multiple class sessions to determine a solution.

The second question type for judging between two competing models has students look at a graph and determine whether a graph of a sample shows an actual difference in the performance on a test of a sub-population or if the discrepancy between the graphs was due to chance (Zieffler, Garfield, Delmas, & Gould, 2007). The task (done as part of a one-on-one interview with college students in their first statistics course) starts by presenting the student with the population graph of student test scores with the mean score provided. The student is then told that a randomly selected group of students was assigned to a review session with a teaching assistant and the graph for that sample is shown. The
interviewer then asks (within a larger series of questions) whether or not the difference was due to the treatment or due to chance.

The recommended analysis of students’ responses focuses on whether one of the models proposed by the task is more appropriate than the other based on the student’s use of data, the extent to which intuitive or prior knowledge was used in arriving at the conclusion, and the rationale for which the model was chosen.

**Context.** GAISE (Franklin et al., 2005) identifies context as the piece of statistics that provides meaning to patterns. One of the key differences cited between statistics and mathematics is that while in mathematics the context obscures the structure, in statistics the context is intrinsically linked to the data and meaning to the structure (Cobb & Moore, 1997). In statistical inference in general, a student’s preconceptions about a context can have a profound influence on their ability to draw appropriate inferences (Ben-Zvi, 2005), especially when the data support conclusions that are contradictory to their preconceptions (Masnick, Klahr, & Morris, 2007; Zieffler et al., 2007). In this way we can see in the broader view that statistically meaningful tasks must make use of context, and that the nature of the context itself is very important.

An unexpected caveat to the intractability of context with data comes from Wild and Pfannkuch (1999), who took a more holistic view of learning IIR through implementation of the inquiry cycle, that of Problem, Planning, Data collection, Analysis of data, and Conclusion (PPDAC). They noted that while context was vital in the development of core concepts and worked in conjunction with PPDAC to
enculturate students into sound statistical practices, there is value in
decontextualizing statistical concepts periodically to facilitate abstraction
(Pfannkuch, 2011). This seems to contradict Cobb and Moore’s (1997) statement
that statistical data are intrinsically and intractably related to context. Pfannkuch
(2011) addressed this by stating:

...that pattern and data-context are inextricably linked appears to be true
when considering the enquiry cycle; however for the building and using of
statistical concepts such as sampling variability, the researcher argues that
the interplay between pattern and data-context may need to be periodically
disconnected in order for students to focus on and consider the patterns that
can be generated by chance alone (p. 44).

Data context here refers to the situation or story out of which the data arose,
which corresponds to the more common use of the word context in mathematics.
Here we see that the recommendation is not in fact for the learning that occurs
through the enquiry cycle, but for the occasional disconnect in order to avoid
contradicting preconceptions in some specific areas of statistics. One example of this
would be that, even though a sample of a given population is skewed, the means of
multiple samples taken as a data set could be normal; in this case the abstraction
from sample graph to a graph of sample means may represent an abstraction that
students find prohibitively difficult if they first encounter it in context. As the
suggestion of Wild and Pfannkuch (1999) advocates the periodic removal of context
in instruction, and only occasionally for specific topics, removal of context was deemed to be unnecessary in this study.

The above research informed the development of items for the AIIR instrument in that context is shown throughout the literature as being a key aspect of the statistical process. Since students’ abilities to draw ISIs are limited if they don’t have an understanding of the context, care must be taken in task creation not to artificially limit the level of understanding that can be elicited from a student by inappropriate choices of context. Further details on how context impacted item creation are presented in Chapter 5.

**Structure of Observed Learning Outcomes (SOLO)**

Research suggests that statistical literacy in general is a hierarchical construct (Watson & Callingham, 2003). One way of analyzing learning that combines cognitive theories of learning with the idea of learning as a hierarchical construct is the SOLO taxonomy (Shaughnessy, 2007). This analysis framework has seen a great deal of attention recently, particularly in statistics research. An overview of the SOLO taxonomy is presented here and how it relates to the current study. Biggs and Collis introduced the SOLO taxonomy in their work, *Evaluating the Quality of Learning* (1982). The SOLO taxonomy is a general hierarchical model of cognitive development in which a student progresses through five modes of thinking: “sensory-motor, ikonic(sic) (images), concrete symbolic, formal, and post formal” (Biggs & Collis, 1982). Because these modes are developmental in nature, the SOLO taxonomy is often referred to as a neo-Piagetian model. These modes of
thinking operate very similarly to Piaget’s modes, and since they describe learning at a much more general level, they are not the facets of the SOLO taxonomy used to analyze student thinking; that work is characterized by cycles within the modes.

The cycles represent progress in a subjects’ ability to reason with increased complexity about a topic, and are comprised of the following levels: Pre-structural (P), Uni-structural (U), Multi-structural (M), Relational (R), and Abstract/Extended Abstract (A). The basic idea, as illustrated in Figure 1 from a writing task implementation (Groth, 2012), is that at P the student has no connection to a learning outcome, at U the student recognizes a single or one-way connection, and at M the student recognizes multiple connections. At the R level, students can coordinate their multiple ways of approaching a learning objective, and at A they can make abstract statements about the objective (Watson & Moritz, 2000). In looking at the development of a conception, P is the beginning point and the A is the ending point; however, there may be more complex learning occurring than can be accommodated by a single cycle of these levels. When this happens the U-M-R cycle can be iterated as needed with what would be the A for the first cycle becoming the U for the next cycle. This would give the general structure (in the case of a two-cycle SOLO-based framework) of P-U₁-M₁-R₁-U₂-M₂-R₂-A. Note that when multiple cycles are incorporated into the framework, the common shorthand is to use the first letter of the level followed by a numerical subscript denoting the cycle, so R₃ would be the relational level in the third U-M-R cycle. These types of two-cycles were used in various studies (see Reading & Reid (2006), Watson & Moritz (2000)), with Reading
and Reid suggesting that a two-cycle SOLO-based framework be used to describe the development of IIR with the first cycle about more naïve usage and the second more about formal usage (Zieffler et al., 2008).

Figure 1. SOLO taxonomy visualization from Groth (2012, p. 31)

Explanations in IIR

Student explanations have been considered important pieces in mathematics education. Research on discourse has valued student explanations for a variety of reasons, including the way teachers can use explanations (e.g., Lampert, 1990), the insight they lend about the student’s thought process (in the propensity for analyzing student explanations as part of the data analysis in empirical studies), and
opportunities for students to examine and learn from peer explanations (e.g., Yackel, Cobb, & Wood, 1991). More recently, researchers have begun looking at the unique and fundamental role explanation plays in statistics education.

Lipton (2004) viewed explanations as delivering causal knowledge, lending this view to inferential thinking. He describes the link largely in terms of scientific exploration, but promotes the idea that explanations are critical when using inference because the result of inference is not firm truth; it’s looking for the most plausible explanation for available data. This line of reasoning carries easily into statistical inference in particular when one considers that statistics is more about providing strong support for a conclusion rather than irrefutable proof of that conclusion. In this way, the uncertainty inherent in statistical questions and findings makes explanations an integral piece of statistics education. Further, student explanations are critical components used in the SOLO level and cycle descriptions.

The SOLO taxonomy focuses on identifying and categorizing student reasoning structures. While inferences can be made about student reasoning based off of conclusions the students draw and the way in which students construct their graphs, explanations have the potential for providing more clear evidence of the learning structures that a student is using when confronted with a task.

This study used student explanations during the course of the observed unit as a way of focusing on student IIR structures. By focusing on eliciting student explanations in classroom discussions, small group work, written student work, and
follow-up interviews, a sufficiently rich data set emerged for the analysis portion of the study.

**Summary**

The importance of IIR is shown in the literature to be due to both its potential impact on the learning of FSI in school (Rubin et al., 2006), and its importance in students’ personal and professional pursuits (Bakker, Kent, Derry, Noss, & Hoyles, 2008) after leaving school. The literature also shows that this type of reasoning doesn’t occur with maturation, but that it needs to be intentionally developed (Zieffler et al., 2008). This showed that there was a real need to understand how IIR is developed and how it influences the learning of FSI, but these topics have only recently begun to be looked at, owing in part to a need for a way of eliciting and describing students’ IIR abilities.

Some work has been done in thinking about how to approach assessing students’ IIR. Zieffler et al. (2008) identified three types of IIR to be assessed (Estimating a Population Graph (EPG), Comparing Two Samples (CTS), and Choosing Between Models (CBM)), along with guidelines for constructing tasks. However, Zieffler et al. (2008) do not construct such an assessment or refine their task criteria empirically.

Furthermore, once an assessment is constructed and refined based on Zieffler et al.’s (2008) framework, there is only the suggestion to use a 2-cycle SOLO taxonomy adaptation for the analysis of students’ IIR. There are examples of the SOLO taxonomy being adapted for similar situations (e.g., Watson & Moritz, 1999),
but nothing that could be applied to IIR directly, particularly the three task types identified by Zieffler et al. (2008).

This chapter has argued the importance of IIR in general, and specifically the need for clearer understandings of how students develop their IIR and leverage it in the learning of FSI. Previous studies indicate that there is a need for a way of assessing and categorizing students’ IIR abilities. The literature provides the background information needed to begin work on such an instrument, along with a starting point for creating a framework for describing students’ IIR. Chapter 3 presents the methodology used in this study to address these concerns.
Chapter 3: Methodology

The review of the literature presented in the preceding chapter provided support for the current research, along with highlighting the gaps in the literature that the research questions address. This chapter will present how the research questions were addressed. The theoretical frameworks are presented first, followed by results from a pilot study, and ending with the data collection and analysis designs. As a reminder the research questions are:

1. What developmental descriptions of students’ IIR abilities are suggested by a SOLO taxonomy and how do they help inform the interpretation of the AIIR?
2. How effective are the items in the Assessment of Informal Inferential Reasoning (AIIR) item bank in eliciting and recording students’ IIR abilities, relative to extensive classroom observations and student interviews?

Frameworks

The frameworks used in this study consist of Zieffler et al.’s (2008) framework for designing items to assess IIR (presented as part of Chapter 2), and an adaptation of the SOLO taxonomy that was used to describe students’ IIR ability.

As mentioned in the literature review, the SOLO taxonomy is a hierarchical cognitive model (Shaughnessy, 2007) that describes and categorizes ability levels according to the types of reasoning structures students are observed using. The sophistication of the reasoning structures determines the ‘cycle’ they belong to, while the number and use of the structures determine the students’ level within that
cycle. This study used a version of the SOLO taxonomy developed in a pilot study and adapted specifically for use in describing students’ IIR ability levels.

In the pilot study, students completed an instrument designed to assess their IIR ability based on recommendations by Zieffler et al. (2008); using a 2-cycle adaptation of the SOLO taxonomy to interpret the results was also included in their recommendations, though they did not create the adaptation themselves. An initial adaptation to the SOLO taxonomy was developed in the pilot study, but due to the small sample size, limited range of responses, and single grade level at which it was tested, further work remained to be done in order to ensure the adaptation was reasonable and useful. More details about the results of the pilot will be discussed in the next section, but it is important to note here that as a result of those findings a third cycle was included, incorporating the three primary pieces of the IIR definition of IIR to define the SOLO cycles. The SOLO cycle definitions that were informed by the previously reviewed literature and informed by the results of the pilot study are shown in Table 2.

The cycle definitions are used to determine the level of the student reasoning structure. Once the structures are identified, the student’s ability is categorized to be in the highest cycle where a structure for that student appears. The student is classified within that cycle by looking at only the structures in the most advanced cycle as follows:

- one structure: Uni-structural (U);
- multiple structures used independently: Multi-structural (M);
• multiple structures used in a coordinated way: Relational (R).

Table 2
SOLO adaptation cycle definitions developed during the pilot study

<table>
<thead>
<tr>
<th>Cycle name</th>
<th>Make judgments and predictions</th>
<th>Use or integrate prior knowledge</th>
<th>Articulate evidence-based arguments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-IIR (Cycle-1)</td>
<td>Judgments and predictions don’t allow for variability (specifically sampling variability)</td>
<td>Personal experience overrules data and context.</td>
<td>No probabilistic language, evidence provided irrelevant to specific context.</td>
</tr>
<tr>
<td>Naïve IIR (Cycle 2)</td>
<td>Allows for uncertainty, draws flawed conclusions</td>
<td>Over-reliance on context over data or vice-versa.</td>
<td>Presence of probabilistic language, flawed or incomplete mathematical justifications.</td>
</tr>
<tr>
<td>Appropriate IIR (Cycle-3)</td>
<td>Allows for uncertainty, draws appropriate conclusions</td>
<td>Considers both data and context together when drawing informal inferences; may start to integrate appropriate mathematics</td>
<td>Presence of probabilistic language, and may begin using statistical measures informally or incompletely.</td>
</tr>
</tbody>
</table>

If a student performs below the expectation in the first cycle he/she is classified as Pre-structural (P), and if a student performs beyond the expectation of the third cycle he/she is classified as Abstract/Extended Abstract (A). As was mentioned in the literature review, common shorthand for SOLO classifications is to use the first letter of the structure type with the subscript indicating the cycle. Work by a student named Reupert (a pseudonym) in the pilot study provides an example of how these cycles and structures were used in the analysis. Reupert was
given a task of taking a graph of a sample and predicting where several additional points would fall on that graph. Reupert used a common cycle-1 strategy where he tried to add the same number of points to each existing data value, labeled a linear-increase strategy. When there weren’t enough points to go around a typical student with this strategy leaves the points off of the higher values (due to working left to right), but he excluded the extra point from the mode. This suggested that he was combining the linear increase strategy with a second cycle-1 strategy where the student tries to make the frequencies as even as possible. This coordination of two strategies from a single cycle is a Relational strategy; since the Relational strategy was cycle-1 thinking, the SOLO code for this strategy is R₁. More details on analysis and coding follow in a subsequent section.

**Pilot Study Results**

The initial work and revisions on the SOLO taxonomy and the AIIR were done in the pilot study. Middle school students were selected to represent a range of general mathematical abilities with the intent of capturing student thought at varying levels of IIR. Middle school students were classified by the cooperating teacher as having high, medium or low mathematics ability. Striated sampling was used to equally represent students from each ability group.

Based on conversations with the cooperating teacher concerning the school’s alignment efforts with the CCSSM (NGACBP & CCSSO, 2010) and the school’s curriculum, the basic assumption of the pilot study was that the students would exhibit reasoning primarily in the Naïve IIR cycle, with the high-achieving students
potentially exhibiting some structure(s) matching the descriptors in higher levels. This basic assumption, however, proved false; the vast majority of observed student-reasoning structures evident in both the interview and the AIIR were classified in the pre-structural category of the Naïve cycle. This made the analysis and comparison based on Zieffler et al.’s (2008) proposed two-cycle SOLO adaptation somewhat trivial in the sense that students were pre-structural on both pieces so there was a superficial appearance of commensurability. While the grain size on the adaptation was likely sufficient to detect differences in students at the Naïve and Appropriate level, the students in the study did not fall into either of those levels.

In light of this shortcoming, data analysis had to instead begin by defining a new level in the SOLO taxonomy, which was labeled the “Pre-IIR” cycle. The interviews were recorded using LiveScribe Sky pens (Livescribe, 2012), which create a video file of the student’s writing synced with the audio from the interview; since an open-coding method was utilized, the interviews were analyzed primarily in transcript form rather than only from the LiveScribe video. Transcriptions also allowed for more capacity to create and collapse codes rapidly and flexibly as compared to only using the LiveScribe video.

The AIIR, based on Zieffler et al.’s (2008) framework (Figure 1), was shown to be fundamentally sound, requiring only minimal changes to the presentation of the problems, such as creating sub-questions, and providing additional, separate space for explanations rather than asking for the response and explanation in the
same space. The SOLO taxonomy was found to be at the appropriate grain size, but not broad enough in scope in order to account for the nature of reasoning evidenced at the middle school level. This resulted in a re-working of the SOLO taxonomy adaptation cycles, and a classification for the structures in the cycles, though only the first cycle was verified during the course of the experiment.

The pilot study produced the initial versions of the AIIR (see Appendix A) and an adaptation of the SOLO taxonomy used in the dissertation study. The results, though preliminary, provided a sufficient basis to begin work on answering the research questions posed for this study.

**Research Design**

The initial piece of this dissertation study involved creating nine new tasks for the AIIR item bank in order to create a more flexible instrument. This development utilized the task type framework, research on context to inform task creation, and pilot study results to complete additional item creation prior to the first AIIR administration date.

Data collection for this study was conducted at a rural midwest middle school (grades 6-8) with approximately 900 students, 11% minority students, 16% qualified for the free or reduced-price lunch program. The school uses *Connected Mathematics Project 2* (CMP2) (Lappan, Fey, Fitzgerald, Friel, & Phillips, 2006) as their mathematics textbook. Participants in this study were Mr. Oak and Ms. Townsend (pseudonyms), the two cooperating teachers, and their (roughly 180)
eighth-grade students. The research began by obtaining signed consent forms from both teachers, as per the HSIRB (#13-05-27).

Both teachers were involved in developing a six-day unit on inferring population parameters from representative samples. Unit construction took place during half-hour, twice-weekly meetings during the first five weeks of the school semester. Mr. Oak presented a hypothetical learning trajectory based on his experience with the students and the learning objectives he needed to accomplish to support student learning in a subsequent unit that focused solely on informally comparing two samples. In the following weeks the focus was on adapting items from IIR research literature (such as the activities found in Gil and Ben-Zvi (2011), Pfannkuch (2011), and Watson (2008) etc.) and items from their CMP2 (Lappan et al., 2006) textbooks to the learning trajectory and objectives.

The remaining teacher responsibilities were split, with Ms. Townsend facilitating the interviews for the initial co-measurability check, and Mr. Oak leading the classes containing the observed student groups and arranging the interviews for the final co-measurability check. Co-measurability was used to compare students’ SOLO levels as determined by the AIIRs with their SOLO levels as determined by interviews and, in the case of the final co-measurability check, classroom observations. Co-measurability addresses the degree to which the students’ IIR ability can be identified with the AIIR instrument as opposed to more in-depth but prohibitively time-consuming methods.
The school year began in early September. Students were given assent forms to be filled out in class and consent forms to be taken home and filled out by their parents/guardians. The first of two administrations of the AIIR was given to the teachers' classes along with their usual assessments of student prior knowledge in the third week of school (the second administration was given after the unit completion on informal statistical inference). This was done for all students as the teachers used the data to inform their teaching, but only the tests of students for whom consent and assent were attained were made available for coding and analysis. The nine new AIIR items were used to create three new versions of the AIIR (A, B, and C) which were administered to students in Ms. Townsend’s first, second, and third classes, respectively; students in Mr. Oak’s class all took the piloted version of the AIIR (Appendix A). Students had part of one class period, approximately 30 minutes, to complete the AIIR assessment.

Eight of Ms. Townsend’s students were selected to participate in follow-up interviews about the AIIR, with interviews occurring before the unit. Ms. Townsend provided a list of her students organized in three categories based off of their general mathematics proficiency (one being lowest proficiency level and three being the highest), and ordered so that students she believed to be the most talkative were at the top of the list. Students were selected to participate in semi-structured interviews exploring their IIR ability, which were recorded using LiveScribe Sky pens (Livescribe, 2012). Interviewees were selected in order to include students from at least two different ability levels from each of the three classes, with
consideration given to selecting students identified as most likely to be talkative during an interview. Student reasoning structures in the AIIR and interview were coded for each student with disparate results used to inform task adjustments prior to the second implementation. Open coding (Creswell, 2012) is described as a way of approaching a topic about which little is known, and was used in this study as it allowed for key elements to be identified based upon their recurrence. These key elements were then placed into descriptive categories that were used to form the codes used in the research. HyperRESEARCH (v. 3.5.4) was chosen as the qualitative data analysis tool due to its ability to incorporate multiple data source types, the versatility in the coding interface that allowed for open coding with subsequent refinements, and the researcher’s familiarity with the software from previous work.

The initial AIIR administration, along with the follow-up interviews, was concluded by the end of the fifth week of school. The initial co-measurability check, which consisted of coding the AIIRs and interviews and checking for commensurability, along with item augmentations based on the results of the co-measurability check occurred during the sixth and seventh week of school. The AIIR data from all of the students were used along with classroom observations from the unit on FSI in order to investigate the SOLO taxonomy adaptation.

The unit (as described in Chapter 4) was begun in the sixth week of the semester; both teachers began the unit on the same day and kept the same schedule. Two of Mr. Oak’s classes were selected for observation based upon Mr. Oak’s recommendation, with the intention of observing the classes whose students are
most willing to share their explanations. The groups within the classes were chosen based upon a review of the results of their pre-AIIR in order to include students from the widest cross-section of IIR ability possible in the final co-measurability check. Student work, LiveScribe (2012) files (from one pen in each group) and class video were collected during the course of the unit.

At the end of the unit, students took the second administration of the AIIR. The twelve items being analyzed (three from the pilot and nine from the dissertation study) were re-organized to create four new versions of the AIIR. Students were assigned a version of the AIIR to take in the second administration such that no two members of a group took the same version of the test. Each of the sixteen students that participated in an observed group was then interviewed in order to elicit richer explanations, which in turn provide further information about their IIR ability.

All data were imported into HyperRESEARCH (v. 3.5.4), and student reasoning structures were then coded and used to assign a SOLO level for the student. The observed and assessed SOLO levels were compared in order to determine co-measurability.

A summary of student involvement is provided in Figure 2, and a brief overview of the data sources and method of analysis for each question is also included in Table 3. The following is a summary of the timeline for the data collection:

- week 1-3, obtain signed consent forms from participants;
- weeks 1-5, plan unit with cooperating teachers;
- week 3, teachers conduct pre-IIR;
- weeks 3-4, conduct initial co-measurability check interviews;
- weeks 5-7, analyze data and modify AIIR items accordingly;
- weeks 6-8, Observe unit, recording four student groups;
- week 8, Second IIR implementation;
- weeks 8-10, Final co-measurability check interviews.

*Figure 2. Sample organization*
Research question two was concerned with refining the SOLO adaptation.

The proposed SOLO taxonomy, however, required extensive re-working from the pilot study since it was only able to capture a limited range of student thinking, so the revision/refinement process here incorporated a larger sample in order to both ensure code saturation and increase the possibility of encountering a broader range of student IIR ability. The AIIR was administered to 180 students as a part of normal classroom practices, with 150 of the AIIRs available to be used in this research. Open coding (Creswell, 2012) was used on student responses to the instrument along with interview data from the eight students in Ms. Townsend’s classes and interview and classroom observation data from Mr. Oak’s classes, to identify any reasoning structure types that were not evidenced in the pilot study. Additional structures that emerged from the coding were used to create more appropriate cycle definitions for a broader range of IIR ability levels.

<table>
<thead>
<tr>
<th>Research Question</th>
<th>Sample</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOLO modification</td>
<td>All students</td>
<td>Coding student reasoning structures</td>
</tr>
<tr>
<td>AIIR item initial</td>
<td>AIIR Interview Group</td>
<td>Comparing reasoning structures between interview and AIIR</td>
</tr>
<tr>
<td>co-measurability check</td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIIR item final co-</td>
<td>Small groups 1-4</td>
<td>Comparing SOLO level between comprehensive student analysis and AIIR results</td>
</tr>
<tr>
<td>measurability check</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3
Research question/sample/methodology brief overview
Summary

This chapter presented the methodology used to address the research questions. Zieffler et al.’s (2008) framework for an IIR assessment was used to construct the AIIR, while an adaptation of the SOLO taxonomy was adapted to analyze and interpret student responses on the AIIR. The first attempt at building the assessment and adapting the SOLO taxonomy were made during the pilot study, and the lessons learned there shaped the starting point for the current research.

Once the initial additional design pieces were completed, the first co-measurability check was conducted. This check was done by analyzing student interviews and by comparing the outcomes to the analysis of the student AIIRs in order to determine whether or not the AIIR was sufficiently eliciting student thought. The results of the initial co-measurability check were used to inform further AIIR and SOLO modifications.

After modifications were made, a final co-measurability check was completed. The final co-measurability check incorporated additional data sources (classroom observations, student work, and classroom artifacts) along with the interviews, and the results of that analysis were compared to analysis of the students' AIIRs.

The next chapter will briefly discuss the unit the students participated in and discuss in more detail the initial item creation and the adaptations to the SOLO taxonomy made throughout the course of the research.
Chapter 4: Unit and Task Design and SOLO Modification

This chapter presents the two elements specifically designed for this study, 1) the instructional unit created with the two cooperating teachers, and 2) additional items for the Assessment of Informal Inferential Reasoning (AIIR) item bank. It also contains a discussion of the analysis and modifications to the SOLO taxonomy used in the pilot study. These are foundational to understanding the analysis and interpretation of the co-measurability check cycles. This analysis is presented separately in this chapter because, while the co-measurability check cycles are best understood when presented chronologically, the SOLO adaptation process was ongoing and incorporated data from every aspect of the study. Since this process was ongoing, it did not fit in at a specific point in the chronological progression of the research so it is presented here to clarify the analysis.

Design of the Teaching Unit and Additional AIIR Items

Based on results from the pilot study, this study called for additional resources to be developed to address the research questions. These resources included developing, 1) a unit of study for the students that more closely aligned with the statistical reasoning goals of the study, and 2) additional assessment items to expand the AIIR item bank. The researcher designed the additional assessment items and, in collaboration with Mr. Oak, produced the unit of study.

**Unit design.** The instructional unit was designed to focus primarily on CCSSM standard 7.SP.2 related to generalizing from a sample to a population. The unit was originally conceived as focusing on leveraging students’ IIR into more
formal statistical inference. After reviewing the CCSSM alignment progress documents that were used by teachers in transitioning to the new standards, considering the IIR levels of students in the pilot, and reviewing the AIIR pre-tests of the students in the classes, we decided that the students were not equipped with the requisite prior knowledge needed to successfully engage in the IIR items. Thus, we designed the unit to focus more on developing their IIR and establishing the conceptual underpinnings of statistical inference. This was also seen as a way of maximizing the data set needed to triangulate students’ IIR ability.

The new focus for the unit presented the opportunity to more directly adapt lessons around developing IIR. Based upon the work of Paparistodemou and Meletiou-Mavrotheris (2008), who researched IIR with 3rd-grade students, we designed a similar unit for the 8th-grade students. The statistical investigative cycle (Wild & Pfannkuch, 1999) of problem, planning, data collection, analysis of data, and conclusion, guided the development of the unit. The outline for the six-day unit is presented in Table 4.

At the end of day two, one Tinkerplot file was created for each class using the respective class data collected from the administered survey. The class data constituted a sample of 30 responses. Students in a class worked from the same sample, with six samples being analyzed across the six classes. This analysis will be presented below in the discussion about sampling, with the random sample selection being analogous to randomly selecting 30 participants.
Additional AIIR item creation. One goal of this study was to create an item bank that could be used to generate different versions of the AIIR. These new items were created by incorporating the lessons learned in the pilot study while also introducing more variety of contexts and more variation of student tasks within each task type. The new items for the item bank were created during the course of the three-week period before the cooperating teachers began their fall semester.
Item creation was based on the framework developed by Zieffler et al. (2008), with consideration given to the lessons learned in the pilot study. The categories for the items correspond to the individual investigation foci of the unit: Estimating a Population Graph (EPG), Comparing Two Samples (CTS), and Choosing Between Models (CBM).

Expanding from a single version of the AIIR to an item bank afforded the opportunity to incorporate additional task types suggested by Zieffler et al. (2008) and diverse contexts for each question type. In the CTS tasks, the additional task type designed was one with sets of graph pairs that tasked students with deciding which pair presented the best and worst support for a statement. For instance, students were presented with the context of student scores on a driving test when texting and when driving regularly (Figure 3). Students were asked to order the three sets of graph pairs by how strongly they supported the idea that texting affected driving ability. Two questions of this type were constructed, with one additional task created that was consistent with the piloted AIIR item (see Appendix C).
Zieffler et al. (2008) also provided examples of various CBM question types. Therefore, alternative task types based on the examples cited in that paper were also explored for the CBM category. One item was created that was similar to the pilot study but was presented in a table rather than a graph to allow for students who may have IIR ability but insufficient graph sense to understand graphical data presentations. A second item presented a population graph and a summary value (presented as a ‘typical’ value without claiming to be any specific summary statistic) from a sample, and asked if the populations were similar.

The three new items for each task type were created in order to provide contexts for which students would have varying levels of access. Task context types are presented in Table 5. The task types were selected with the intention of providing some variation in the levels of student context expertise.
These additional items were administered twice, and were taken with classroom observation data, student interviews, and student artifacts to provide the data for further refinement of the SOLO taxonomy. The next section presents the refinement process and results.

Table 5

<table>
<thead>
<tr>
<th>Table 5</th>
<th>Contexts for new AIIR items</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Direct Access</td>
</tr>
<tr>
<td>EPG</td>
<td>Students’ texting habits</td>
</tr>
<tr>
<td>CTS</td>
<td>Drivers Education obstacle course</td>
</tr>
<tr>
<td>CBM</td>
<td>Amusement park</td>
</tr>
</tbody>
</table>

**SOLO Taxonomy Adaptation**

The Structure of Observed Learning Outcomes (SOLO) taxonomy was introduced in Chapter 2, and presented in Chapter 3 as the analytical framework that was used to interpret student responses on the Assessment of Informal Inferential Reasoning (AIIR), as per Reading’s recommendation (Zieffler et al., 2008). Initial adaptations to the SOLO descriptions were made in the pilot study and a three-cycle model was decided on. This section presents the refinements and justifications for those refinements made during this dissertation research.

The purpose of this portion of the research was to further refine the version of the adapted SOLO taxonomy that was developed in the pilot study (see Table 2 on page 29). That initial version provided guidelines for classifying student reasoning structures/strategies. In the original table the variability indicators were:
• Cycle-1: Judgments and predictions don’t allow for variability
• Cycle 2: Allows for uncertainty, draws flawed conclusions.
• Cycle-3: Allows for uncertainty, draws appropriate conclusions.

However, these initial descriptors had the unintended consequence of limiting the range of how students might use variability (recall that the use of variability here refers specifically to sampling variability), which was incompatible with the findings in the initial co-measurability check. The initial descriptors only allowed for a discrete rather than graduated allowance for students’ ability to implement variability into their strategies, which ran counter to what was observed. In the original descriptor, a cycle-2 student would be expected to fully understand the nature and role of variability, but that understanding would then be overridden by some other aspect of their reasoning resulting in flawed conclusions about the item. In this study, cycle-2 students were not found to make flawed conclusions with an allowance for variability, but exhibited limited allowance for variability with conclusions appropriate to their proficiency.

An example of this limited allowance for variability can be seen in Hannibal’s (all names are pseudonyms) response to part A of task CBM3 in the final co-measurability check. In this problem a person is going to an amusement park with some friends, and the amusement park is offering guests the opportunity to win a “quick pass” (which allows them to skip the line for any attraction once) by drawing a ball out of a bag. If an orange ball is drawn, the individual wins a pass. The group of friends wants to see if any line has an advantage, so they count the number of
successes per 15 people in each line. Hannibal is presented with the findings in a
graph shown in Figure 4. He doesn’t appear bothered by the difference of size one
between Lane 1 and Lane 2. This shows that he does have some concept of sampling
variability. However, while he can see that Lane 1 and Lane 2 could have the same
actual probability, he believes that Lane 2 and Lane 5 must have different
probabilities. On the same problem another student was able to identify that the
lanes with sevens could have a probability of 6/15 to 8/15, and that the lane with
nine could have a probability of 8/15 to 10/15, but was adamant that the lanes with
seven successes had lower probabilities for success than the lane with nine
successes. This was defined as a limited-variability strategy because, while the
student did not insist that outcomes must be exactly identical (as would be the case
if the student had no allowance for variability), the student only allowed for very
small amounts of variability in the data, and even the small amount of variability
allowed for was applied inconsistently.

Figure 4. Graph form CBM3 part A

The cycle descriptors for the use of prior knowledge (see Table 2 on page 29)
was expanded to explicitly include context knowledge in the title. While the
descriptors defined in the pilot study were seen as consistent with the findings in this current research, during the initial co-measurability check this category was also found to contribute an important descriptor to the pre-structural level of understanding. For a student whose IIR capacity is in cycle-1 (the Pre-IIR cycle), personal data are seen as completely over-ruling other data and becomes the only factor in answering the problem. For instance, when faced with the previously mentioned problem regarding lines at an amusement park, Larry said that the lanes must all be fair because a business wouldn't treat its customers unfairly. A pre-structural student tends to expand and invent new aspects of a context in order to allow him to arrive at a conclusion (Watson & Moritz, 1999). A hypothetical example of this in the same problem would be if a student said ‘no they aren’t fair, because the park probably wants to see if people will catch on to them not being fair and pick a better lane.’

The probabilistic language indicator for cycle-1 was modified so that it specified the presence of deterministic language rather than simply the absence of probabilistic language. This was done for two reasons, 1) it was noticed that the absence of something was not particularly useful in the coding, and 2) it was desirable to more closely mirror the language in the GAISE framework, particularly when describing student reasoning about variability (Franklin et al., 2005, pp. 11-12). Mathematical justification language at cycle 2 and 3 was replaced with probabilistic language, with the emphasis shifted to how well the probabilistic language was integrated into statements of certainty (when requested) and how
probabilistic language in explanatory statements was reflected in what the student presented as the solution to the problem. The final SOLO level descriptors are shown in Table 6. Examples of the more common student reasoning structures, the SOLO level those structures indicate, and the code that was used for them in this study are shown in Table 7, Table 8, and , for the EPG, CTS, and CBM task types, respectively

Table 6

<table>
<thead>
<tr>
<th>Construct Cycle</th>
<th>Use of Variability</th>
<th>Use of Context</th>
<th>Certainty and argumentation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Structural No cycle</td>
<td>No concept or use of variability.</td>
<td>Modifies context to answer question</td>
<td>Deterministic language</td>
</tr>
<tr>
<td>Pre-IIR (Cycle-1)</td>
<td>Limited conception or use of variability.</td>
<td>Answers based solely in context.</td>
<td>Probabilistic language with over-use of context over data or vice-versa.</td>
</tr>
<tr>
<td>Naive-IIR (Cycle 2)</td>
<td>Limited conception or use of variability.</td>
<td>Over-use of context over data or vice-versa.</td>
<td>Probabilistic language with minimal impact on conclusions</td>
</tr>
</tbody>
</table>

Once a student’s reasoning structures were coded, the corresponding SOLO codes for the structures were identified. Students were seen as reasoning at the highest SOLO code in which they exhibited those reasoning structures.
Table 7  
**SOLO taxonomy examples and codes for EPG question type**

<table>
<thead>
<tr>
<th>SOLO Code</th>
<th>Pre-Structural</th>
<th>Structure example</th>
<th>Structure Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Structural</td>
<td>Student expands/personalizes context to provide answers</td>
<td>Estimating a Population Graph (EPG) Structure example</td>
<td>Personal-context</td>
</tr>
<tr>
<td>Pre-Structural</td>
<td>Student selects numbers from the context and/or problem and combines them arithmetically in an inappropriate or unfocused way</td>
<td>SOLO taxonomy examples and codes for EPG question type</td>
<td>Undirected-arithmetic</td>
</tr>
<tr>
<td>U₁</td>
<td>Student increases all data values by a fixed amount</td>
<td>Linear-growth-vertical-only</td>
<td></td>
</tr>
<tr>
<td>U₁</td>
<td>Student increases all data values to level off the graph</td>
<td>Leveling-off-data</td>
<td></td>
</tr>
<tr>
<td>M₁</td>
<td>Student uses both strategies in U₁ independently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R₁</td>
<td>Student coordinates both U₁ strategies</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U₂</td>
<td>Student doubles existing data value frequencies</td>
<td>Doubling</td>
<td></td>
</tr>
<tr>
<td>U₂</td>
<td>Student increases existing data values such that the general shape of the sample is preserved without including data at new values</td>
<td>Shape-preserving-growth-vertical-only</td>
<td></td>
</tr>
<tr>
<td>M₂</td>
<td>Student uses both strategies in U₂ independently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R₂</td>
<td>Student coordinates the use of a U₂ strategy with consideration to the context</td>
<td></td>
<td></td>
</tr>
<tr>
<td>U₃</td>
<td>Student grows the sample such that the general shape of the sample is preserved both vertically and by including appropriate additional data values</td>
<td>Shape-preserving-growth-vertical-horizontal</td>
<td></td>
</tr>
<tr>
<td>U₃</td>
<td>Student explanations include appropriate statements of uncertainty that reference (explicitly or implicitly) measures of spread or center</td>
<td>Appropriate-variability</td>
<td></td>
</tr>
<tr>
<td>M₃</td>
<td>Student uses both strategies in U₃ independently</td>
<td></td>
<td></td>
</tr>
<tr>
<td>R₃</td>
<td>Student coordinates the use of U₃ strategies</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Table 8

**SOLO taxonomy examples and codes for CTS question type**

<table>
<thead>
<tr>
<th>SOLO code</th>
<th>Comparing Two Samples (CTS) Structure example</th>
<th>Structure Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Structural</td>
<td>Student expands/personalizes context to provide answers</td>
<td>Personal-context</td>
</tr>
<tr>
<td>U₁</td>
<td>Student compares the modes to see which has the greatest frequency</td>
<td>One-irrelevant-aspect</td>
</tr>
<tr>
<td></td>
<td>Student compares outliers only</td>
<td>One-irrelevant-aspect</td>
</tr>
<tr>
<td>M₁</td>
<td>Student uses both strategies in U₁ independently</td>
<td>Two-irrelevant-aspects</td>
</tr>
<tr>
<td>R₁</td>
<td>Student coordinates both U₁ strategies</td>
<td>Two-coordinated-irrelevant-aspects</td>
</tr>
<tr>
<td>U₂</td>
<td>Student compares the location of the modes</td>
<td>One-relevant-aspect</td>
</tr>
<tr>
<td></td>
<td>Student compares the location of the medians</td>
<td>One-relevant-aspect</td>
</tr>
<tr>
<td>M₂</td>
<td>Student uses both strategies in U₂ independently</td>
<td>Two-relevant-aspects</td>
</tr>
<tr>
<td>R₂</td>
<td>Student coordinates the use of a U₂ strategy with consideration to the context</td>
<td>Two-coordinated-relevant-aspects</td>
</tr>
<tr>
<td>U₃</td>
<td>Student compares the modes of the data with consideration to graph shape</td>
<td>One-appropriate-comparison</td>
</tr>
<tr>
<td></td>
<td>Student compares the median of the data with consideration to graph shape</td>
<td>One-appropriate-comparison</td>
</tr>
<tr>
<td>M₃</td>
<td>Student uses both strategies in U₂ independently</td>
<td>Coordinated-appropriate-comparisons</td>
</tr>
<tr>
<td>R₃</td>
<td>Student uses a strategy in U₂ along with an appropriate use of context</td>
<td>Appropriate-comparison-context</td>
</tr>
</tbody>
</table>
Table 9
SOLO taxonomy examples and codes for the CBM question type

<table>
<thead>
<tr>
<th>SOLO code</th>
<th>Choosing Between Models (CBM) Structure example</th>
<th>Structure Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Structural</td>
<td>Student expands/personalizes context to provide answers</td>
<td>Personal-context</td>
</tr>
<tr>
<td>U₁</td>
<td>Student reads the data, but uses his/her own experience with the context to over-rule what can be seen in the data</td>
<td>Experience-trumps-data</td>
</tr>
<tr>
<td></td>
<td>Student’s response shows that the statistic being in the context must match the model of the sample data exactly. For instance, if the statistic being modeled is a typical score of 35, the mode must be exactly 35</td>
<td>No-variability</td>
</tr>
<tr>
<td>M₁</td>
<td>Student uses both strategies in U₁ independently</td>
<td></td>
</tr>
<tr>
<td>R₁</td>
<td>Student coordinates both U₁ strategies</td>
<td></td>
</tr>
<tr>
<td>U₂</td>
<td>Student allows for only small variations between the statistic being modeled and the sample graph. For instance, if the situation being modeled is a fair die and the sample is 60 throws; the student states that a frequency from 9-11 is ok, but 8 or 12 indicates the die is unfair</td>
<td>Limited-variability</td>
</tr>
<tr>
<td></td>
<td>Student expresses that he is uncertain of his result, and when asked what would make him more certain he says he would need to see data for the entire population</td>
<td>Uncertainty-population-needed</td>
</tr>
<tr>
<td>M₂</td>
<td>Student uses both strategies in U₂ independently</td>
<td></td>
</tr>
<tr>
<td>R₂</td>
<td>Student coordinates the use of a U₂ strategy with consideration to the context</td>
<td></td>
</tr>
<tr>
<td>U₃</td>
<td>Student expresses that he is uncertain of his result, and when asked what would make him more certain he says he would like data from a larger sample</td>
<td>Uncertainty-larger-sample</td>
</tr>
<tr>
<td>M₃</td>
<td>Student uses both strategies in U₂ independently</td>
<td></td>
</tr>
<tr>
<td>R₃</td>
<td>Student coordinates the use of U₂ strategies</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 5: AIIR Modification and Co-Measurability Analysis

Since new items were created for the current study, those items needed a check on the co-measurability of the results they produced. This chapter discusses findings from the first co-measurability check for these additional items, the subsequent process used in item modification, and the final item co-measurability check.

For the sections on the co-measurability checks, the SOLO taxonomy will be referenced throughout in terms of levels, cycles, and structure codes. Summaries of these codes can be found in Table 6 through Table 9, presented in Chapter 4.

Initial Co-Measurability Check

Initial Co-Measurability Analysis of AIIR Items and administration of the AIIR was focused on accomplishing two goals. The primary goal was to serve as an initial check of co-measurability for the newly created AIIR items; findings from this initial check informed the modifications made to the items prior to the final AIIR administration. A second goal was to generate data for use in the SOLO modification as was discussed in Chapter 4.

During the third week of school, students completed the initial administration versions of the AIIR. The AIIR item bank used in the initial administration can be found in Appendix B. Item names include the question type (EPG for Estimating a Population Graph, CTS for Comparing Two Samples, and CBM for Choosing Between Models), and the item number, and a letter for the version (P for pilot, I for initial co-measurability version, F for final co-measurability version,
and R for release version). For example, EPG2I is the initial co-measurability check version of the second question in the estimating a population graph task type. Items were organized into three versions of the AIIR, each version including one of each task type. The organization of the items into versions is presented in Table 10; note that the first question in each question type was omitted due to them undergoing an initial co-measurability check in the pilot study.

Table 10

<table>
<thead>
<tr>
<th>Version 1</th>
<th>EPG2I</th>
<th>CTS2I</th>
<th>CBM2I</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version 2</td>
<td>EPG3I</td>
<td>CTS3I</td>
<td>CBM3I</td>
</tr>
<tr>
<td>Version 3</td>
<td>EPG4I</td>
<td>CTS4I</td>
<td>CBM4I</td>
</tr>
</tbody>
</table>

Initial co-measurability of AIIR items was based upon the student interviews conducted two weeks after the initial administration. The process mirrored that done in the pilot study; students took the AIIR, participated in task-based semi-structured interviews, then the results of the AIIR were compared to the results of the interview to see if they were commensurable.

Interview transcripts and written work during the interviews were imported to HyperRESEARCH (v. 3.5.4), identified by the version of the assessment taken by the student, the student’s number, and the origination of the material; for example “V1 S2 Interview Transcript” would be the transcription of the audio file from the interview with student 2, who took version 1 of the AIIR. Two cases were created in HyperRESEARCH (v. 3.5.4) for each student: one had the interview data and the second case had the written responses taken from the classroom administered AIIR.
Recall LiveScribe Sky pens (Livescribe, 2012) were used in the classroom and interview data collection process (see the Research Design section in Chapter 3). Data collection for the interview was hampered by a technical difficulty with the LiveScribe Sky pens (Livescribe, 2012); while the audio was recorded correctly and the pen was able to generate a picture of what the student work looked like, it was unable to sync the audio and the student work as they were being created. Because of this the choice was made to transcribe the entire interview and code the transcriptions and image files in HyperRESEARCH (v. 3.5.4).

Student interviews were coded first. When the interview transcript stated that the student was writing, the relevant piece of writing was coded, then coding resumed on the transcript. For example, during Alexis’ interview, the student was asked to attempt a problem from the AIIR with different parameters (growing the sample by a different amount). When the transcript indicated that Alexis wrote her response, the written response was coded before continuing with the coding of the transcript. These codes were under the student’s ‘interview’ case.

Student AIIRs were coded next. The AIIR was administered with one question type on each page, and were imported to HyperRESEARCH (v. 3.5.4) as three files each containing a single question. These codes were under the student’s ‘AIIR’ case.

Coding was done by looking for two key items: student reasoning structures based on identifiable strategies used by the student and student level indicators indicated by cues in their descriptions and justification that indicated the degree to
which they, for instance, allowed for variability. Student reasoning structures relate directly to the methods that students used to complete items. For example, a common strategy that was used when students needed to double the amount of points on a sample graph was to double the frequencies at the pre-existing data values. Another example occurred when Betsy was presented with a display of 15 students’ test scores and was asked to include 15 additional students’ test scores. She identified that the size of the original data set was equal to the number of data points to be added. Her response is shown in Figure 5, where it can be seen that the set of data she added was identical to the original set of data. This doubling of the data at existing values was considered a reasoning structure, and was coded as doubling. Doubling is considered a structure in the Naïve cycle because while it did result in proportional vertical growth in the graph, it did not allow for new values along the axis. The criteria for determining the cycle for the structure can be found in the discussion about revising the SOLO taxonomy (page 45).

*Figure 5. Example of the doubling strategy Betsy’s pre-AIIR*
By contrast, level indicators are more in line with the definitions of the SOLO cycles in Table 6 on page 49. An example of this would be from the amusement park context (see pp. 46 and 47 for further item description and corresponding graph). In this context a group of friends monitor six lines at an amusement park to see the rate at which prizes are won. Two of the lines resulted in seven successes, three resulted in 8 successes, and one resulted in 9 successes. The student expressed certainty that the line with 9 successes had a higher overall probability of success because “the numbers are all scattered out.” This was evidence of the student not allowing for variability in thinking about the situation, and that at least some of his reasoning about the problem was in the ‘Pre-IIR’ cycle.

After the coding was completed, reports were generated and used to identify the interview SOLO level based upon the interview, and the assessed SOLO level based solely on the AIIR responses. The interviewer based the interview around the AIIR questions and the individual student’s responses. Because of this, if a student provided enough detail in a response, provided enough justification for decisions or used a strategy that could be identified, that aspect of the AIIR was not covered in the interview. This was not seen as problematic because the assessment would only be seen as deficient when it failed to identify reasoning structures that a student was capable of, not when it identified reasoning structures that were not identified in the interview. For this reason, the co-measurability check needed only to identify a deficiency when the assessed SOLO level was lower than the interview SOLO level. Because the only necessity was identifying AIIR deficiencies, and because the length
of the interviews resulted in focusing on clarifying ambiguity in student responses and performing extension tasks, the interview SOLO level was seen as having a lower bound at the assessed SOLO level. Table 11 shows the interview SOLO level and assessed SOLO level for each student on each question type. As a reminder, the letter signifies the three different SOLO levels of Unistructural (U), Multistructural (M), or Relational (R), while the number signifies whether the student is in the Pre-IIR (1), Naïve IIR (2), or Appropriate IIR (3) cycle. In the following discussion the label “Student’s Initial/Aspect” will identify the rows from the table: so Alexis’ work (A) on the question type “Estimating a Population Graph” (EPG) will be written A/EPG.
### Table 11

*Assessed and interview SOLO level results for initial co-measurability check*

<table>
<thead>
<tr>
<th>Version</th>
<th>Student</th>
<th>Aspect</th>
<th>Assessed SOLO Level</th>
<th>Interview SOLO Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alexis</td>
<td>EPG</td>
<td>U₂</td>
<td>U₂</td>
</tr>
<tr>
<td>1</td>
<td>Alexis</td>
<td>CBM</td>
<td>U₂</td>
<td>U₂</td>
</tr>
<tr>
<td>1</td>
<td>Alexis</td>
<td>CTS</td>
<td>U₁</td>
<td>U₂</td>
</tr>
<tr>
<td>1</td>
<td>Betsy</td>
<td>EPG</td>
<td>U₂</td>
<td>U₂</td>
</tr>
<tr>
<td>1</td>
<td>Betsy</td>
<td>CBM</td>
<td>U₁</td>
<td>U₁</td>
</tr>
<tr>
<td>1</td>
<td>Betsy</td>
<td>CTS</td>
<td>U₁</td>
<td>U₂</td>
</tr>
<tr>
<td>2</td>
<td>Chad</td>
<td>EPG</td>
<td>U₂</td>
<td>M₂</td>
</tr>
<tr>
<td>2</td>
<td>Chad</td>
<td>CBM</td>
<td>U₁</td>
<td>U₁</td>
</tr>
<tr>
<td>2</td>
<td>Chad</td>
<td>CTS</td>
<td>U₁</td>
<td>U₁</td>
</tr>
<tr>
<td>2</td>
<td>Devin</td>
<td>EPG</td>
<td>U₂</td>
<td>P</td>
</tr>
<tr>
<td>2</td>
<td>Devin</td>
<td>CBM</td>
<td>U₁</td>
<td>R₁</td>
</tr>
<tr>
<td>2</td>
<td>Devin</td>
<td>CTS</td>
<td>U₁</td>
<td>R₁</td>
</tr>
<tr>
<td>3</td>
<td>Erica</td>
<td>EPG</td>
<td>U₃</td>
<td>U₃</td>
</tr>
<tr>
<td>3</td>
<td>Erica</td>
<td>CBM</td>
<td>P</td>
<td>P</td>
</tr>
<tr>
<td>3</td>
<td>Erica</td>
<td>CTS</td>
<td>R₁</td>
<td>R₁</td>
</tr>
<tr>
<td>3</td>
<td>Frank</td>
<td>EPG</td>
<td>U₂</td>
<td>U₂</td>
</tr>
<tr>
<td>3</td>
<td>Frank</td>
<td>CBM</td>
<td>U₁</td>
<td>U₁</td>
</tr>
<tr>
<td>3</td>
<td>Frank</td>
<td>CTS</td>
<td>R₁</td>
<td>R₁</td>
</tr>
</tbody>
</table>

The following sub-sections will present examples from the initial co-measurability check analysis. The three students presented here were chosen for inclusion in this chapter based on the variety of reasoning levels presented by the students and the opportunities for presentation of student examples for each question type. Analyses for students not selected for this section are included in Appendix C. Each student analysis here will be broken down into three subsections, one for each question type. The analysis will focus on students’ interview SOLO level, and how it compared to their assessed SOLO level for each question type.
Details about the composition of the initial-co-measurability-check AIIRs can be found in Table 10.

EPG1, CBM1, and CTS1 went through initial co-measurability check as part of the pilot study. The items in the initial version are presented in Appendix B.

**Devin EPG3.** Devin was selected as the first student to be presented here in order to show an example of a student whose assessed and interview SOLO levels did not match. During the interview Devin exhibited the same strategy as he did on the AIIR, doubling, when growing a sample of 15 data points to 30 data points.

When asked to estimate the population graph, Devin kept the shape similar to the sample graph, saying “if there was a higher point I made it higher, and if it was lower I made it lower.” He was thinking in terms of stretching existing data upwards (a proportional strategy) rather than linearly increasing which, when taken with his ability to use a doubling strategy, placed his reasoning on the EPG task type at M2 as he was using multiple strategies independently.

Devin’s performance on the AIIR was markedly different. Devin began the AIIR by looking at the question where the context is someone trying to figure out how many text messages other people his age are getting in a typical day. In the context, the person requests that schools in his area have their students email him the number of texts they send. The first task on the assessment related to the context is to “add 15 more points to the graph to represent the new responses,” which Devin responds to by writing “because he got 15 emails and 15/3=5 so I did five dots.” He placed all five of these new “dots” in the 12-15 bin. The normal code
would be undirected-arithmetic, a pre-structural code. In this case it appeared as if it could also be a misunderstanding of the graph and intent of the question rather than a misconception of the statistics, in that Devin saw that the person in the context (Evan) received 15 emails, therefore the new piece of data was that Devin received 15 emails that day. This conclusion was supported in the interview when he explains that he saw that Evan received 15 emails and decided that the new points should all go at 15, though when asked he was unable to support his decision to divide by three. The determining factor was then seen to be his written justification for his population graph, which was the final task of that question. He, like many students, attempted to draw a dot plot for the population graph, only he chose to do so by growing the sample that he created in the first task (which had the three additional points at 12-15 that distorted the original shape of the graph). The graph contained no new points on the horizontal axis, and appeared at first glance to be a vertical growth strategy; however, his AIIR justification for why he chose that shape was “I doubled everything.” Because of his understanding that doubling could be an effective strategy his response was coded as a U₂. This assessed level did not match the interview level. The U₂ coding suggests that Devin’s informal inferential reasoning on questions where he chooses between models is at a Naïve level.

**Devin CTS3.** Devin used a single strategy in both the interview and on the AIIR. Devin based his conclusion on which graphs had the highest frequency at the mode, and chose that as signifying that particular pair as exhibiting the most
evidence. Since he was using one reasoning structure, this placed him at U₁ for both the assessed and interview SOLO level for this task type.

**Devin CBM3.** During the interview Devin explains that since the six lines did not have identical results, some lines had a higher probability than others. This reasoning was mirrored in his AIIR, and was coded as U₁ in both cases. Devin also used undirected-arithmetic in both cases when attempting to create a population graph that modeled his decision, which is a pre-structural code. The result was that Devin is considered to work at a U₁ level for the CBM question type based on both the interview and the AIIR.

**Erica EPG4.** Erica’s case was selected for this section to provide an example of a student with cycle-3 reasoning. When given the task of growing a sample from 10 data points to 23, Erica not only grows the sample proportionally vertically, but includes an additional point at a value that was not previously in the data set; a strategy that she also used on the AIIR (see *Figure 6*). This strategy was coded ‘shape-preserving-growth-vertical-horizontal,’ a U₃ code. U₃ indicates that Erica’s informal inferential reasoning on this question is appropriate though it is Unistructural as she uses only a single strategy.

Erica was limited to Unistructural because when she attempted to connect the context to the graphs her responses were incongruous, and she mentioned while drawing the graph in the interview that she thought that maybe if things were working perfectly the time could get down to 2-3 minutes (the stack of three dots on her interview response) which was unsupported in the data. Then she added an
additional dot at 0-1 even though a perfect time would have been 2-3 minutes according to her previous statement. This shows that her use of context was still somewhat naïve, and that it was not coordinated with her ability to manipulate the graph, resulting in a U₃ code for Erica on the EPG task type.

Figure 6. Erica's AIIR (left) and interview (right) growing a sample task responses.

**Erica CTS4.** In this problem Erica began by saying that since the data weren't the same for both graphs, there had to be a difference in the underlying populations, exhibiting no allowance for variability, a pre-IIR cycle code. Her reasoning was somewhat inconsistent because she said that since the samples were different the populations were different, but when asked if samples of 400 were taken from each population she said they would probably be the same.

Her responses on the AIIR were also self-contradictory. In the AIIR she provided the following written explanation when asked to justify her response that the populations were the same:

If you look at the striped theres (sic) dot from 3.0 to 11.0 but if you look at the spotted the dots go from 4.0-10.0 so you know that there is a difference.
Also Lafayette is wrong from what his graph shows there are 20 dots for striped and 20 dots for spotted so in other words there (sic) equal.

So her reasoning appears to be that the populations are different because of a minor difference in range, but she also seems to contradict herself where she uses the fact that the number of data points are the same in each graph as evidence that the populations are the same. In this case it is particularly interesting because she seems to be using the two strategies, range and cardinality, in a coordinated way, regardless of the fact that they give her conflicting results. Despite the contradictory results, she is seen as using multiple Pre-IIR reasoning structures in a coordinated way, suggesting relational reasoning thus resulting in an R₁ coding.

**Erica CBM4.** Erica exhibited the same strategy in both the interview and the assessment for the CBM problem type: personal-context. In the problem she is told that farmers are raising purple pigs, and that the more purple a pig is (quantified by its ‘purple score’) the more it’s worth. She is presented with a population graph for the past year’s pigs and information that a sample from this year’s pigs had a certain purple score. When asked if this year’s pigs are going to be more purple than last year’s she says that they would be because as purple pigs get older, they’re more likely to have darker purple babies. This was seen as a pre-structural level response because she expanded on the context in order to be able to answer the question solely based on the context. These three responses from Erica illustrate that a student’s IIR can exhibit multiple levels and different cycles, depending upon the type of problem posed.
**Frank EPG4.** Frank’s case was selected for presentation here because of the opportunity to show an example of relational thinking (in the CTS task type). For the EPG task type, Frank’s response in the interview showed him still able to double the existing data (as he did in the AIIR), but when asked to grow a sample of 10 by an additional 7, he made some unusual choices. His results (shown in Figure 7) show that while he did increase the number of data points at the mode by a greater amount than most other points, he also tripled the number of data points at the set’s maximum value. The mean and median of his additional points were also significantly higher than the mean and median of the existing data points. When asked to justify his response, he responded “I think as they did further research that it would get better and better and less time that it would take as they work out kinks…” Even though his new points indicated an increase in typical transportation time rather than a decrease, the reasoning behind altering the central tendency of the subsequent week’s measurements was grounded in an alteration he made to the context of the problem (specifically because research is ongoing and improvements are constantly being made). This is a pre-structural code.

In the interview, Frank was presented with a problem where doubling was not a possible strategy, and in this case Frank exhibited pre-structural reasoning when confronted with that version of the task. This confirms that his only cycle-2 strategy is that of doubling. In light of his ability to use a doubling strategy when growing a sample, he was seen as having a U₂ ability level for the EPG question type.
Frank used a single strategy in both the interview and the assessment when comparing two samples. Frank picked a cut-off score for ‘low values,’ and for ‘high values,’ and counted how many points occurred below his low value cut-off and above his high value cut-off. In the AIIR, for example, he selected the number six as his cut-off score for low values, and counted how many threes, fours, and fives each graph has. He concluded that the graph that has fewer scores below his cut-off score of six must have higher scores in general, which he generalized to the population. Since the cutoff values were not appropriate to the data his strategy was seen as a cycle-1, Pre-IIR strategy. Since his strategy coordinated two cycle-1 strategies (inappropriate low cut value and inappropriate high cut value) his reasoning was assigned a code of R₁ on the CTS task type.

Frank CBM4. Frank’s response on the CBM question (the same purple pig question as Erica) during both the AIIR and the interview exhibited a lack of allowance for variability because when he saw that a sample of 10 pigs resulted in a 76 he assumed that the typical score for the entire population would be exactly 76. When asked to elaborate on his explanation in the interview his chain of reasoning
was essentially that because the new typical score was 76 rather than 72, the farmers must have changed their method; since the method change had a positive impact on the sample, it means the pigs this year will be better than last year, and that next year’s will even be better than this year’s. In both the assessment and the interview the cycle was determined by Frank’s lack of allowance for variability, with his use of a personal-context strategy becoming apparent in the interview. In both cases his reasoning was seen as overuse-of-context and personal-context; since overuse-of-context is a cycle-1 code, Frank’s SOLO level for the CBM problem type was U₁.

**Coding discrepancies.** The degree to which the students’ interview SOLO levels and assessed SOLO levels match provides a measure of co-measurability for the AIIR. What needed further examination were those items where there was not a match. This occurs in Alexis’ and Betsy’s CTS task responses. In these two instances, the discrepancy was determined by the student interview explanations and resulted from the students’ misunderstanding of the question during the written assessment. Both students changed their responses in their interviews.

A second type of discrepancy was the difference between Unistructural and Multistructural reasoning within a cycle. This occurs at D/EPG. The discrepancy in Devin’s work on the EPG task occurred when Devin was able to use doubling and, when presented with the alternative task where the number of new cases and number of existing cases were relatively prime, included the new cases in a way that was vertically proportional. In the initial co-measurability check versions of the
AIIR, students were only presented with either a problem where doubling was possible or one where doubling was not possible, but never both. This resulted in the possibility of students exhibiting the strategy shape-preserving-growth-vertical-only or doubling strategies but not both, which in turn led to a deficiency in the test for identifying Multistructural students (particularly for cycle 2).

**Modifications**

After considering the pilot and initial co-measurability check results, it was decided that a single short assessment to accurately identify students’ IIR across all three question types was untenable. Instead the decision was made to focus on creating a short assessment that could accurately identify students’ IIR in a single question type.

It was posited that rather than a single question for any given question type, students may need a break and transition to a new context in order to facilitate them exhibiting Multistructural (M) capability (when it existed). To this end, the assessments administered for final co-measurability check consisted of only a single question type. The assessments each contained three questions of a single problem type. Items in a single question type could then be analyzed individually, as a set of two items (referred to as pair-wise analysis), and as a set of three items (referred to as three-item analysis). The pair-wise and three-item analyses were performed in order to identify the number of questions the AIIR should have of each question type in order to be co-measurable. This allowed for several different results to come from the analyses. As a reminder, the question types were Choosing Between
Models (CBM), Comparing Two Samples (CTS), and Estimating Population Graphs (EPG). Four questions of each question type underwent an initial co-measurability check (one in the pilot and three in the dissertation study), and the questions will be referred to by their type, number, and version; for instance, the third question in the final co-measurability check item bank from the EPG question type will be called EPG3F. The versions of the assessment that are to be administered are shown in Table 12.

<table>
<thead>
<tr>
<th>Version Number</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>EPG 1F, 2F, 3F</td>
</tr>
<tr>
<td>2</td>
<td>EPG 1F, 2F, 4F</td>
</tr>
<tr>
<td>3</td>
<td>CTS 1F, 2F, 3F</td>
</tr>
<tr>
<td>4</td>
<td>CTS 1F, 2F, 4F</td>
</tr>
<tr>
<td>5</td>
<td>CTS 2F, 3F, 4F</td>
</tr>
<tr>
<td>6</td>
<td>CBM 1F, 2F, 3F</td>
</tr>
<tr>
<td>7</td>
<td>CBM 2F, 3F, 4F</td>
</tr>
<tr>
<td>8</td>
<td>CBM 1F, 3F, 4F</td>
</tr>
</tbody>
</table>

This study was constricted in its administration because the state standardized testing that had been originally scheduled for a later date in the year was rescheduled in such a way as to interfere with data collection. The revised schedule left the cooperating teachers pressed for time, so the assessments were held to three questions rather than four or six. This administration is more consistent with the idea of short focused assessments, allowed for focusing the interviews on a single question type to more thoroughly explore student reasoning, and fit the needs of the participating students and teachers given the constrained
schedule. Eight versions were made in order to maximize the number of pair-wise analyses while still allowing each version to be administered twice (four pairs of students in each class were followed closely, resulting in 16 students for whom sufficient data were collected to perform the final co-measurability check).

From Table 12 it can be seen that the only question combination missing is EPG3 with EPG4. This pair was omitted because both of these questions allow for the doubling strategy. As was mentioned at the end of the initial co-measurability coding section (page 68), it seemed most useful to combine a doubling question with a question where growing the sample has to be done in some way other than doubling, so analyzing the two items that allowed for doubling as a pair was deemed least important and was therefore chosen as the pair to be omitted.

**Final Co-Measurability Check**

The final co-measurability check used additional data sources that were unavailable in the initial co-measurability check in order to more accurately identify students' observed SOLO level. In addition to interview transcripts and artifacts, the final co-measurability check data sources also included classroom video, LiveScribe (2012) data to capture classroom conversations, and unit artifacts in the form of LiveScribe (2012) files for written work and TinkerPlots (Konold & Miller, 2005) files for computer work.

Coding of the final co-measurability check versions of the AIIRe began with the target groups of students. Each three-page assessment was scanned as a PDF then converted to three .png files, one file for each question, so that they could be
imported to HyperRESEARCH (v. 3.5.4). Once imported, they were coded for student reasoning structures using the SOLO descriptors. Codes for students were analyzed to produce a SOLO code for each of the three possible pairs of items, and a SOLO code that took all of the items into account.

Each interview was then transcribed, checked against the transcript for accuracy, then imported to HyperRESEARCH (v. 3.5.4). Pictures were taken of the student-created posters and screenshots were taken of their digitized interview work from the LiveScribe (2012) files and imported into HyperRESEARCH (v. 3.5.4).

Coding began once these data sources were imported and only for those data from students whose question types were completed on the AIIR. For instance, Echo took Version 1 of the AIIR (see Table 12) which contained only EPG tasks. When coding Echo’s poster, only reasoning structures related to the EPG capability of the student were coded. Interviews were coded first, then the posters, then the interview artifacts. When coding the interview artifacts, the LiveScribe (2012) audio file was used to provide context for the student drawings, and to be able to observe the manner in which students constructed the graph (observing things such as the order in which points were added to a graph or seeing at what point in a process students got ‘stuck’). Next, the classroom interactions were coded using primarily the LiveScribe (2012) audio files (due to better audio clarity than the video files), using the video files to provide context. Finally, the students’ classroom artifacts were coded including student scratch work, assignments, and any other writing captured by the LiveScribe Sky pens (Livescribe, 2012).
Once all data were coded, the codes were analyzed to produce the observed and assessed SOLO levels. Observed SOLO levels were based off of all of the information gathered about the student except for their written work on the AIIR administrations. Four assessed SOLO levels were identified for each student: one for each possible pair-wise analysis and one for the three-item analysis (see page 32).

The following sections present the results of the analysis for one student from each task type. These students were chosen in order to present a wide range of student thinking, the opportunity to present student work, and to provide examples of common student reasoning structures. The analyses of the thirteen students not presented here can be found in Appendix D. Both here and in the appendices students are organized by the task type on which their version of the AIIR focused. The analysis for each student is split into 1) a section presenting the analysis for the student’s observed SOLO level, and 2) a section presenting the analysis for his or her assessed SOLO levels. Each analysis is further separated into the type of AIIR test the student took: again, tests were designed to focus on a single facet of IIR, that of Estimating a Population Graph (EPG), Choosing Between Models (CBM) or Comparing Two Samples (CTS). Throughout, the term ‘variability’ will continue to refer to sampling variability. Students’ understanding of variability will be classified here as “not present,” “limited,” or “appropriate.” Limited variability in the context of this study denotes student conception of variability that is not developed enough to draw appropriate informal statistical inferences. Appropriate
variability is not intended to imply that students have complete understandings of
variability but rather that their informal understanding of variability is sufficient to
allow them to make appropriate informal inferences.

For each student, the section containing the observed SOLO level occurs first.
Evidence and analysis is presented, and each subsection closes with the resulting
SOLO code for the student; details on structure codes and how they informed the
SOLO codes can be found in the
SOLO Taxonomy Adaptation section (see page 45).

The presentation of a student's observed SOLO level section is followed by a
section on his or hers assessed SOLO level. Each assessed SOLO level section begins
by presenting evidence regarding the structure codes identified for that student, and
the items from which those structure codes came. Recall that each student was
administered three items of a single question type, and that there are three pair-
wise analyses and one three-item analysis. As such, each student is assigned SOLO
levels four times; the final part of each section describes the SOLO levels attributed
to the student by each of these four analyses. Each question type has a summarizing
table that displays the results of the four analyses for each student (for the EPG
question type see Table 13 on page 83, for CTS see Table 14 on page 84, and for CBM
see Table 15 on page 90).

**Estimating a Population Graph (EPG).**

*Reupert observed.* Reupert was paired with Echo for instruction and the
project, and as a group, they had some difficulty with variability. In the interview,
Reupert was shown his sketch of the population graph from EPG4. In EPG4 the students are told that Priya lives on the moon, and her commute time to school is somewhat unpredictable. The first part of the task presents a graph that Priya made of commute times after ten trips, and the second task requires the student to estimate the shape of the graph after a year. Reupert drew the graph shown in **Figure 8.**

This graph was unique in what had been observed amongst students in the study because the shape of the population graph is identical to the shape of the sample graph. This is generally what occurs with students who have a cycle-2 (Naïve structure) strategy for sample growth, except that Reupert also linearly increased values between the minimum and maximum which were previously zero. When he was asked if the decision to include those values was intentional he said it was not, but in a follow-up question asking whether all of the data would occur at values where data already existed he said “probably, maybe one or two will be in other categories, but based off this, that is what I inferred.” His graph when taken on its own suggests the linear-growth-vertical-only code, Pre-IIR cycle-1 reasoning. His explanation, however, suggests that he is capable of some cycle-2 reasoning. Since his cycle-2 reasoning is not reflected in his solution to the question (the additional values were said to be a mistake), his limited allowance for variability is useful for identifying which cycle he is reasoning in, but isn’t used as a strategy for finding a solution (and therefore not a code-able strategy).
Reupert’s interview did show he was able to use the doubling strategy for growing a sample. When doubling was not possible he would work left to right increasing each column of data by one point, which he would repeat until he ran out of data points, increasing each stack by three points. This is a cycle-1 strategy called linear-growth-vertical-only, and shows that while he does have one strategy in the cycle-2 (doubling) and some allowance for variability, when his single cycle-2 strategy doesn’t work he reverts to Pre-IIR (cycle-1) strategies. Since the coding strategy always takes the highest level coded on student reasoning, Reupert’s ability level in EPG tasks based upon the observed data collected was classified as U₂.

**Reupert assessed.** Reupert’s version of the AIIR included items EPG1, EPG2, and EPG4. Reupert’s primary strategy was linear-growth-vertical-only, and it appeared on a part of every single problem, either in growing the sample or sketching the population graph. In this strategy, a student attempts to grow each value by the same number of data points. In watching the LiveScribe (2012) files of students who used only the linear-growth-vertical-only strategy, it was seen that they typically begin at the lowest value and added one point at each value working left to right. When they’ve added one point to each value, they begin again at the
lowest value and repeat this pattern of adding points until they run out of points to add. If they run out of points before they finish their last cycle of adding points, students simply leave off where they run out of points, resulting in lower values having their frequency increased by one more than higher values. Reupert didn’t follow this pattern. In both cases where he grew a sample using linear vertical growth he omits the point at the mode, which seems to be more consistent with another cycle-1 strategy of leveling-off-data. However, there was insufficient evidence in his written explanation to confidently code him as coordinating these two cycle-1 strategies.

Reupert was able to use a doubling strategy on EPG4, so two of the three pair-wise analyses, along with the 3-item analysis, resulted in SOLO codes of U₂. The remaining pair-wise analysis can confidently be stated as at least U₁ because of his use of a linear growth strategy, and because he appears to be coordinating the leveling and linear growth strategies, likely R₁.

**Reupert’s results compared.** Reupert was identified as being U₂ in the observed portion. Of the pair-wise and three-item analyses, only a single pair-wise analysis contradicts the observed SOLO level: the one in which he is identified as reasoning in cycle-1. This supports a larger trend that will be seen in subsequent sections and in the task type analysis at the end of this chapter; single item assessments would not have co-measurably measured students’ IIR.
Choosing Between Models (CBM).

Scoot observed. Scoot’s case is presented here due to her having one of the greatest discrepancies between observed and assessed SOLO levels. Scoot was paired with Larry for the instruction and project part of the research. Scoot and Larry produced very little usable data during instruction; they would also answer questions quickly and with little thought, and were often off topic. They also accidentally turned the recorder on the pen off early in day 3, and placed it out of range on day four, leading to limited available data of conversations that they had. They did not complete their poster, however, their available data do prove illustrative.

Scoot’s strategy in the interview was primarily undirected-arithmetic. For example, in the amusement park problem she is presented with the number of times a person wins in each of six lines after 15 people have gone through each line. He is later asked to create a table for what the data could look like after 50 people had gone through the lines. To answer this question he says he took the number 50 and divided it by the line number (lines are labeled one to six). During an interview question she was asked what the graph of a bank of snow machines would look like if all the machines were running properly. The following exchange occurred:

I:...you say ‘might all be even,’ is that what you would expect, that if they were all operating correctly it’d just be an even line?

S: Yeah

I: So when you say ‘might be’ do you mean there are other possibilities?
S: <nods>

I: Ok, so what else could it...

S: Like, if they were all the same brand, and they were like, one brand was more productive than others?

I: Ok, well let’s say that they’re all the same brand and the same model, so they’re all expected to produce the exact same amount of snow on a typical night... and that they’re all working properly, what would that graph look like?

S: Um, even?

I: Just completely straight even?

S: Or maybe a little higher than the others because they're expected to be but some might produce more than others.

Here Scoot showed two strategies. First she suggested that there may be some variability in the data due to a mix of different types of machines; this is an example of personal-context. Later she restates that they might not be completely even, but instead acknowledges that there may be some variability. This is the first instance of Scoot making some allowance for variability, and the allowance for variability had an influence on the answer to the question. As Scoot shows some allowance for variability, she is seen as U2 for the CBM question type.

**Scoot assessed.** Scoot’s AIIR included items CBM2, CBM3, and CBM4. Scoot exhibited difficulty reading the graphs; in the purple pig context where population scores for a single year are represented (CBM4), she noted that “the purple pig’s
(sic) purple rate tends to gradually increase.” Because her response was based on a graph misinterpretation and therefore irrelevant to the question, this item was recorded as having a pre-structural response.

Scoot also exhibited no understanding of variability. In CBM3 she said the amusement park lines weren’t fair “because some rates are higher than others.” When asked in CBM2 to describe what a graph would look like when all machines are operating correctly she said they “might be all even,” which seems to include some probabilistic language in that she expresses uncertainty, but because it did not impact her answers to the tasks it was not enough to place her into the Naïve cycle. Because of the lack of allowance for variability in her strategies, all pair-wise and three-item analyses identified her SOLO level as U₁ for the CBM problem type.

**Scoot’s results compared.** Scoot’s case shows an indication of being U₁ from the AIIR and U₂ from the observations. This case was particularly important because a seemingly sparse data set from the observed portion produced a higher SOLO level than the AIIR. An explanation for this difference is that she did in fact develop new understandings through the assessment and interview. This is supported in the interview transcript beginning on page 77, where she began by implying that only different brands could have different outputs, then in the end allowed for the possibility that two machines of the same brand could have some small differences in their output.
Comparing Two Samples (CTS).

**Johny observed.** Johny was paired with Haley during the interview and project portions of the research. The group work showed a limited understanding of variability and an over-reliance on context. However, Johny’s interview showed a much greater variety and higher level of reasoning capability.

Johny used a variety of strategies to compare two samples. One strategy was that he would find some score that he saw as central, and count how many points were above or below that score, then use that to draw conclusions; this is similar to Frank’s strategy in the initial co-measurability check, only with a larger range (recall, this is a cycle-1 strategy). He used a similar strategy later in the interview, but instead of a simple counting strategy he began talking more about the general shape of the graph as determined by the frequencies at different points; he coordinated this comparison based on shape with a focus on the location of the modes. This coordination of two Naïve cycle strategies (Naïve cycle because they lacked consideration of context) resulted in Johny’s reasoning ability being classified as R₂, relational, for the CTS task type.

**Johny assessed.** Johny’s AIIR test includes items CTS2, CTS3, and CTS4. In CTS4 he was asked to explain what the graphs would look like if two treatments had the same affect, and the graph he produced, shown in Figure 9, was unlike any other in either co-measurability check or the AIIRs reviewed for the purposes of refining the SOLO adaptation. It can be inferred that R is for reading and V is for video games. From the context of the problem it seems likely that the S is for same, and
that perhaps in this graph he is showing that all of the data favored the same impact and none favored reading or the video game.

![Graph Image]

*Figure 9. Johny’s CTS4 response*

In CTS4, Johny was asked to determine whether there is a difference in two populations based on two samples. He said there was a difference because “it has less on the low and more on the high.” This use of two cycle-2 strategies where the number of high values are compared and the cycle-2 strategy where the number of low values are compared, in a coordinated way in a single explanation resulted in a SOLO code of R2 for Johny in every analysis except for the pair-wise analysis CTS2 and CTS3.

**Problem-Type Analysis**

**Estimating a Population Graph (EPG).** The administration and sampling method allowed for 12 analyses of 2-item sets and two analyses of 3-item sets. The 2-item sets had a 66% rate of co-measurability in identifying the SOLO level of a student, whereas the 3-item sets were successful 100% of the time.

Two versions of the EPG assessment were given; one containing items 1, 2, and 3, and one version containing items 1, 2, and 4. The item combinations and their results are shown in Table 13. During the analysis it was noted that EPG 1
consistently resulted in very minimal amounts of usable data, when it resulted in no usable data at all. In the table, we can see that three of the four pair-wise combinations in which errors occurred took place in a combination that included item EPG1. Because of this, EPG1 was omitted from the final item bank.

The remaining inaccurate combination is 2,3 from the student Alice (pseudonym). This result came from improperly following the directions. EPG2 and EPG3 are designed as a pair where students grow a sample from 15 to 28 and from 15 to 30, respectively. The design intent was that students in the Naïve cycle would be able to demonstrate a doubling strategy in EPG3, but EPG2 would force them to either adopt a second cycle-2 strategy (such as proportional vertical growth) or a cycle-1 strategy (such as linear growth) in order to differentiate between $U_2$ and $M_2$ students. Alice completed EPG3 by doubling, but completed EPG2 by also doubling, by adding 15 additional data points rather than the 13 requested in the problem. However, she did demonstrate an ability to use the shape-preserving-growth-vertical-only strategy on EPG1 (where the sample is grown from 15 to 26), and demonstrated in interviews and classroom work that she was unable to coordinate the two approaches or proceed to cycle-3 reasoning. Because of this, her EPG 2,3 result was inaccurate whereas her EPG 1,2 and 1,3 were accurate. This was decided to be more a result of the student misreading the question or not following directions than a fault in the item, so item three was retained in the item bank.
Table 13
**EPG final co-measurability check results by item combinations**

<table>
<thead>
<tr>
<th>Item numbers</th>
<th>N total</th>
<th>N accurate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1,2</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1,3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1,4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2,3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>2,4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1,2,3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>1,2,4</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

**Comparing Two Samples (CTS).** The administration and sampling method allowed for 18 analyses of 2-item sets and six analyses of 3-item sets. The success rates are a little more complicated to talk about here. Overall, five of the 18 2-item set analyses resulted in Assessed IIR levels lower than Observed levels; this returns a success rate of 72%. For the 3-item set analyses one of the six resulted in Assessed IIR levels lower than Observed, resulting in a 83% success rate. The breakdown is provided in **Table 14**, where shaded cells indicate a discrepancy where the observed SOLO level exceeds the assessed SOLO level.

The results in Table 14 for the pair-wise analyses are not as accurate as hoped. It is helpful, then, to take a second look at the data organized by student name rather than analysis set. Here we can see that over half of the inaccurate measurements came from one student: Sofia.
Table 14

<table>
<thead>
<tr>
<th>Name</th>
<th>1,2</th>
<th>1,3</th>
<th>1,4</th>
<th>2,3</th>
<th>2,4</th>
<th>3,4</th>
<th>1,2,3</th>
<th>1,2,4</th>
<th>1,3,4</th>
<th>2,3,4</th>
<th>Observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avalana</td>
<td>U₁</td>
<td>M₁</td>
<td>M₁</td>
<td>M₁</td>
<td>M₁</td>
<td>M₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td></td>
</tr>
<tr>
<td>D’Jackson</td>
<td>U₁</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td></td>
</tr>
<tr>
<td>Dynklfrin</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₂</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td></td>
</tr>
<tr>
<td>Johny</td>
<td>U₁</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td>R₂</td>
<td></td>
</tr>
<tr>
<td>Phylicia</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td></td>
</tr>
<tr>
<td>Sofia</td>
<td>U₁</td>
<td>U₁</td>
<td>M₁</td>
<td>M₁</td>
<td>M₁</td>
<td>M₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td>U₁</td>
<td></td>
</tr>
</tbody>
</table>

Sofia’s case is somewhat straightforward. Sofia is what Biggs and Collis (1991) would call transitioning, which means they appear to be in the process of moving from one cycle to the next. Sofia’s consistently exhibits a Naïve level of understanding about variability throughout the AIIR. When asked in CTS2 to describe what graphs of two samples would look like if their populations scored the same on a test she said “the graphs would be the same,” and when asked to select which pair of graphs showed the least evidence of a difference in population parameters, she justified her response by saying that the pair of graphs selected were “...the closest to the same...” The classification of her IIR level as Naïve was further supported by choosing school A as having the most evidence from the set of graphs from Figure 10. Her response from CTS4 further reinforced the classification since she decided that the populations underlying the graphs were different and that she was “very certain?” of her response.
All of Sofia’s AIIR responses, classroom artifacts, and classroom observation data support the notion that she was definitely in the Pre-IIR cycle; she consistently exhibited no allowance for variability and her interpretation of ‘most evidence’ was based solely on the frequency of data comprising the mode. The single contrary piece of evidence came from the post-interview that occurred two weeks after the unit and assessment were finished. In the interview she was asked to draw a pair of graphs that represented two samples of size ten from a single population (the Spotted Apox, in this case). As predicted, she drew two identical graphs. However, when asked to talk about the graphs the following exchange occurred:

I: Okay, so down here in this second set we are going to pretend Lafayette gets 20 more spotted, right, and he just looks at spotted, and he has 10 of them do the test. And he graphs the first 10 here, and he has the next 10 do the test and graphs them here. So they’re both spotted so they have the same intelligence, right? So could you make those two graphs for me? <Student draws> Okay, can you tell me why you chose these points?

S: Well, I kind of tried to make it the same as the spotted ones from both of
them, but I guess I did it a little bit different but I kept these two graphs the same.

I: Okay, so you kept the spotted one graph and the spotted two graph exactly the same; do you think that's important?

S: No, they are like the same thing, so if they weren’t the same they should at least be pretty close because they’re the same animal I guess.

I: Okay, and how close is pretty close?

S: Like, not have too much off, but I wouldn’t think they’d have any below four because of like how they did on the first test, and I think they’d have more within the eight or seven to 10 range and not have that many below it.

Here she appears to have made some progress in having some small allowance for variability between samples from similar (in this case identical) populations. This is classified by the code limited-variability, and is a Naïve cycle code.

In this case the evidence suggests that her reasoning has progressed since taking the assessment. In the pilot and initial co-measurability checks when there was a discrepancy it was most often that a new question created an opportunity for the student to exhibit some reasoning structure that they hadn’t had an opportunity to present before, and the new reasoning structure would fall into a higher SOLO cycle. In this case Sofia provided multiple pieces of evidence about her use and understanding of variability, and that evidence showed progress in her thinking between the assessment and the interview.
Another inconsistency came from Johny. In Johny’s case, he reflected on the AIIR during the interview and chooses to change his answer on item CTS2. His original answer was the distractor in which there was no change to any measures of center, but the distribution was changed so that the measures of variability would increase (see Figure 11). This choice has been coded one-irrelevant-aspect, and is seen as a reasoning structure in the Naïve IIR cycle. His second choice was also coded one-irrelevant-aspect because the only change involved the outlier. Also, his third choice was the intended first choice because the different mode and median were recognized informally by their general shape within the data sets.

Johny decided to change his answer to A:2, B:1, C:3. His justification showed that he identified the change in measures of center (informally) in school B, and considered the relative importance and magnitude of the changes in distribution and the importance of single outliers and coordinated his understandings of the two to order them the way he did. This was coded as ‘Coordinating two relevant aspects,’ (since once measure of center has been identified as determining the correct answer, the two aspects that were irrelevant to that comparison become the only relevant aspects for differentiating between the remaining options), but Johny did not express uncertainty or relate the graphical data to the context, limiting his response to the Naïve cycle.

Johny’s new thinking about the problem came with minimal prompting from the interviewer, and his decision to revise his response was spontaneous and not connected to any line of questioning posed by the interviewer. Because of this it is
not possible at this point to identify a problematic aspect of the assessment that led to the disparity. Potential causes are that the student didn’t take enough time answering the questions during the AIIR administration, that he overlooked some aspect of the question that he only noticed on second viewing, or that he furthered his understanding about IIR between the time of the assessment administration and the time of the interview.

Figure 11. Johny’s CTS2 response

The remaining inconsistency comes from D’Jackson’s assessment. This is attributed primarily to the difficulties with the CTS1 item not being able to generate enough data. This is a case where the AIIR item was problematic, and was removed from the final item bank.
Ultimately there is enough evidence through the analysis of classroom data, artifacts, and student presentations to support the assertion that Sofia's observed IIR level was the same as her assessed IIR level at the end of the statistics unit. In Johny's case it appears that the difficulty is more a result of the format/implementation of the assessment than with the item from the assessment. Other inconsistencies are attributed to the lack of information gathered from CTS1, which will be omitted from the final item bank.

**Choosing Between Models (CBM).** The administration and sampling method allowed for 18 analyses of 2-item sets and six analyses of 3-item sets. The comparisons are shown in Table 15, where cells where assessed SOLO levels that are exceeded by the observed SOLO level are shaded; the Xs are to denote cases where the items were administered but did not collect sufficient data to produce a SOLO code. From the table we can see that every comparison from the AIIR met or exceeded the Observed IIR level for the student except for one, Scoot. This shows that the CBM question set met or exceeded the accuracy of the observation data 75% of the time for the pair-wise analysis and 83% of the time for the three-way analysis.
Table 15

*CBM item combination results organized by student pseudonym*

<table>
<thead>
<tr>
<th>Choosing Between Models</th>
<th>Haley</th>
<th>Hannibal</th>
<th>Larry</th>
<th>MJ</th>
<th>Scoot</th>
<th>Steve</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As is shown in Table 15, the primary source for inconsistencies was from the case of Scoot. Scoot’s situation mirrors fairly closely what happened with Sofia in the CTS question type. Every piece of evidence prior to the interview, along with the bulk of the evidence in the interview, indicate that she has only cycle-1 strategies except for one tentative answer at the end of a line of questioning in the interview. In this way, her cycle-2 reasoning can be interpreted as due to progress in the student’s thinking between the assessment and the interview (slightly longer than a week), rather than inaccuracy of the instrument.

This leaves only a single inconsistency shown in the case of Larry. The difficulty with Larry’s AIIR was his extreme brevity; in question one the only answer provided on the entire question was “it is good,” which was not useful in identifying student reasoning. The lack of evidence of student reasoning from Larry in CBM1 and CBM2 is represented as an X in Table 15, indicating that no SOLO code was assigned for that pair-wise analysis. For further details about Larry, see Appendix D.
Summary

The main focus of this chapter has been on describing the co-measurability checks and their consequences. Findings of the first co-measurability check for new items developed for the AIIR show that the modifications based on the pilot were effective in eliciting student reasoning structures, but that the format of the assessment (a single assessment of all three task types) was not appropriate. Some of the most interesting student cases were presented to highlight key student thinking, to provide examples of student work, and to set up the discussions of modifications and limitations. Several items were modified to allow for a lower reading level, to present more variety in task types, to clarify graph presentation, and to accommodate the change in format. For example, in the initial validation interviews several students noted that the gradient shading on the items was distracting; as a result, graphs were changed to be solid monochrome.

A final co-measurability check was completed by looking at items analyzed as pairs and items analyzed as sets of three. The items analyzed as pairs matched or exceeded the observed SOLO level of the students 70.8% of the time, and the items analyzed as sets of three matched or exceeded the observed SOLO level of the students 87.5% of the time. Two students were identified as having advanced their IIR ability prior to or during the interview, and 42.9% of the errors from the two-item analysis, and 100% of the errors from the three-item analysis came from these two students. This suggests that the AIIR can co-measurably elicit students'
reasoning structures about IIR with the same or at a better rate than the classroom observations and student interviews.
Chapter 6: Conclusions

The research focus for this study centered upon the development of an assessment tool for characterizing informal inferential reasoning (the AIIR) and upon the modification of a SOLO taxonomy for describing how students' informal inferential reasoning progresses developmentally. Previous chapters presented the motivation for this work, the background literature, the methods for addressing the research questions, and the analysis of the data. This final chapter presents the results, limitations, and implications of this research study. The chapter is divided into five sections that address results from the two research questions on the development of the SOLO taxonomy and the AIIR, discuss the limitations of the design of the study, provide implications and open questions from the research findings, and present a final discussion.

Question 1: SOLO Taxonomy Results

The overarching goal of this research was to produce a way of eliciting and describing students' IIR abilities. The second part of that goal, describing students' IIR abilities, required the creation or adaptation of some analytical framework that could be applied to student data in order to categorize each student. Research in statistics has supported the idea that statistical learning is hierarchical (Watson & Callingham, 2003) and, being a cognitive hierarchical model, the SOLO taxonomy made sense as a starting point. The choice of the SOLO taxonomy was further supported by a recommendation by Reading that IIR could be looked at using a two-cycle SOLO taxonomy adaptation (Zieffler et al., 2008). This led to establishing the
first research question posed in this study: “what developmental descriptions of students’ IIR abilities are suggested by a SOLO taxonomy.”

This research began by piloting a small number of items combined with interviews with a small group of 6th-grade students. The initial SOLO taxonomy adaptation mirrored Reading’s (Zieffler et al., 2008) recommendations, with the first cycle being Naïve reasoning and the second cycle being Appropriate reasoning. This quickly proved problematic, as it resulted in nearly all students being classified as pre-structural. Since one of the needs of the educational community addressed in this study was developing an ability to look at the longitudinal development of students’ IIR, having all 6th-grade students (and by extension, likely all younger students) in a single classification ran counter to the research intention. However, data showed that the pre-structural reasoning was quite varied, and also showed a potential progression towards Naïve reasoning. Through this analysis of the pre-structural responses, a new cycle of reasoning was identified that would become the new first cycle, Pre-IIR reasoning. This new SOLO taxonomy adaptation to the pilot data was then reapplied and found to be sufficient as a starting point for the current research study.

During this research study, the fundamentals of the SOLO adaptation produced in the pilot study proved to be sound; the cycles differed from each other significantly enough to prove useful, the hierarchy of thought suggested by the adaptation was consistent with what was observed, and the level descriptors provided a useful starting point despite needing some further refinement. Details
about the adaptation can be found in Chapter 4, but the bulk of the modifications focused around more rigorously defining the cycles and the levels within those cycles by reviewing student work in order to identify reasoning structures for which the initial adaptation did not provide a clear way of categorizing. The larger sample size in the current study contributed to a better understanding of what student reasoning structures were prevalent and this gave rise to a more efficient way of describing the nature of the reasoning structures at each level. The resulting level descriptors are shown in Table 6 in Chapter 5.

Prominent student reasoning structures were presented throughout the AIIR co-measurability check sections and in Appendices III and IV. The most common ones for the task types are included in Table 7 through Table 9 on pages 50-52.

**Question 2: AIIR Results**

The first part of the over-arching goal required the development of a way to elicit student thinking about IIR in a way that would be convenient to researchers and minimize the intrusion on students and classroom teachers. Based on the literature review, Zieffler et al. (2008) provided guidelines to what such a tool might look like. Those guidelines served as the basis for constructing the Assessment of Informal Inferential Reasoning (AIIR) used in the pilot study. This prompted the second research question in this study: “how effective are the items in the Assessment of Informal Inferential Reasoning (AIIR) item bank in eliciting and recording students’ IIR abilities, relative to extensive classroom observations and student interviews?”
Work on this question began with the creation of the original AIIR items. The original AIIR (which can be found in Appendix A) was a three-question assessment. The first context in the original version was that of a student categorizing animals from a science-fiction book based on how many feet they had, and was used to frame the Estimating a Population Graph (EPG) task type. The second context was that of students performing science experiments to figure out how many calories were in a single serving of various types of potato chips, and was used to frame the Comparing Two Samples (CTS) and Choosing Between Models (CBM) task types. While these items seemed to work well with the 6th-grade students they were piloted on, the same concern that lead to the creation of a third SOLO taxonomy cycle cast some doubt over the co-measurability of the instrument: since the 6th graders’ IIR was fairly undeveloped it was impossible to know how useful the items would be with students that have more developed IIR abilities.

With the pilot serving as confirmation that such a tool was possible to develop, this study began by expanding the assessment from a single item for each task type to four items for each task type. This allowed for the exploration of a wider variety of contexts and tasks within each task type. These new items were used to create new versions of the AIIR, where each version had one question from each of the three task types.

After the initial co-measurability check it was observed that using a single question for a question type was unreliable in identifying Multistructural reasoning types. Even questions with multiple sub-questions were not able to reliably identify
Multistructural reasoning, which was potentially attributed to students becoming ‘stuck’ on one mode of reasoning for an entire context. For this reason the idea of creating a single short instrument to measure student reasoning for all three question types for the purpose of developing an AIIR was abandoned in favor of investigating the efficacy of short tests that focused on a single question type.

Analysis for the final co-measurability check identified pair-wise and three-item SOLO levels for students. Only two- and three-item analyses were done in the final co-measurability check. Single-item analyses were not done, though it can be inferred from the two- and three- item analyses details that the rate at which single items would reliably identify students’ SOLO levels would be quite low. In particular there were significant problems with the accuracy and amount of data produced by EPG1, CTS1 and CBM1. These items were originally constructed in the pilot and were designed as a single coherent set of items that elicited reasoning across all three question types; as such, it is not particularly surprising that those items were not as successful as the items that were designed with this different implementation in mind. They were also not modified between the initial co-measurability check and the final co-measurability check, a point at which the researcher had a better understanding of what task contexts were appropriate for this population, and what attributes of the tasks were conducive to eliciting student reasoning. These led to the items being eliminated from the final item bank.

When the previously mentioned items are removed from the item bank, three items for each question type remain. These items have been analyzed in pairs
and as sets of three, with all analyses only including items from a single question type. The pair-wise analysis had a success rate of 66.7% while the three-item analysis had a co-measurability or an agreement rate of 87.5%. In the discussion it was also shown that for two of the students for whom the AIIR was inaccurate, the cause was not an error in the instrument but perhaps due more to their ability level increasing in the time between the assessment and the interview. When these students’ results are omitted, the pair-wise analysis had an agreement rate between the assessed and observed SOLO level of 88.9% of the time and the three-item analysis had an agreement rate of 100%.

One potential reason for the lower co-measurability rate of the pair-wise analyses was that some students had difficulties reading some types of graphs. There are two potential remedies to this shortcoming: the AIIR could begin with a short primer on graph reading, or graph sense could be seen as a limiting factor in determining students’ IIR. The data collected was insufficient to identify the appropriate resolution to this, and it remains an open question at the time of this writing.

**Other Results**

This research also presented other results worth noting but not directly related to the research questions. These results came in the forms of a refinement of the definition of IIR, general findings about the construction of items for assessing IIR, and analytic procedures for coding data.
Refining the definition of IIR. A definition of IIR was presented in Chapter 1 as “the use of informal statistical knowledge to make inferences about unknown populations wherein generalizations beyond data are made, data is used as evidence, and probabilistic language in expressing uncertainty about generalizations is present.” This definition appeared sufficient at the beginning of the research in that it captured the essence of the majority of the published research to that point, which was shown in the literature review to be primarily with K-5 students.

Through the course of this research the difficulty of using this definition to differentiate between informal and formal statistical inference became apparent. This difficulty is related to the ‘data is used as evidence’ component of the definition; in particular, when the data being used as evidence is a result of data analysis on the data, what level of analysis is permitted before the type of reasoning is no longer able to be classified as informal? For example, a student in 10th grade who is only able to speak generally about where the bulk of the data is located without being able to estimate center and spread and show how that evidence supports his/her claim is not satisfying the ‘data is used as evidence’ requirement as well as a fourth grader with the same argument.

When looking to clarify the difference between formal and informal inferential reasoning, it appears that a fundamental differentiating factor is in the use of estimation; for example, a student comparing two samples by looking at their graphs could estimate a reasonable range for each of the population means and use
that information to make an informal inference. Other ambiguity in the definition could be addressed by stipulating that grade-appropriate analysis of data should be used as evidence rather than simply data as evidence, with grade-appropriate analysis being determined by the class curricula. To that end I propose the following definition of IIR: the use of informal statistical knowledge to estimate inferences about unknown populations wherein generalizations beyond data are made, grade-level appropriate analysis of data is used as evidence, and probabilistic language in expressing uncertainty about generalizations is present.

**Constructing assessment items and performing analysis.** Several findings of this nature were noted during the methodology and assessment sections of this dissertation. These findings include:

- readability and brevity in the posing of questions increased the students’ likelihood of comprehending and answering the questions,
- realistic (rather than fantasy or direct access) contexts were less likely to lead students to disregard data and focus solely on finding an answer from the context,
- a shift in context set up by multiple problems per task type better facilitates identification of students’ multistructural thinking by the AIIR, and
- a mix of growth factors, in terms of whether or not the number of new data points are multiples of the original number of data points, are needed in growing-a-sample tasks.
In the analysis it can become cumbersome to code everything then find the top-cycle codes, then see how they interact. Instead, the most efficient way of coding is to only look for evidence of thinking at or above the cycle level of the last coded structure. For instance, once cycle-2 reasoning is evidenced it may only be valuable to look for other cycle-2 reasoning or cycle-3 reasoning; cycle-1 reasoning becomes irrelevant in determining the SOLO level of the student (though it may be useful for other purposes depending on the nature of the research).

Limitations

The primary limitation of this study was the limited sample size available in order to design a feasible study, given the constraints on time, number of researchers and funding. These constraints limited the site selection to a single Midwestern middle school with low percentages of students on free or reduced lunch, and low percentages of minority students. As was mentioned in the literature review, the context of problems may only be regionally appropriate and future researchers looking to implement the instrument may need to take this into account.

Another limitation was that unanticipated questions arose after the initial co-measurability check that were not answered in the literature and for which data sources were insufficient to answer. The graph sense/IIR question is one such problem. There seems to be strong implicit support for graph sense and IIR to be closely linked in the literature, but not enough to draw a conclusion. Another unanticipated question came from the range of student SOLO levels. During the
analysis for the final co-measurability student reasoning varied from pre-structural
to $U_3$. Since the literature claims that IIR must be intentionally developed, it stands
to reason that some students have had more opportunities to develop their IIR than
others. If more was known about how such varying levels of reasoning occurred
within a single cohort of students it could inform interventions at lower grades, but
the disparity wasn’t realized until after data collection had finished so there was no
opportunity to collect even the most basic background information on the students.

Another limiting factor of the study was that the range of student ability
levels was insufficient to fully investigate cycle-3 (the Appropriate cycle), along with
the AIIR’s capacity to elicit these types of reasoning. Only one of the 24 students
from the co-measurability check portions was judged capable of cycle-3 reasoning
and that student was Unistructural. Of the additional students coded, cycle-3
reasoning rarely occurred (the classification of their reasoning required analysis
beyond simply identifying reasoning structures, which was not within the scope of
this study, so the exact number is not presented here). When cycle-3 reasoning was
exhibited, it was always in the question type involving estimating a population
graph (EPG). A possible explanation for the cycle-2 reasoning being only in the EPG
task type is that most curricula introduces single-sample single-population
questions or examples first, and that through their previous work in constructing
graphs or being introduced to measures of center that they may have had some
exposure to tasks of this type. However, without further research it is not possible
to determine this with any certainty.
Along with the limitation dealing with the scarcity of observed cycle-3 reasoning on the AIIR and SOLO adaptation, it is also worth noting that students at the Abstract level of reasoning were not observed at all. Without a clear view of what cycle 3 looks like, it is difficult to identify what types of reasoning would be classified as Abstract. However, it can be assumed that this type of reasoning would not be IIR as that is what the adaptation is meant to categorize, so there would need to be success in FSI tasks in order for students to be categorized as Abstract. Since the AIIR was only designed to elicit and record IIR it is unlikely that the AIIR would be able to identify students at the Abstract level.

**Implications and Open Questions**

In the limitations section it was noted that only a limited range of SOLO levels was observed in students. Based upon the limited range of observed SOLO levels, two implications emerge: more research is needed in order to confirm the SOLO adaptation in the Appropriate cycle (cycle 3), and middle school students’ IIR is underdeveloped. The GAISE report (Franklin et al., 2005) indicates that informal notions of statistical inference should be developed beginning in early elementary school. Of the 16 students from the final co-measurability groups, five were in the Pre-IIR cycle (cycle 1), 10 were in the Naïve IIR cycle (cycle 2), and one was in the Appropriate IIR cycle (cycle 3). This, taken with the findings of the literature review regarding statistical understanding as hierarchical (Watson & Callingham, 2003), suggests that students are not being prepared for the learning of formal statistical inference, which the CCSSM (NGACBP & CCSSO, 2010) and GAISE (Franklin et al.,
2005) designate as beginning in high school. This is an implication rather than a finding because of the sampling method used in the study.

The state of middle school students’ IIR from this research raises further questions: we don’t know how indicative these results are of middle school students in general. This particular sample was chosen to equally represent students at various mathematical ability levels. The sample was generated in order to maximize the opportunity to observe students at a wide level of IIR ability, and not to be representative of student abilities at that site. Through an informal review of 30 random student IIRs from the final co-measurability check, it appears as though students in the Naïve cycle (cycle 2) are over-represented, and students from the Pre-IIR cycle (cycle 1) are underrepresented. Again, since the group of students whose AllIRs were made available to the researcher was not random, generalizations can’t reliably be made from this data set. Further research is needed in order to understand what the state of middle school students’ IIR is.

As was mentioned earlier, the range of students’ ability levels was noteworthy. In Chapter 2 literature was presented claiming that IIR must be intentionally developed, specifically that it is not connected to students’ maturation, their level of education in related subjects, or developed through their life experiences (Zieffler et al., 2008, p. 4). However, among the 150 students whose AllIRs were made available, several cases of cycle-3 reasoning emerged. Given that IIR doesn’t spontaneously develop, what occurred in these students’ background that developed their IIR abilities? It is possible that answering this question could
provide ideas that could be leveraged into learning opportunities for students, or lead to a re-thinking of how IIR develops in students.

The under-developed IIR in middle-school students also raises the problem that, while work has been done particularly at lower grade levels, we don't know how we can get students’ with under-developed IIR abilities caught up in the upper grades. In reviewing the literature, there are several examples of researchers examining how to develop elementary-level students’ IIR (e.g., Langrall, Nisbet, Mooney, & Janssen, 2011; Makar, Bakker, & Ben-Zvi, 2011), but IIR development in older children was somewhat scarce (e.g., Jacob, 2013). Combined with this study’s implication that middle-school students’ IIR abilities are under-developed, there is a real need for better understanding of how IIR can be developed in upper grades.

The significance of under-developed IIR in students relates to another key aspect of research that remains to be conducted. This study showed progress in developing a way of measuring and classifying students’ IIR and provided potential tools for beginning to look at how students’ IIR proficiency affects their learning of formal statistical inference. Further progress might begin by looking longitudinally at how students’ IIR develops and what methods are most effective in developing it. This longitudinal type of research could expand past the cycle focus of this research to include the more Piagetian elements of the SOLO taxonomy in order to look at whether there are developmental factors that limit the growth of informal reasoning; while we know that IIR doesn’t develop with maturation, it remains to be determined whether or not maturation plays a limiting role in IIR development.
There was also a potential gender divide found in the results. Only 58% of female students involved in the final implementation were identified as having SOLO levels two or higher, whereas 89% of male students were identified as having SOLO levels two or higher.

These implications also raise concerns about the way in which inferential reasoning is developed in the CCSSM (NGACBP & CCSSO, 2012). There is evidence suggesting a severe deficiency in students’ IIR abilities that could interfere with their future learning of formal inference. However, the only analysis asked of students prior to middle school is arithmetic comparisons of data, which will in no way begin to correct the deficiencies seen in students’ preparations for the learning of formal statistical inference. Also, with the clarifying additions to the definition of IIR, it becomes even more apparent that appropriate IIR is grade-level specific; when this is taken with the idea of statistical learning being hierarchical, the new expectation of middle school students making an informal inference between two populations with no learning trajectory at lower grade levels to support IIR appears inappropriate.

In addition to the many research implications, the AIIR also has utility for practitioners. As an assessment, the AIIR can be administered prior to a sequence of instruction designed to develop students’ IIR, to gauge their prior knowledge, or as a formative or summative assessment for such a unit. The items could be adapted for task-based instruction at lower grades as IIR is being introduced, or could be used to engage or evaluate prior knowledge in later grades where the focus is on developing
formal statistical inference. The next two subsections present examples of how the AIIR items could be implemented as a formative assessment during a middle school unit, and as a measure of prior knowledge in a high school statistics unit.

**AIIR as formative assessment.** The task archetypes presented by Zieffler et al. (2008) were drawn from the literature on statistics learning at various grade levels. As a result, lessons grounded in similar literature will have goals and tasks that align well with AIIR tasks. For example, consider the Schoolopoly task first presented by Stohl and Tarr (2002). The Schoolopoly task is a recurring item in statistics education research (i.e. Rider & Lee, 2006; Weber, Maher, Powell, & Lee, 2008; Zieffler et al., 2008), and uses the statistical investigative cycle (Wild & Pfannkuch, 1999) to have students discover whether or not dice from a certain manufacturer are fair.

Identifying whether or not a set of dice is fair relates to the CBM task type because, in Schoolopoly-type investigations, students generate (or are given) data sets and are asked to identify whether the data could be a model of a fair or unfair situation. Task CBM2 in the final item bank (see Appendix E) could be assigned after the Schoolopoly investigation in an IIR unit to assess students’ understanding. In CBM2, students are presented with the context of a group of friends at an amusement park. The group of friends is about to get into one of six lines to enter the amusement park, and as people go through the lines they have the opportunity to win a prize. The group of friends wants to know if the chances of winning a prize are the same for each line, so they gather some data that are presented in the
problem. The task has the same goal of determining fairness as determined by comparing frequencies of six categorical variables, but is different enough in context to pose a new challenge to students. This would help teachers gauge the degree to which the learning objectives of the investigation were achieved before moving on to subsequent pieces of the unit.

**AIIR as a pretest.** In the Common Core State Standards for Mathematics (CCSSM) (NGACBP & CCSSO, 2010) for high school there is a strand called “Making Inferences and Justifying Conclusions” (p.81). The second subsection is as follows:

**Make inferences and justify conclusions from sample surveys, experiments, and observational studies**

3. Recognize the purposes of and differences among sample surveys, experiments, and observational studies; explain how randomization relates to each.

4. Use data from a sample survey to estimate a population mean or proportion; develop a margin of error through the use of simulation models for random sampling.

5. Use data from randomized experiments to compare two treatments; use simulations to decide if differences between parameters are significant.

6. Evaluate reports based on data. (NGACBP & CCSSO, 2010, pp. 81-82)

These four standards require students to have a developed understanding of variability, and represent further development of all three IIR task types. These items are a small subset of the 31 statistics and probability standards (including
standards for college and career ready students) at the high school level. Given their similarity, and that they represent a relatively small portion of total standards, it is possible that they would be addressed in the same unit.

Approaching a unit with those learning objectives would require careful consideration of student prior knowledge. This will be particularly pronounced in the years immediately following full CCSSM implementation while teachers are still trying to become accustomed to the changes in prior knowledge brought about by the changes in curriculum due to CCSSM in earlier grades. While a version of the AIIR assessing all three task-types was seen as insufficient for research purposes, it may suit the purposes of the classroom teacher preparing to teach such a unit.

In selecting the tasks for the version of the AIIR to be administered, the teacher could think carefully about what context types (see Table 5 on p. 45) or student tasks within a question would be most appropriate. In this hypothetical scenario the teacher notices that the students have difficulty seeing how statistics will be useful for them outside of high school. The teacher would draw primarily from the items with realistic context or contexts for which students have direct access. In this case, the teacher could use EPG1, CTS2, and CBM1. Item 3 from this version of the AIIR is presented in Figure 12; this item addresses the choosing between models task type and, like the other items in the set, uses a context that is designed towards realistic uses of IIR. The teacher could also have selected items that fit with tasks in the upcoming unit. This version of the AIIR would fit the assessment goals of the teacher and the versatility of having several items in each
task type provides the teacher with the ability to select items that best fit the needs of the students.

This year 150 students enrolled in Mathematics 101 at Cityville College. The students are divided into 10 classes of 15 students each. Math 101 is designed so that the typical student should receive a B, which is an 85%.

1A) The graph below shows the final percentage grades in one 15-student class. Fill in the graph with an additional 13 student.

Explain why you chose to add those points:

1B) In the rectangle below, draw an outline of what you think the shape of the graph would be if you included all of the students for this year. Be sure to include appropriate labels.

What made you choose this shape?

Figure 12. Item CBM1 from the final item bank

Student response analysis by the teacher would not likely use the analytical framework presented in this study, which is primarily useful as a research tool.
Instead, the teacher would likely be looking for common misconceptions or weaknesses in student reasoning that he/she could use to inform instruction.

**Closing Remarks**

I began the pilot study for this line of research with the intention of leveraging Zieffler et al.'s (2008) work into a useable instrument and applying it in the dissertation to investigate the development of students’ IIR in a unit about formal inference. Through further review of the literature related to context (i.e. Ben-Zvi, 2005; Cobb & Moore, 1997; Wild & Pfannkuch, 1999) and explanations (i.e. Gill & Ben-Zvi, 2011; Lampert, 1990; Lipton, 2004) I started to get a clearer picture of what work lay in front of me. At the end of the pilot I had created a single version of the IIR that used two contexts for the three task types. Through the course of the dissertation work the AIIR has undergone significant changes. The contexts were somewhat limiting, so an item bank was created using a variety of context types and more probing questions. From there it became apparent that multiple contexts and varieties of questions were needed, particularly for identifying Multistructural reasoning in students.

While I was unable to find any examples of other people attempting to create assessments based on the Zieffler et al.’s (2008) task framework, several examples exist in the literature of people adapting the SOLO taxonomy for a variety of uses (i.e., Reading & Reid, 2006; Watson & Moritz, 2000). By reviewing these examples, considering the recommendations by Zieffler et al. (2008), and reflecting on previous work I have done with the SOLO taxonomy, I was able to produce an initial
adaptation for the pilot study that required less modification than the AIIR. However, even with this background, student performance necessitated the inclusion of a third cycle. In this dissertation study the primary changes were in clarity of language, and in expanding the definitions to account for unpredicted student reasoning.

While the AIIR and SOLO adaptation were the ‘ends’ of this research study, they are intended as a ‘means’ for a much broader set of research that needs to be done in order for the education community to understand students’ learning of statistics. The GAISE report states that:

The surest way to help students attain the necessary skill level is to begin the statistics education process in the elementary grades, and keep strengthening and expanding students’ statistical thinking skills throughout the middle- and high-school years. (Franklin et al., 2005, p. 3)

This research provides a means for furthering our understanding of how to strengthen and expand those statistical thinking skills.
References


HyperRESEARCH (Version 3.5.4) [Computer software]. Randolph, MA: Researchware, Inc.


Appendix A: Initial AIIR Items

Chloe has been reading a science-fiction series about a family that is traveling through space and crashes in a jungle on a strange world. In the books the family describes some of the thousands of strange and interesting animals they find in the jungle.

Chloe wants to write a book report about the series, and decides that she wants to describe fifteen random animals that the family finds during their travels. The father in the book is a famous tracker, so there are a lot of good descriptions on the animal tracks that the creatures make. She decides to organize the animals based off of how many feet they have. Of the ten creatures she picks, three only had one foot, four had 3 feet, two had 6 feet, and one had eight feet, three had 5 feet, and two had 0 feet.

1A) The graph below shows the number of animals in her group that had a certain number of feet (there are four dots at the three, so there were four animals with three legs). Fill in the graph with what you think would happen if she included ten additional types of animals.

![Graph showing animal feet distribution]

Explain why you chose to add those points:

1B) On the graph provided below, draw what you think the shape of the graph would be if you included all of the animals in the jungle.

![Graph showing animal feet distribution with additional points]

What made you chose this shape?
Scientists can determine how many calories are in a piece of food by using a method called bomb calorimetry. Basically they set food on fire, and let it completely burn up. By measuring how much heat it produces, they can figure out how many calories are in a piece of food (don’t try this on your own, though).

Two 7th graders, Jose and Jacquelyn are at a summer science camp, and get to experiment with bomb calorimeters. Their teacher gives them each a bag of chips and they need to figure out which type of chips has more calories. They decide that they will each test 11 samples, using the bag’s serving size to determine the size of the sample. Jose runs out of time, though, and only gets to test nine samples. The graph below shows their results.

2) Is there a difference between the number of calories in Jose’s chips and Jacquelyn’s chips?

How certain are you?

Explain your answer.

3) Naomi was testing the same chips as Jacquelyn, and according to her data she estimated the number of calories for that brand of chips as 98. What do you think of Naomi’s estimate?

Explain your answer
Chloe is writing a book report. In her book a family is exploring other planets. The father in the book is a famous tracker, so there are a lot of good descriptions on the animal tracks that the creatures make. She decides to organize the animals based off of how many feet they have. Of the ten creatures she picks, three only had one foot, four had 3 feet, two had 6 feet, and one had eight feet, three had 5 feet, and two had 0 feet.

1A) The graph below shows the number of animals in her group that had a certain number of feet (there are four dots at the three, so there were four animals with three legs). Fill in the graph with what you think would happen if she included 11 additional types of animals.

Explain why you chose to add those points:

1B) On the graph provided below, draw what you think the shape of the graph would be if you included all of the animals in the jungle.

What made you chose this shape?
EPG 2

This year 150 students enrolled in Mathematics 101 at Cityville College. The students are divided into 10 classes of 15 students each. Math 101 is designed so that the typical student should receive a B, which is an 85%.

1A) The graph below shows the final percentage grades in one 15-student class. Fill in the graph with an additional 13 student.

![Graph showing final percentage grades]

Explain why you chose to add those points:

1B) In the rectangle below, draw an outline of what you think the shape of the graph would be if you included all of the students for this year. Be sure to include appropriate labels.

![Rectangle for drawing graph]

What made you choose this shape?
EPG 3

Evan is trying to figure out how many texts the typical teenager sends in a day. He sends surveys to the local schools asking students how many texts they send in a typical day.

1A) On the first day, Evan receives emails from 15 students and plots the data for the first day, shown below. If he receives 15 more responses tomorrow, what do you think his graph would look like? Add 15 more points to the graph to represent the new responses.

Explain why you chose those points:

1B) In the box below, draw what you think the shape of the graph would be if 500 students responded. Be sure to include appropriate labels.

What made you choose this shape?
EPG 4

Priya lives on the moon with her family, and takes the gravity tube to and from school each day. Since the tube is so new, it’s still somewhat unreliable. It has taken her as little as 5 minutes to get to school, but some days it can take up to 20 minutes!

1A) Priya keeps track of how long it takes her to get to and from school for a week (for a total of 10 trips), and graphs them. Her graph is shown below. If she graphs her trip time for an additional week, what do you think her graph would look like? Add 10 more points on the graph to represent the second week of travel time.

![Graph of travel time in minutes]

Explain why you chose these points.

1B) In the box below, draw what you think the shape of the graph would be if she kept track for an entire school year (assuming no improvements are made). Be sure to include appropriate labels.

What made you choose this shape?
CTS 1

Two 7th graders, Jose and Jacquelyn are at a summer science camp, and they are testing bags of potato chips to see how many calories they really have. Their teacher gives them each a bag of chips and they need to figure out which type of chips has more calories. They decide that they will each test 11 samples, using the bag’s serving size to determine the size of the sample. Jose runs out of time, though, and only gets to test nine samples. The graph below shows their results.

1) Is there a difference between the number of calories in Jose’s chips and Jacquelyn’s chips?

How certain are you?

Explain your answer.
CTS 2

40 students are testing a new educational video game. 20 students play the video game for an hour each day, while the other 20 read for an extra hour each day. At the end of one month they take a 20 point quiz to see whether the video game or the extra reading was more effective.

1A) Explain what the graphs would look like if playing the video game for an hour a day has the same affect as reading an extra hour a day.

2A) The testing was done at three additional schools, and the pairs of graphs are shown below. Number the graphs from 1 to 3, where 1 is the school where there is the best evidence that Reading and Video Game had different results, and 3 is the school where there is the least amount of evidence that Reading and Video Game had different results.

Explain why you chose what you did for “1”.
CTS 3

Cityville High School’s driver education program is trying to see if texting affects students’ driving ability. They set up an obstacle course and grade each student drives through once regularly and once while texting. The highest score possible was 30 points. The instructors made two graphs: one graph with the class scores while driving regularly, and one with the class scores while texting and driving.

1A) What do you think the graphs would look like if the students got lower scores while texting?

2A) The testing was done at a total of three schools and the pairs of graphs are shown below. Number the graphs from 1 to 3, where 1 is the school whose graph shows the best evidence that driving was affected by texting, and 3 is the school where there is the least evidence.

Explain why the graph pair you labeled as ‘1’ represents the best evidence.
Lafayette is an alien biologist working on Perseus 6. His job is to study an animal called an Apox. Apox are divided into two groups: the striped and the spotted. Lafayette thinks the spotted Apox are smarter. He has designed puzzle, and will measure how quickly the Apox are able to solve it. He makes one graph with the results for 20 striped Apox, and one graph with the results for 20 spotted Apox, shown below.

1) Is one type of Apox smarter than the other?

How certain are you?

Explain your answer.
CBM 1

Two 7th graders, Naomi and Jacquelyn are at a summer science camp, and they are testing bags of potato chips to see how many calories they really have. Their teacher gives them each a bag of chips and they need to figure out how many calories are in the bag. They decide that they will each test 11 samples, using the bag’s serving size to determine the size of the sample. Jacquelyn graphs her results, which are shown below.

1) Naomi was testing the same chips as Jacquelyn, and according to her data she estimated the number of calories for that brand of chips as 98. What do you think of Naomi’s estimate?
CBM 2

Kora is in charge of the snow machines at the Cityville Indoor Ski Slope. She keeps track of the snow produced each night, and plots the data on a graph. She suspects that one of the machines is beginning to wear out. While even perfect machines don’t produce exactly the same amount of snow every night, she knows that if the machines were operating correctly she should be able to get about 15 centimeters (cm) on a typical night.

1A) What would the graph look like if the machines were all operating correctly?

Why?

1B) The graph below shows the amount of snow made for each of the last 14 days. Do you think Kora has a broken snow machine?

Explain your answer.
CBM 3

You and a group of friends are going to the amusement park! This weekend when you enter the park, you pull a random ball out of a bag. If it’s orange you win a pass that lets you skip the line at any ride once. You and your friends want to get into the line with the greatest chance of getting a pass. Each of you picks a line and watches the first fifteen people go through, and counts how many people get an orange ball.

1A) There are six lines. Eight people win on lines 1, 4, and 6. Seven people win on lines 2 and 3. Nine people win on line 5. A graph of this data is shown below.

Do you think the lines all have the same chance?

Explain your reason.

1B) If you counted the winners for the first 50 people in each line instead of 15, how many winners do you think there would be in each line? (fill in the table)

<table>
<thead>
<tr>
<th>Line 1</th>
<th>Line 2</th>
<th>Line 3</th>
<th>Line 4</th>
<th>Line 5</th>
<th>Line 6</th>
</tr>
</thead>
</table>

Explain your reason.
Laura’s family farm is one of the only places in the galaxy that raises purple patchwork pigs. The value of the pigs are in how purple they are, and every year each pig is rated from 1-100 on how purple it is before it is sold. Last year they sold 1,000 pigs, and their purple scores are graphed below. The typical purple score was 72. This year the inspector came around and checked 10 random pigs, and their typical purple score was 76.

1A) Do you think Laura’s pigs will be generally more purple than last year?

Explain your reason.

1B) How confident are you with your answer in 1A?

What information would you need in order to be more confident?
Appendix C: Additional Initial Co-measurability Analyses

Alexis EPG2.

Alexis took version 1. When explaining her strategy in the EPG2 task she stated:

Alexis: I just added all my circles... the same as that one
Interviewer: So you just doubled the number of circles?
Alexis: Yes.

The student was then asked to present an answer for a situation where the number of additional data points used to grow the sample was not equal to the original number of data points. Alexis begins by adding one data point to each existing data value, leaving her with three additional points to distribute. She uses two of these points to make all of the data values other than mode have the same frequency, then adds the last data point to the mode. In this way she is coordinating a linear-increase strategy with a leveling-off strategy. Using these two cycle-1 strategies in a together identifies her strategy use in this problem as R₁, with her selection of the mode for the last data point suggesting a capacity for cycle-2 reasoning that went un-realized in this question, likely due to the item’s complexity.

This interview SOLO level was exceeded by the SOLO level from the AIIR, where her use of the doubling strategy resulted in her reasoning ability being classified as U₂. As was explained earlier in this section, the interview SOLO level is seen as having a lower bound at the assessed SOLO level, which resulted in interview SOLO level and assessed SOLO level of U₂.
Alexis CTS2.

The context for this item is seeing whether an hour a day spent playing an educational video game had a better affect on student reading than simply spending an extra hour a day reading a book. Alexis begins by expanding on her justification for a pair of graphs she created; her explanation shows that she extended the context and altered the question in order to arrive at a completely context-dependent response (coded as ‘personal-context’). However, when comparing pairs of graph to identify which pair shows the most evidence for population differences she is able to identify the mode as an important characteristic to look at, though she does not consider the shape in general or the distribution of data points, resulting in a $U_2$ code for her ability to work on CTS question types.

This contradicts her classification from the AIIR as $U_1$. The reason comes from the student’s interpretation of the question. In the task, the student was presented with the context of seeing whether an educational video game improved students’ reading ability more than an equal amount of time doing additional reading. The student was presented with three pairs of graphs each, containing a graph of outcomes from students who did the additional reading and a graph of outcomes from students who played the video game. She was asked to order them from showing the least evidence that the video game helped students improve more than additional reading to the most evidence that the video game helped students improve more than additional reading. Her response was unclear, and when asked to clarify what she meant she says “this one I think is different, because these two right next to this one are the same height, but this one only has two on it.” From her
explanation she was clearly comparing three graphs instead of three sets of graph pairs. I asked a question to clarify whether she was looking at which pair had the most difference between the two graphs, or if she was looking at which reading graph was the most different from a set of three. She had in fact been comparing the reading graphs to see which school had the most different reading graph from the other two. Since she did not support her answer when prompted to on the AIIR, her reasoning strategies were classified as being in the Naïve cycle.

Alexis CBM2.

In the CBM task type Alexis began by sketching a graph of what it would look like if there were four machines, one of which was broken, resulting in a lower output. The three machines working properly were drawn with identical output, with the broken machine showing about two-thirds the output of the properly operating ones. When asked whether or not the machines would have the same output, she responded ‘not every night,’ showing that she did allow for some variability, though it didn’t factor into her answers.

This limited allowance for variability placed her at the U2 level for CBM. The AIIR also identified her as U2 due to her being able to make general judgments about symmetry/shape (students at the Naïve cycle would say they should all be the same number), though her conclusion was incorrect.

Betsy EPG2.

Betsy also showed that she was able to use a doubling strategy, which is a cycle-2 (Naïve) strategy. When confronted with a situation where doubling was not possible, she added almost all of the new data points to the mode. She had been
counting all the data points outside the mode, and then counting the ones at the mode, and ran out of room above the mode; she is told that she can put additional points that belong at the mode below the graph, and ended up with the graph shown in Error! Reference source not found.. When asked why she picked those points she said “well, because the majority of people are supposed to be at that... uh, in between 83 and 88, and people that are over and under should be less than people that have Bs.” Taken with the graph it seemed she was trying to make sure at least half of the data was located at the mode. When asked directly she confirmed that this was her strategy. Since this is not a proportional strategy, her ability on the EPG task type was determined to be U₂. The same was found in the AIIR, where she was able to use a doubling strategy, though no other cycle-2 strategies were present.

Figure C-1. Betsy’s EPG response

Betsy CTS2.

Betsy’s strategies in comparing two samples focused solely on the location of the mode. She did not look at graph shape, and when asked what two samples from the same population would look like she said her goal in making the drawing was to
“make them exactly the same.” Her response was coded as one-relevant-aspect, a U\textsubscript{2} code.

This result did not match her AIIR result due to the same graph misinterpretation as Alexis, and her lack of a written response when asked to justify her decisions. The data that was collected from the AIIR was Betsy making the two graphs identical, which was coded as no-variability, a U\textsubscript{1} code.

**Betsy CBM2.**

In both her interview and AIIR, Betsy insisted that if the four machines are working properly they should be precisely the same, and that if they aren’t exactly the same then one of them is broken. This was coded as no-variability in both cases, a U\textsubscript{1} code.

**Chad EPG3.**

Chad was also able to use the doubling strategy when growing a sample, which was a U\textsubscript{2} code. When presented with growing the sample task where doubling was impossible, Chad also used the linear/leveling strategy that Alexis used, resulting in R\textsubscript{1} for that question. Overall, since he was able to effectively implement the doubling strategy he was considered to be U\textsubscript{2} for EPG tasks. This finding was the same in his AIIR, where he was also able to use the doubling strategy, but no other cycle-2 strategies.

**Chad CTS3.**

Chad’s strategies on the AIIR and interview for the CTS type were identical. He said that the pair of graphs with the highest modes had the most evidence, and the lowest modes had the least evidence. On the AIIR he writes “one has the highest
dots and three has the lowest,” and when he presented similar reasoning in the
interview he clarified that by higher dots he meant that there was a greater
frequency of responses at the data value. His reasoning was classified as $U_1$ for both
the assessed and observed SOLO levels because.

**Chad CBM3.**

Chad was consistent in his strategies in CBM also. He was working on the
amusement park problem where a group of friends counted the number of people
out of 15 who received a prize at each of six lines. He claimed that there was no
difference in the success rates of the six lines because they weren’t identically the
same height, and when asked to estimate what the graph would look like for a much
larger sample he attempted to combine whatever numbers he found listed in the
problem arithmetically in a somewhat undirected manner. Both responses were
seen as no-variability, a $U_1$ code.
Appendix D: Additional Final Co-Measurability Analyses

Estimating a Population Graph (EPG).

Alice observed. Alice was paired up with Avalana during instruction and posters. Their classroom interactions that resulted in codes predominantly exhibited a lack of allowance for variability and a lack of proportional reasoning. Their deterministic language on the class assignments and poster indicated a view of statistics as having a single set answer; for example, when analyzing student responses to the survey question ‘how would losing technology influence your life,’ they concluded that “there will be a very big impact on people,” which disregarded the bulk of the data and the shape of the graph in order to arrive at a single finding that they felt they could state unequivocally. Alice played a more minor roll in the group work, however, and exhibited higher levels of ability on individual work.

During the interview, Alice was presented with a series of questions centered around the contexts from the AIIR items. During the interview Alice was able to perform and explain growing samples using both the doubling strategy (cycle-2) and the shape-preserving-growth-vertical-only strategy (cycle-2). However, when working in direct access contexts or realistic contexts (see Table 5 on p. 45) she would occasionally lapse into the experience-trumps-data strategy; for example, when asked how she would grow a sample of test scores she said:

...there's usually two or three people with 100% on mostly everything, and then there's a good percent of the people that get As on everything or most everything, and then there's people that get Bs or higher, and then there's always the people that don't really do anything so they get lower.
She is observed to have access to two strategies in the Naïve cycle: shape-preserving-growth-vertical-only, and doubling. Since she was unable to coordinate the two strategies, and her understanding of variability didn’t indicate that she had reasoning structures at a higher level, her ability level on the EPG question type was considered to be \( M_2 \), or Multistructural in the Naïve IIR cycle.

**Alice assessed.** Alice’s AIIR test includes items EPG1, EPG2 and EPG3. In the EPG question types, the most telling response is in how students chose to complete the sub-items concerned with growing the sample. Each of these questions begins by presenting a context, then having the student grow a sample with 15 data points by 11, 13, and 15 for EPG1, EPG2, and EPG3, respectively. Alice’s responses are a good example of why three problems are preferable to two. In Alice’s response she shows that she is able to use the doubling strategy in EPG3, where it would be expected; however, she also uses it in EPG2 by including 15 additional points instead of 13. In EPG1, she does see that she needs 11 additional points instead of 15, and uses the shape-preserving-growth-vertical-only strategy. Since there were no more complex strategies used by Alice, her 2-item comparison utilizing EPG2 and EPG3 produced a code of \( U_2 \), whereas her observed SOLO level and the other three analyses of her AIIR (the two other pair-wise analysis and the 3-item analysis) indicate that her SOLO level for the EPG task type is \( M_2 \).

**Arthur observed.** Arthur was paired with Steve during the instruction and posters. During the group work, the pair showed little to know allowance for variability. While the class was coached on what types of questions could be productively answered with data, one of the questions suggested by Arthur was
‘how does the score board in the gym work?’, a question that is not suited to statistical investigation. During the statistical investigation using TinkerPlots (Konold & Miller, 2005) they were tasked with determining whether carnivores or herbivores were generally bigger. They drew all their conclusions from comparing a single carnivorous dinosaur to a single herbivorous dinosaur, even though they had already made the carnivore and herbivore graphs. On a separate TinkerPlots (Konold & Miller, 2005) task, they chose to base their answers completely on their personal experience rather than considering the graph they had constructed.

Finally, when they were asked on their poster what their graph would look like if all of the students from all six classes were included in the data set, they stated that ‘the graph will have six times our information,’ which was taken to mean that the frequencies in the resulting graph would be exactly six times the frequencies in their current graph, which exhibits no allowance for variability.

The allowance for variability is slightly improved on in Arthur’s interview. When asked further about why the graph for all of the classes resembled the graph for a single class (on the survey project), Arthur said ‘...our class’s data would probably be five times less than, about, than the whole team’s data,’ which, while showing little allowance for variability, doesn’t use completely deterministic language, and expresses some doubt that the frequencies will be precisely 1/5 of those in the population graph. Arthur was also able to grow samples of data using both the shape-preserving-growth-vertical-only strategy and the doubling strategy. Since he was unable to coordinate the two strategies, and his understanding of variability didn’t indicate that he had reasoning structures at a higher level, his
ability level on the EPG question type was considered to be M₂, or Multistructural in the Naïve IIR cycle.

**Arthur assessed.** Arthur’s AllIR test included items EPG1, EPG2, and EPG4 (see Appendix B). EPG1 and EPG2, as stated earlier, tasked the student with growing a sample of 15 by 11 and 13, respectively. EPG4 is an item that allows for a doubling strategy. On the first problem (EPG1) Arthur grows the mode by three data points, values with only one data point present by one data point, and all other existing data by 2. He does not include new data values in this sample growth, which shows the shape-preserving-growth-vertical-only strategy was used.

He uses this same strategy on EPG2, only without the same rule (he grows one value by two even though there is only one data point there initially, and grows a second value by one even though there were two there initially). This is again a shape-preserving-growth-vertical-only strategy, but one that relies less on a rule and more on the shape of the graph; while this is the same code as in EPG1, it does suggest that he is beginning to look at the way the shape grows rather than the frequencies at discrete data points, which could mark the beginning of a transition to allowing for horizontal growth of a sample. However, in EPG4 he simply uses the doubling strategy. There is no evidence in his explanations that any other strategies were used, or that the context was coordinated with thinking about the data in the student’s strategies. Since the strategies in EPG1 and EPG2 were the same, he was marked as U₂ for that pair-wise analysis, with all other analyses matching his observed SOLO level of M₂.
**Echo observed.** Echo worked with Reupert during the instruction and the poster. They showed a limited-variability strategy in their group work, and during their poster presentation they concluded that “a lot of people would have a larg (sic) impact when losing technology.” By having said “a lot of people,” this shows slightly more allowance for variability than Alice and Avalana’s statement about the same problem ‘there will be a very big impact on people.’ These preliminary signs of the students’ allowance for variability suggest they are operating at least at the Naïve cycle. However, in Echo and Reupert’s prediction for what the graph for all of the sections would look like, they say ‘6 more classes will make 6x more people per group,’ which indicates that while they may have shown some allowance for variability on the previous problem, it is something they still struggle with as a group.

During the interview, Echo shows that she is able to use the doubling strategy for growing a sample when the number of original data points matches the number of additional data points. She was then presented with two tasks for growing the sample when the number of additional points is different from the number of original points. In both cases she identified a general shape that the data formed, and grew the sample proportionally both vertically and horizontally to maintain the shape of the graph. This was seen as a cycle-3 strategy. However, when growing a sample within the context of EPG2, she created a graph with points exceeding 100%, which lead to the following exchange:

I: ...so what would this point here be?

E: 105
I: So you have one student getting 105%?

E: <laughs> yeah.

I: And why did you put that point there?

E: I don't know.

In this exchange she realizes that she has placed a point that defies the context and is unable to rationalize the point with the context. After the interviewer suggests that maybe there was extra credit Echo agrees that that would make the point reasonable, but from the exchange it is apparent that the context was not considered in her strategy for growing the sample. Because she used a single strategy her ability in the EPG task type was considered to be $U_3$.

**Echo assessed.** Echo completed the same version of the AIIR as Alice, which included EPG1, EPG2, and EPG3. In EPG1, the student is tasked with growing a sample, where the context is a record of alien animals organized by how many feet they have. In the data the majority of data points were at odd numbers, and in the pilot a common strategy was that a student would impose their own personal context and grow only even numbered values. In this case, while all but one data point was at or below 6, the student included data points only at or above 7. Her explanation was that it made a pattern, but since the new pattern did not reflect the shape of the original data it was coded as a cycle-1 strategy of experience-trumps-data.

On EPG3, where doubling is a possible strategy, Echo used doubling; this is a cycle-2 strategy. However, on EPG3 she noticed that the general shape of the data was a triangle, so she grew the sample both vertically and horizontally to keep the
triangle shape. This is a cycle-3 strategy, and the highest level strategy that she exhibits. In her response, she does include points on the graph that defy the context, so her reasoning took into account the data only. Because her cycle-3 strategy only showed up on one question, her pair-wise analysis that didn't include EPG3 resulted in a SOLO level of U₂. All other analyses resulted in a SOLO level of U₃, which was consistent with the observed SOLO level.

**Choosing Between Models (CBM).**

**Haley observed.** Haley worked with Johny on the instruction and presentation portions of the class. Haley and Johny showed a tendency to base conclusions solely on context during the TinkerPlots (Konold & Miller, 2005) section. One example was when the students were asked to use a graph to decide if herbivorous or carnivorous dinosaurs tended to be larger. To answer the problem they didn’t even attempt to construct a graph, they simply stated that “the one who eat meat will be bigger because they get more protein.” They also showed a limited understanding of variability during their class work on the poster by attempting to answer a question that required a statistical investigation with a deterministic answer (people said they used their phone the most).

Haley’s difficulties are further shown in the interview. Haley consistently showed no allowance for variability. For example, in the amusement park context a group of friends count how many people win a prize out of the first 15 people in each line.

I: So this is saying that after fifteen people went through lane one, eight of those people got an orange ball.
H: Okay.

I: So they’re watching the first line, and eight people get an orange ball. If they watch another fifteen people, not the same, a new fifteen people, how many orange balls do you think they will get?

H: Eight, the same amount.

I: The same amount, so here will be the same way, so fifteen people go through lane two... (lane two had seven people win of the first fifteen people)

H: Seven.

She also tended to expand or alter the context of a problem so that she could draw a conclusion without using the data, and when she was confronted with situations where her expectation based off of experience contradicted what she identified in the data she would back the conclusion based off of her experience over what could be seen in the data. This was particularly evidenced in talking about the purple pig problem. In the purple pig problem, a family raises and sells pigs whose price is determined by how purple their skin is. Students are given the ‘purple score’ of a small sample, along with the population graph of the pigs’ purple scores from the previous year, and asked whether the current year’s pigs are more purple than the previous year’s. The sample from the current year had a score of 76. When asked what some reasonable values to expect for the population score Haley noted that they could be as low as 67 or as high as 80. She was then asked whether the sample suggested the pigs would be more purple this year, and she that she would need more information to attempt to answer the question, and the information she wanted was what made the pigs purple. In this way, her valid reasoning about the
problem isn’t used to answer the problem because it is over-ruled by the context. This experience-trumps-data strategy is seen as a Pre-IIR strategy because, while it does over-use the context it also requires the student to be able to identify some relevant aspects of the data (though in this strategy they are disregarded). Haley was able to use only one strategy in the Pre-IIR cycle, leading to her ability classification as U₁ for the CBM question type.

**Haley assessed.** Haley’s AIIR test includes items CBM1, CBM3, and CBM4. Haley’s reasoning showed no allowance for variability. When presented with the amusement park context, where all of six lines had 6-8 successes of 15 trials, and asked if the lines have the same chance for success, she responds that they don’t because some lanes have more (successes). Generally she tries to answer questions solely from the context. In the purple pig context, she says that there is a pattern, but to be more certain of whether or not this year’s pigs are better than last year’s she would need to know what causes pigs to be purple; students who consider the data in their response respond that in order to be more certain they need a bigger sample (in cycle-1 responses typically wanting to see data for the entire population). In the calorimetry context rather than wanting more samples of chips she wants to know what brands the chips were. This over-reliance on context and lack of understanding of variability place Haley’s SOLO level at U₁ for CBM question types.

**Hannibal observed.** Hannibal worked with Phylicia during the instruction and project portions of the study. During their group work they answered questions very quickly with minimal conversation. Typically their decisions were based
around the context. For example, on the poster they did use probabilistic language when describing their results, which would indicate that they may have some capacity for Naïve or Appropriate reasoning, but their reasoning structures didn’t show any evidence of reasoning at these levels. For example, The primary place CBM reasoning showed up in general was when one student would suggest a statement for a set of data and the other student would evaluate whether the statement fit the graph/model, but what primarily occurred in this group was that Hannibal would give his answer without reasoning (primarily to EPG type tasks), and Phylicia would record them without discussion or challenging Hannibal’s response.

Hannibal’s reasoning during the interview showed further evidence of an inability to reason at higher levels for the CBM problem type. A primary difficulty Hannibal encountered was an inability to read graphs. Hannibal was presented with histograms and bar graphs, and consistently interpreted them to be change-over-time graphs. Once the graphs were explained and he expressed understanding and could read data from the graphs, his strategies were undirected-arithmetic and experience-trumps-data, a pre-structural and a Pre-IIR cycle code respectively. For this reason Hannibal was classified as U₁ for the CBM problem type.

**Hannibal assessed.** Hannibal’s AIIR included items CBM1, CBM2, and CBM3. Hannibal’s reasoning showed evidence of being at cycle 2 on two questions. On CBM3, when presented with the lane data graphs he did say that the lines did not have the same chance, but in his explanation he gives the reason that “lane five has nine wins while two and three have seven.” His reasoning wasn’t that the lanes
weren't all the same, it was that there was a difference of two wins between some lanes and others. This shows that he does have some early conception of variability. Again in CBM2 he generates a graph to model the situation where all six machines are working properly, and he allows for some variability in machine performance. Because this limited-variability was used as a strategy for answering questions, all of the item analysis on Hannibal’s AIIR resulted in a SOLO level of U₂.

**Steve observed.** Steve worked with Arthur during the instruction and project portions of the research. Their group work is presented in Arthur’s observed section of the EPG question type, but essentially the group work was primarily in the Pre-IIR cycle.

During the interview, Steve also predominantly used strategies typical for Pre-IIR cycle students. He was able to demonstrate a limited understanding of variability, however. While working in the amusement part context, Steve was asked what would happen if a single line with 8 successes was re-counted for the next 15 people. In this case a student with no allowance for variability would be expected to say exactly eight, whereas a student with a strong grasp of variability would be expected to say a wider range with different probabilities within the range. Steve showed a limited-variability strategy by stating that it would either be 7, 8, or 9, without commenting on the probability of any of the answers. Since Steve was able to incorporate his limited understanding of variability into responding to a question, he is seen as U₂ for the CBM question type.

**Steve assessed.** Steve’s AIIR, like Haley’s, included items CBM1, CBM3, and CBM4. Steve’s response to CBM1 highlighted a weakness in the problem in that
there wasn’t enough probing for student reasoning. In CBM1 students are told that Naomi performed an experiment and estimated the calories per serving of a type of chips to be 98. They are then presented with data organized in a histogram from a second experiment done by Jacquelyn on the same type of chips. When asked what he thought of Naomi’s estimate, Steve states “I think Naomi’s estimate may be off because Jacquelyn’s is higher.” On one hand this could show no allowance for variability due to Steve implying that since Jacquelyn’s chips appear to be typically higher there is a problem with Naomi’s estimate regardless of how large the difference is (the difference is impossible to precisely determine since the data is presented as a histogram only), or he could be suggesting through the use of probabilistic language that it is difficult to determine if there is a difference given the sample sizes.

This ambiguity is dealt with in CBM4 where the item further prompts the student to express how confident they are in their answer and what could make them more confident. In CBM4 Steve states that a larger sample size could make him more confident. This shows that there is some type of understanding of variability. However, in the same problem he also claims that a sample mean could not come from a sample of the population graph and that he was very certain of the fact, showing that while he does have some concept of variability and that it is integrated into his solution strategy, it is a limited understanding. This limited concept of variability was further evidenced in the amusement park line context (CBM3). The amusement park is giving guests the opportunity to win a prize when they enter the park, and a group of friends wants to see if there is a line they can go
through that would give them a higher probability of winning. The friends count how many out of 15 people win in each line, and the result is that either 7, 8, or 9 people win in the various lines. He only mentioned the highest frequency line as having a greater probability, implying that the single win difference between the other lines did not seem like a large enough difference to him. Because his use of the limited-variability strategy in his responses in CBM3 and CBM4, Steve’s SOLO level for the CBM question type is $U_2$.

**Larry observed.** Larry was Scoot’s partner during the instruction and project portion of their research. Their group work is explained in more detail during Scoot’s subsection (page 77), with the result that there was not much usable evidence from this portion of the research from these two students. Larry, like Hannibal, had fundamental difficulties in reading and interpreting graphs, tending to interpret all graphs as change-over-time graphs. With some help, Larry was eventually able to understand the graphs in the interview, but he continued to struggle with the material. An example of his reasoning comes again from the amusement park context. He expressed that the lines must be fair because they are a business and they should be fair to their customers. He says it’s not that he saw that in the graph, but it’s just how things should happen. When asked which line he would go in he said he would go in one of the lines that had more winners, but still insisted that the lines all had the same chance. This resulted in the ‘Altering question for personal use’ code. However, later in the interview he did note that if a line got 8 winners? for the first 15 guests he would expect it to get 7, 8, or 9 over the course of the next 15 guests. In this way his reasoning is similar to Steve, which
shows that he has a Naïve IIR strategy. Because of his ability to use his limited understanding of variability to form a conclusion, his ability level on CBM tasks is $U_2$.

**Larry assessed.** Larry’s AIIR test, like Hannibal’s, included items CBM1, CBM2, and CBM3. Larry’s responses highlight another limitation of this type of assessment in that only one of his responses had more than three words. His entire response to all of CBM1 (the chip calorie context) was “that is good,” and there simply isn’t much that can be inferred from that response about the student’s reasoning.

On CBM2 when asked what a graph of six fully functioning machines would look like he responded “lots, like 45” and when asked to justify his response he notes that “because $3 \times 15 = 145$.” While the 15 likely comes from the problem stating that the machines produce 15 cm of snow on a typical night, it is unclear where the three came from as neither the number three or the numeral 3 appear anywhere on the page. Without more of an explanation it’s impossible to say with any reasonable amount of certainty what his strategy was on that portion of the assessment. His longest response comes later in that question when asked to justify why he thinks there is a broken machine; he says “because it is very large at times and very small at times,” which would ideally mean there was a lot of snow sometimes and only a little other times, but commonly pre-IIR and cycle-1 students often confuse frequency with data, and if that is the level he is operating at he might not be considering data values at all.
On CBM3 he could have been showing the capacity for higher cycle thinking, though. When presented with the graphs of success rates in the different lines and asked if they have the same chance he simply says “yes.” In this problem when a student was able to identify that the lines could have the same success rate it was evidence that the student had some concept of variability. The difficulty in this case being that without any justification for the answer it is impossible to tell whether the student’s strategy incorporated some concept of variability, or if the student’s strategy was, perhaps, that he had a 50% chance of getting the ‘correct’ answer and chosen randomly. The decision was made in this case to not give the student the ‘benefit of the doubt’ in order to hold the assessment to a higher level of scrutiny. Without further explanation and given that he uses an incorrect algorithm to populate a second sample table, his reasoning is seen as being limited to cycle-1. For this reason he is identified as having SOLO level U1 for the CBM task type for all analysis except for the pairwise analysis of CBM1 with CBM2, from which there was insufficient evidence to assign a SOLO level.

**MJ observed.** MJ worked with Sofia on the instruction and project portions of the research. During the TinkerPlots (Konold & Miller, 2005) activity, they began by making a prediction about how cereal would be arranged on the shelf. They were then directed to construct a relevant graph and evaluate how well their prediction matched the sample. In their discussion on how well the model fit their prediction they used probabilistic language to describe the graph, which indicates the potential for 2nd or 3rd cycle reasoning structures. They did tend to favor context for reaching
conclusions from the data, but not to the extent that data was ignored. This is also more indicative of 2\textsuperscript{nd} cycle reasoning.

Second cycle reasoning is supported further by their poster results. On their poster their prediction is fairly deterministic: “You can abuse technology by dropping it and breaking it. Also you could look up bad things or do bad things on it.” This type of deterministic response generally indicates a low or non-existent allowance for variability in the data by indicating that there is one exactly correct answer for questions that require statistical investigation. However, they also note that by performing the survey you could get a large list of ways people could abuse technology since the responses could vary, and summarizes the results of the study by listing the ‘top contenders.’ This shows a progression in their understanding of variability and the statistics question distinction through the course of the unit, finishing in reasoning that is typical of students in the Naïve IIR cycle.

MJ’s interview shows further evidence of early Naïve cycle thinking. He still tended to over-use the context in arriving at conclusions, and he was seen as letting experience trump data. However, he also showed that he was able to implement a limited view on variability into his reasoning structures, and his responses benefited from this increased reasoning ability. On one hand, when asked how he could be more certain about how well a sample supported a prediction for the population he said he’d need to have data on every single member of the population; on the other hand, he was able to identify that if the population parameter was a certain number, samples could result in slightly higher or lower numbers. This
ability to implement his moderate understanding of variability in his reasoning about CBM tasks placed MJ at the U₂ reasoning level.

**MJ assessed.** MJ’s AIIR included items CBM2, CBM3, and CBM4. MJ’s reasoning is a mix of cycle-1 and cycle-2 thinking. On CBM2, MJ explained why he thought a snow machine was broken by saying “because if they shoot 15cm and only shoot about 8cm in 14 days then it does not work right.” Students with no conception of variability, or cycle-1 students, typically say that a machine must be broken because there are a variety of values, but MJ focused on the lowest value. In CBM3 he also identified that the lines could all have the same probability of success, though his justification “if they are in a line with a lot of winners” did not provide enough evidence to support anything higher than cycle-2 reasoning.

He did have some struggles though. In CBM3 he is presented with the amusement park context, in which a group of friends record the number of passes won out of 15 people in each of six lines. He correctly posited that the lines could all have the same probability for winning a prize, but when asked to estimate how many people would win in each line for a sample size of 50 rather than 15 he produced the table shown in Error! Reference source not found.. In the original sample lines 1-6 had 8, 7, 7, 8, 9, and 8 successes respectively. His model of a larger sample size kept roughly the same number of successes as a smaller sample, and also included more variability than the smaller sample; for this reason his concept of variability is seen as limited in CBM3.
Figure D-1. MJ’s response on CBM3 at increased sample size task.

In CBM4 there was further evidence that his conception of variability was limited. In this question he based most of his answers around the context. When asked what information he would need to feel more confident of his answer, rather than asking for a larger sample size or population information, he asks for a graph that shows the next year’s data. Because he did have some conception of variability that was exhibited in both CBM2 and CBM3, MJ was classified as having \( U_2 \) ability level for CBM tasks.

Comparing Two Samples (CTS)

**Avalana observed.** Avalana worked with Alice during the instruction and project portions of the study. A more thorough description of this section of the research is shown in Alice’s subsection of the EPG portion, but essentially the group work, in which Avalana played a prominent role, showed no allowance for variability and a lack of proportional reasoning.

During the interview, Avalana was presented with three pairs of graphs, each pair having one graph of a control type situation and one graph of a treatment. She was asked to indicate which pair of graphs showed the most evidence for a decrease in performance from the control to the treatment. She selected a pair of graphs that were completely identical except that one had a low outlier and one had a high outlier. Her justification was that she picked the one that “had more dots at the beginning than at the end,” pointing on the paper to the outliers. No other strategies
were evident in the interview, though it is worth noting that class announcements, an interruption, and a non-standard schedule for the day resulted in the interview being approximately 10 minutes shorter than intended. Regardless, the evidence indicated that Avalana’s capability for CTS question types was at the $U_1$ level.

**Avalana assessed.** Avalana’s AIIR includes items CTS1, CTS2, and CTS3. Her responses gave very clear examples of reasoning from a cycle-1 student. On CTS1 Avalana showed that she had personalized the context to something she was able to answer without data by responding that “Jose did not eat as many chips as Jacquelyn did,” where the context is concerned with testing chips for the caloric value of a serving size; this shows that she was creating a new personal context in order to answer the question, which is a pre-structural strategy. In CTS2 she says that graphs for two different treatments would indicate no difference in their affect if they were completely identical, and her graph comparison strategy in CTS3 is to pick a small range on the graph (15-20 in this case) and count how many data points fall in that range for each graph. In CTS1 and CTS2 she uses a strategy of assessing whether graphs are identical or not, a cycle-1 strategy, and in CTS3 she uses the compare-an-irrelevant-interval strategy, also a cycle-1 strategy. Because of these two different strategies her reasoning is classified as $M_1$ for all analysis except for the pair-wise analysis utilizing CTS1 and CTS2, in which she is classified as $U_1$.

**D’Jackson observed.** D’Jackson was intended to work with Dynklfrin during the instruction and project portion of the class. Dynklfrin was going through a distressing social situation and spent most of the class time off topic or in another area of the room. D’Jackson’s work indicated a very limited allowance for
variability, particularly in the poster where he focuses on only the modal survey response and presents his findings using deterministic language. He also says that a graph containing all classes would be ‘much larger,’ but doesn’t mention anything about the features of the graph.

D’Jackson’s interview continued to show his difficulty with variability. When asked to construct a graph for intelligence test results from two separate populations, where the populations had the same intelligence, D’Jackson constructed the graph shown in Error! Reference source not found.. The graphs are both completely identical, and have uniform frequencies for each point at which there is data. This showed that D’Jackson did not have any implementable conception of variability. However, when asked to judge whether to graphs presented the same results for two different subsets of a population, he was able to identify that they did both have the same primary result, and based his comparison on a relevant feature: the mode. Because he was able to compare two samples using the one-relevant-aspect strategy, he is operating at the $U_2$ level for the CTS task type.

Figure D-2. D’Jackson’s CTS interview response

**D'Jackson assessed.** D’Jackson’s AIIR test includes items CTS1, CTS2, and CTS4. D’Jackson has two strategies he uses for comparing two samples of data. In CTS2, D’Jackson was tasked with identifying the pair of graphs that represented the
most evidence that one treatment was better than the other. D’Jackson wrote that he made his decision “because the scores are a lot more spread out.” While looking at the variability is certainly something we hope students take into account, the variability in all six graphs were fairly similar (the graph with the least variability had a standard deviation of approximately 2.4, and the graph with the most variability had a standard deviation of approximately 3.3). Since the spread of the data was not a relevant aspect to base an entire decision on, this was seen as a cycle-1 strategy.

D’Jackson did use one cycle-2 strategy when comparing two samples. In CTS4 D’Jackson uses the value eight as a central value of both graphs, then counts how many data points each graph has above 8. He concluded that the graph that had more data values above 8 must be from a population that had higher scores. While this strategy would not work for comparing samples of different sizes, it was seen as a cycle-2 strategy in this case. Since CTS4 was the only item where a cycle-2 strategy was used, all analysis except for the pairwise analysis that omitted CTS4 resulted in a SOLO code of $U_2$ for D’Jackson in the CTS problem type.

Dynklfrin observed. Dynklfrin was partnered with D’Jackson for the instruction and project portion of the research. Dynklfrin began the instructional unit as the recorder, but due to his mistreatment of the technology the first day’s data was lost (at various points he dropped, played drums with, kicked, and stepped on the LiveScribe Sky pen (Livescribe, 2012), causing all data from that day for that pen to be lost); since this was the only day he was engaged in the group work, there is no usable data gathered on Dynklfrin outside of the interview.
Dynklfrin’s strategy during the interview was based primarily around comparing frequencies. For instance, when presented with two graphs, each symmetrical around the same point with the same range and only slightly different shapes, Dynklfrin stated that the graphs were ‘much different’ because one had higher numbers. Because of this, Dynklfrin’s reasoning on CTS tasks was seen as U1. However, it is worth noting that Dynklfrin’s Observed SOLO level is based off of only interview data and interview artifacts.

**Dynklfrin assessed.** Dynklfrin’s AIIR test includes items CTS1, CTS2 and CTS3. Dynklfrin’s response to CTS2 showed somewhat mixed results. On part A he describes what the graphs for the samples would look like if the two treatments did not produce different results by stating that they would be “the same grade, and the same accuracy (sic),” showing a lack of allowance for variability. However, on part B when asked to identify the pair of graphs that are most different, he selects the pair that has the largest difference in the mode and lower variability than is present in the other pairs of graphs. For his explanation he simply writes his decision was “based on the shape of the graph’s (sic).” His reasoning in this problem was seen as U2 based on the strength of his choice but limited by the lack of a detailed explanation.

In CTS1 we see again find evidence that Dynklfrin does have some cycle-2 strategy for dealing with the comparison of two samples, and again it is comparing the mode. This time it is more explicitly explained by his reasoning that in the graphs he is comparing one has a mode of 110 and one has a mode of 105 (which is a slight misreading of the histogram wherein one has a mode in the 100-105
interval and the other has a mode in the 105-110 interval). This is again coded as a cycle-2 strategy, particularly because while the modes were different, they were in adjacent intervals on the histogram, which would not have been seen as significant if the student had a more appropriate concept of variability. Since CTS1 and CTS2 both showed the same cycle-2 strategy, all analysis for Dynklfrin result in a SOLO level of U₂ for the CTS problem type.

Phyllicia observed. Phyllicia was partnered with Hannibal for the instruction and project portions of the research. As was stated in Hannibal's subsection of the CBM section, while the group did appear to have some limited understanding of variability, it never factored into the decision making for answering questions about data. Instead the pair tended to make most of their decisions based on the context of the problem.

Phyllicia continued to have difficulties in the interview portion. Like Hannibal and Larry, Phyllicia had difficulties interpreting and reading data from graphs. She, too, attempted to interpret the graphs as change over time rather than data frequency graphs. Once the graphs were explained and she expressed understanding, the strategy she used when comparing graphs was based on comparing the height of the graphs. She considered a pair of graphs to have more evidence if the overall height of the graph was greater. Because this was the only strategy she could use, her ability level was classified as U₁ for the CTS task type.

Phyllicia assessed. Phyllicia’s AIIR test includes items CTS1, CTS2, and CTS4. Phyllicia’s reasoning was very consistent across all items; she compared graphs entirely based off of which one had the highest frequency. In CTS1 she judged that
there was a difference “because Jose’s graph is a little shorter,” in CTS2 she noted that if the treatments had the same result “the graphs would be a lot higher,” and in CTS4 she says there is a difference because one group had more scores than the other (since the sample sizes and modes were the same this was interpreted as being about the difference in the ‘height’ of the graphs). Because of the consistent use of this single cycle-1 strategy, Phylicia’s SOLO level for the CTS question type was classified as U₁.

**Sofia observed.** Sofia was paired with MJ for the instruction and project portion of the research. Their reasoning is written up in greater detail under MJ’s subsection of the CBM section, but their dominant strategy was an over-reliance on context, with very limited allowance for variability.

During the interview Sofia had multiple strategies for comparing two samples. Her most-often used strategy was the same frequency strategy as Phylicia: she would say that the taller a graph was, the more evidence it had, and that the greater the difference in height, the greater the difference in the data. In one question she also mentions that in addition to how tall the stacks were she also looked at how spread out they were. However, since she was looking at the spread of the data set with respect to how tall the graphs were, this was still seen as a Pre-IIR cycle strategy. However, when asked to draw two new graphs where the samples came from similar populations (in this case samples from two types of animals that had similar intelligence measures), she drew graphs that made sense with her previous strategies, but then mentioned that since one graph had a point that was higher than usual she put a lower one in to balance it out. This shows that
she was making comparisons based on some informal idea of central tendency, which lead to her reasoning about CTS tasks to be classified as $U_2$.

**Sofia assessed.** Sofia’s AIIR test includes items CTS2, CTS3, and CTS4. In CTS2 Sofia shows a lack of allowance for variability in that when explaining what the graphs would look like if the treatments had the same affect she said that “the graphs would be the same.” She makes the comparison in part B of that problem providing only the justification that “it seemed the closest to the same for me.” No SOLO code was identified for Sofia on CTS2.

In CTS3 she makes the same decision as described earlier of using only the spread of the data to make her decision. In CTS4 she used the strategy of selecting a cut-off and counting the number of data points above that point. Recall, in previous cases when that cut-off was seen as near the middle it was a cycle-2 strategy, but when it was more arbitrarily or inappropriately chosen it was a cycle-1 strategy. In this case she selects a point that lies in or near the lower quartile for both of the graphs and counts the number of data points above the cut-off; this is seen as an inappropriate choice for a lower cut-off value and therefore a cycle-1 strategy. This use of two different cycle-1 strategies in two distinct problems resulted in an $M_1$ SOLO code for the three-item analysis and the pair-wise analysis of CTS3 and CTS4. The remaining two pair-wise analysis resulted in a SOLO code of $U_1$. 


Appendix E: Final Item Bank

EPG 1

This year 150 students enrolled in Mathematics 101 at Cityville College. The students are divided into 10 classes of 15 students each. Math 101 is designed so that the typical student should receive a B, which is an 85%.

1A) The graph below shows the final percentage grades in one 15-student class. Fill in the graph with an additional 13 student.

Provide a description of the graph and why you chose to add those points:

1B) In the rectangle below, draw an outline of what you think the shape of the graph would be if you included all of the students for this year. Be sure to include appropriate labels.

Explain what made you choose this shape:
EPG 2

Evan is trying to figure out how many texts the typical teenager sends in a day. He sends surveys to the local schools asking students how many texts they send in a typical day.

1A) On the first day, Evan receives emails from 15 students and plots the data for the first day, shown below. If he receives 15 more responses tomorrow, what do you think his graph would look like? Add 15 more points to the graph to represent the new responses.

![Graph showing text counts per day](image)

Explain why you chose those points:

1B) In the box below, draw what you think the shape of the graph would be if 500 students responded. Be sure to include appropriate labels.

![Blank graph](image)

What made you choose this shape?
EPG 3

Priya lives on the moon with her family, and takes the gravity tube to and from school each day. The tube is somewhat unreliable. It has taken her as little as 5 minutes to get to school, but some days it can take up to 20 minutes!

1A) Priya keeps track of how long it takes her to get to and from school for a week (for a total of 10 trips), and graphs them. Her graph is shown below. If she graphs her trip time for an additional week, what do you think her graph would look like? Add 10 more points on the graph to represent the second week of travel time.

![Graph](image)

Explain why you chose these points.

1B) In the box below, draw what you think the shape of the graph would be if she kept track for an entire school year (assuming no improvements are made). Be sure to include appropriate labels.

![Box](image)

What made you choose this shape?
CTS 1

40 students are testing a new educational video game. 20 students play the video game for an hour each day, while the other 20 read for an extra hour each day. At the end of one month they take a 20 point quiz to see whether the video game or the extra reading was more effective.

1A) Explain what the graphs would look like if playing the video game for an hour a day has the same affect as reading an extra hour a day.

2A) The testing was done at three additional schools, and the pairs of graphs are shown below. Number the graphs from 1 to 3, where 1 is the school where there is the best evidence that Video Games had better results, and 3 is the school where there is the least amount of evidence that Video Game had better results.

Explain why you chose what you did for “1”.
CTS 2

Cityville High School’s driver education program is trying to see if texting affects students’ driving ability. They set up an obstacle course and grade each student drives through once regularly and once while texting. The highest score possible was 30 points. The instructors made two graphs: one graph with the class scores while driving regularly, and one with the class scores while texting and driving.

1a) What do you think the graphs would look like if the students got lower scores while texting?

1b) The testing was done at a total of three schools and the pairs of graphs are shown below. Number the graphs from 1 to 3, where 1 is the school whose graph shows the best evidence that driving got worse while texting, and 3 is the school where there is the least evidence.

Explain why the graph pair you labeled as ‘1’ represents the best evidence.
CTS 3

Lafayette is an alien biologist working on Perseus 6. His job is to study an animal called an Apox. Apox are divided into two groups: the striped and the spotted. Lafayette thinks the spotted Apox are smarter. He has designed puzzle, and will measure how quickly the Apox are able to solve it. He makes one graph with the results for 20 striped Apox, and one graph with the results for 20 spotted Apox, shown below.

2) Is one type of Apox smarter than the other?

How certain are you?

Explain your answer.
CBM 1

Kora is in charge of the snow machines at the Cityville Indoor Ski Slope. She keeps track of the snow produced each night, and plots the data on a graph. She suspects that one of the machines is beginning to wear out. While even perfect machines don’t produce exactly the same amount of snow every night, she knows that if the machines were operating correctly she should be able to get about 15 centimeters (cm) on a typical night.

1A) What would the graph look like if the machines were all operating correctly?

Why?

1B) The graph below shows the amount of snow made for each of the last 14 days. Do you think Kora has a broken snow machine?

Explain your answer.
CBM 2

You and a group of friends are going to the amusement park! This weekend when you enter the park, you pull a random ball out of a bag. If it’s orange you win a pass that lets you skip the line at any ride once. You and your friends want to get into the line with the greatest chance of getting a pass. Each of you picks a line and watches the first fifteen people go through, and counts how many people get an orange ball.

1A) There are six lines. Eight people win on lines 1, 4, and 6. Seven people win on lines 2 and 3. Nine people win on line 5. A graph of this data is shown below.

![Graph showing passes won in 15 attempts](image)

Do you think the lines all have the same chance?

Explain your reason.

1B) If you counted the winners for the first 50 people in each line instead of 15, how many winners do you think there would be in each line? (fill in the table)

<table>
<thead>
<tr>
<th>Line 1</th>
<th>Line 2</th>
<th>Line 3</th>
<th>Line 4</th>
<th>Line 5</th>
<th>Line 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Explain your reason.
Laura’s family farm is one of the only places in the galaxy that raises purple patchwork pigs. The value of the pigs are in how purple they are, and every year each pig is rated from 1-100 on how purple it is before it is sold. Last year they sold 1,000 pigs, and their purple scores are graphed below. The typical purple score was 72. This year the inspector came around and checked 10 random pigs, and their typical purple score was 76.

1A) Do you think Laura’s pigs will be generally more purple than last year?

Explain your reason.

1B) How confident are you with your answer in 1A?

What information would you need in order to be more confident?
Appendix F: Human Subjects Institutional Review Board Approval

Date: June 10, 2013

To: Christine Browning, Principal Investigator
Joshua Goss, Student Investigator for dissertation

From: Amy Naugle, Ph.D., Chair

Re: HSIRB Project Number 13-05-27

This letter will serve as confirmation that your research project titled “Informal Inferential Reasoning’s Effect on the Learning of Formal Inference in Middle School Students” has been approved under the expedited category of review by the Human Subjects Institutional Review Board. The conditions and duration of this approval are specified in the Policies of Western Michigan University. You may now begin to implement the research as described in the application.

Please note: This research may only be conducted exactly in the form it was approved. You must seek specific board approval for any changes in this project (e.g., you must request a post approval change to enroll subjects beyond the number stated in your application under “Number of subjects you want to complete the study”). Failure to obtain approval for changes will result in a protocol deviation. In addition, if there are any unanticipated adverse reactions or unanticipated events associated with the conduct of this research, you should immediately suspend the project and contact the Chair of the HSIRB for consultation.

Reapproval of the project is required if it extends beyond the termination date stated below.

The Board wishes you success in the pursuit of your research goals.

Approval Termination: June 10, 2014