ABSTRACT

Title of dissertation: MATHEMATICAL SENSEMAKING
VIA EPISTEMIC GAMES

Mark Eichenlaub
Doctor of Philosophy, 2018

Dissertation directed by: Professor Edward F. Redish
Department of Physics

In this thesis, I study some aspects of how students learn to use math to make sense of physical phenomena. Solving physics problems usually requires dealing with algebraic expressions. That can take the form of reading equations you're given, manipulating them, or creating them. It's possible to use equations simply according to formal rules of algebra, but most students also learn to interpret the equations and use the equations as ways to bolster their physical understanding. Here, I report on three years of studying this mathematical sensemaking an introductory physics for life sciences course at the University of Maryland. There are both qualitative and quantitative threads to this work. The qualitative work analyzes a series of problem-solving interviews. First, I use case studies from these interviews to survey the variety of rich cognitive tools students bring to bear on problems around use of algebraic expressions and equations and make observations on potential applications to instruction. Next, I draw a connection between the ontological metaphors students use for equations and the epistemic games they play while solving problems. I show that certain ontological metaphors are used significantly more often in playing certain e-games, and describe the significance of this finding for problem solving. The quantitative thread of this thesis describes how my collaborators and I created and analyzed the Math Epistemic Games Survey, a math concept inventory that studies how students' uptake of problem-solving strategies such as "check the extreme cases" progressed over the year-long physics course. I show that students on average make little progress on the MEGS over a semester, which suggests that curriculum development in this area has great potential upside. Finally, I test several different methods of analyzing the multiple-choice test data that go beyond counting correct and incorrect answers to extract lessons from the distractors students choose. Using these methods on computer-simulated data and real data from the MEGS, I caution against drawing too-strong conclusions from their results.

MATHEMATICAL SENSEMAKING VIA EPISTEMIC GAMES

by

Mark Eichenlaub

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2018

Advisory Committee:
Professor Edward F. Redish, Chair/Advisor
Professor Andrew Elby
Professor Michelle Girvan
Professor Ayush Gupta
Professor Eric Brewe

Preface

I wrote Chapter 2 in collaboration with my advisor, Professor Redish. We co-developed the themes in this chapter by analyzing video data together. Additionally, Professor Redish proof-read this chapter, suggested edits, and rewrote certain sentences, especially relating to the background on epistemic games. This chapter has been accepted for publication in an upcoming book [Pospiech, in press], and is quoted here with at most minor changes. We intend to develop chapters 4, 5, and 3 into future publications.

Chapter 5 is based on a class project I conducted in a class taught by Professor Michelle Girvan, a member of my committee. Several important ideas there, including looking into modularity maximization algorithms on bipartite networks and using the variation of information as a metric of clustering accuracy, are due to her, and the chapter has been significantly improved by her feedback.

Portions of chapter 3, especially section 3.6.2, are adapted from blog posts I wrote on the blog "Reading Physics"[Liu and Eichenlaub, 2015].

Acknowledgments

I'm grateful that so many people have contributed ideas and support throughout the process of writing this thesis.

This work studies how students learn in hopes of supporting them as they grow as learners and scientists. Fortunately, I worked with so many insightful, engaged students as a researcher, teaching assistant, and instructor of PHYS131 and 132 that I could only address a small part of their contributions in this thesis. Thank you to all the students who participated in this research by sitting for interviews, taking the MEGS and MAX, talking to me about this work, and most of all, bringing their insights and perspectives to their physics courses.

My advisor, Joe Redish, introduced me to PER, created the research project under which I did this work, and consistently guided me along the way with advice, feedback, stories, and encouragement.

Deborah Hemingway collaborated closely with me throughout the thesis, it has been a pleasure to have her to brainstorm ideas with, read papers with, revise work together, and create new material.

The entire PER group at Maryland supported me throughout the writing of this thesis. Thank you for the conversations and analysis to Benjamin Dreyfus, Andrew Elby, Ayush Gupta, Brandon Johnson, Vijay Kaul, Christina Krist, Hannah Jardine, Kim Moore, Alice Olmstead, Gina Quan, Jennifer Radoff, Hannah Sabo, Stephen Secules, Katie Shirey, Erin Sohr, Elijah Tabachnik, and Chandra Turpen. You built a community where I could share ideas and grow as a researcher.

For advice and guidance on various stages of this work in while in progress, thank you to Eric Brewe, Michelle Girvan, Janet Walkoe, and Justyna Zwolak.

My thanks to the instructors, teaching assistants, learning assistants, and staff who ran physics 131 and 132 from 2015 - 2018. Via weekly meetings and informal conversations, you taught me a great deal about biophysics and teaching. I hope this work can give something back.

For her advice, feedback and support while writing this thesis, I would like to thank Jiajia Dong.

My family has always given me the deepest support. Thank you to my mother and father for inspiring me to explore the beauty of nature and of mathematics, and for inestimable amounts of support of all kinds over my life.

This work was supported by NSF Grant DUE-15-04366.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

131	first semester of UMD IPLS
132	second semester of UMD IPLS
ATPSA	Assessment of Textbook Problem-Solving Ability
e-game	epistemic game
ECD	Evidence-Centered Design
FCI	Force Concept Inventory
IPLS	Introductory Physics for the Life Sciences
IRT	Item Response Theory
LA	Learning assistant
MAMCR	Module Analysis for Multiple Choice Responses
MAX	Math Attitude and Expectations Survey
MEGS	Math Epistemic Games Survey
MIRT	Multi-Dimensional Item Response Theory
MPEX	Maryland Physics Expectations Survey
NEXUS	National Experiment in Undergraduate Science Education
NEXUS Physics	The NEXUS course on physics, i.e. PHYS131/132 at UMD
NSF	National Science Foundation
p-prim	phenomenological primitive
PERG	Physics Education Research Group (at UMD)
PHYS131	first semester of UMD IPLS
PHYS132	second semester of UMD IPLS
SCL	Scientific Community Labs
TA	Teaching Assistant
UMD	University of Maryland, College Park
Under/Over	Understanding and Overcoming Barriers to Using Math in Science

"We" generally refers to the research team of the author, Edward Redish, and Deborah Hemingway. In chapter 2 it refers to the author and Edward Redish. "I" refers to the author.

Chapter 1: Introduction

There are rules for how to manipulate mathematical equations, but using math in physics involves much more than learning these rules. As educators, we value helping and encouraging students to use math as a tool for better physical understanding. As they develop a sense of how to do this, students learn to associate meanings with variables and with operations. They manipulate equations to new forms with a goal in mind, check the if their answers are sensible, and use particular features of equations (for example their dimensions or their functional forms) to extract useful information without needing to use every aspect of the equation. In this thesis, I report on several threads of research related to understanding how students come to use and value mathematical sensemaking behaviors in introductory physics.

The context of this work is a large introductory physics for the life sciences (IPLS) course. Students in this population have a wide variety of attitudes towards math. They vary in their affect around mathematics, in their opinion of its usefulness and applicability to their personal and (prospective) professional lives, and in the epistemological resources they associate with mathematical problem solving. Although all the students have completed a year-long calculus sequence, they also vary considerably in their level of background in mathematics and the extent to which they've applied it to the sciences and used it for modeling. This diversity of viewpoints and backgrounds affords a rich opportunity for studying how instruction interacts with problem-solving.

We begin in chapter 2 by describing case studies from a series of problem-solving interviews with IPLS students. Building on the analytical framework of epistemic games to describe the way students frame a problem-solving activity, we find many seeds of expert-like thinking in how students approach problems.

The first time you try to ride a bicycle, you're likely to fall. Without encouragement, you might well abandon it and stick to walking. In these interviews, we saw that very often, students attempting expert-like sensemaking behaviors, perhaps as tentative explorations of whether they were useful and valid, failed to solve the problem they were working on. In class, students might attempt to examine extreme cases of a formula and thereby tie mathematical forms to physical intuition, but won't get an answer or receive positive feedback on their perhaps-experimental use of these techniques. We wonder whether current physics instruction misses an opportunity to reinforce these behaviors.

Much of our focus in this project is on student use of equations, and so it makes sense to study what types of objects equations are in students conceptual systems. We call this the ontological metaphors used to think about equations. For example, are equations mutable? Do they consist of individual pieces which one can reason about apart from the rest of the equation, or are the unitary black boxes? In chapter 3 we propose that students fluidly switch between these viewpoints in the way that seems most effective to them while solving problems, so that skill at changing your view of what equations are is another one of many pieces of expert problem-solving cognition, alongside content knowledge, metacognition, or self-efficacy. We develop a coding scheme to categorize ontological metaphors used to think about equations, and by coding and analyzing student use of equations during problem solving, we examine the link between equation ontologies and the epistemic games students are playing. We find that playing certain epistemic games (extreme cases, dimensional analysis) is associated with certain ontological metaphors (symbols as parameters and grouping of symbols, respectively), so that having and accessing these ontological metaphors for thinking about equations may be a prerequisite to using these productive problem-solving strategies, which we refer to as epistmeic games (e-games).

To investigate student uptake of e-games quantitatively, we built, tested, and validated a new concept inventory, the Math Epistemic Games Survey (MEGS), described in chapter 4. The MEGS consists of questions best-solved using common problem-solving strategies such as dimensional analysis or estimation. We wanted to know whether, after a year of instruction in a reformed-pedagogy course, students had gained control of these problem-solving strategies as part of a personal toolbox which they used freely on challenging problems. The IPLS course was a good target because its fundamental goal is to teach students skills that fall broadly under "thinking like a physicist" - skills related to building, understanding, and evaluating quantitative models - which in recent years have been increasingly valued in the biological sciences and medicine and health.

After giving the test to IPLS students before and after the semester over the course of three years, we found that even in reformed-pedagogy, active-learning based classrooms which show good improvement on concept inventories such at the Force and Motion Concept Evaluation, students often make little to no progress on the MEGS. However, after the targeted efforts of an instructor based on the specific strategies built into the MEGS, the class made modest gains, which we hope will be replicated and significantly expanded in the future, both by classes at UMD and elsewhere.

MEGS questions were built around four targeted e-games: dimensional analysis and scaling, estimation, examining extreme cases, and mapping symbols onto meaning. After building the survey around those ideas, we wanted to know to what extent is the survey a valid, reliable instrument to measure constructs corresponding to student use and facility with each of these e-games. Classical test theory provides one approach to this question, and I present the results of classical test theory on MEGS data in chapter 4.

Factor analysis provides another approach to validating tests and gleaning new insights from them. Factor analysis traditionally tries to find questions on a test that are correlated to each other. In this context, two questions being "correlated" means that if a student answers one question correctly, they are likely to answer the other question correctly as well. This approach doesn't gain anything from the data on which incorrect answer choices students chose, so following a recently-published extension of factor analysis [Scott and Schumayer, 2017], I apply factor analysis to all the answer choices on the MEGS, not just the questions, in chapter 5.

Recently, network-based approaches have allowed understanding the role of distractors on the FCI [Brewe et al., 2016]. This approach models test responses as a network with both students and answer choices as nodes. In chapter 5, I use this model to explore the idea of clustering MEGS answer choices together. I compare the results to those from factor analysis in an attempt to both learn new things about the MEGS and what it can tell us about how students perceive the questions on it.

Also in chapter 5, I use factor analysis and network techniques on simulations of test data that come from four different models I've built. This explores what I do and do not learn from quantitative analysis of MEGS data. I try to answer the question, "for which models of student cognition does our network-based technique give greater insight into the structure of a test and how students are responding to it than factor analysis does"? My conclusions in this chapter are mostly preliminary; they mostly demonstrate the caution I need to take in interpreting the results of statistical analyses. Factor analysis and network-based techniques have a number of parameters and choices to make during the analysis process; they need to be fine-tuned before they give sensible results. The methods I apply to MEGS data in chapter 5, while giving some genuine results, don't tell me anything I didn't already know.

Before formally beginning graduate school in PER, I wrote an essay [Eichenlaub, 2013] on my views on the learning process, cognition, and physics. I've taken the final chapter, chapter 6, as an opportunity to update this essay with a personal reflection on things I've learned over the course of writing this thesis and becoming part of the physics education research community.

Chapter 2: Blending physical knowledge with mathematical form in physics problem solving

2.1 Introduction

Physicists and educators have long held problem-solving to be one of the key tools to help students understand physics [Meltzer and Otero, 2015]. If problem-solving is a bridge to expert-like understanding, we should find ways to let students experience expert-like thinking in as many dimensions as possible while working problems. This includes learning new physical concepts and mathematical techniques, because experts and novices differ greatly in the amount of physics and math they know. But experts also diverge from novices in their problem-solving strategies, their patterns of metacognition [Schoenfeld and Sloane, 2016], their epistemological stances towards their work (and abilities to negotiate between various stances), their conception of what mathematical entities are, and their expectations for how to derive meaning from their work. These differences between experts and novices are part of a "hidden curriculum" that students need to learn as they progress in physics, but which we rarely teach explicitly [Redish et al., 2010]. In this chapter, we document a variety of ways that, in a problem-solving interview setting, we saw college students in an introductory physics class trying out expert-like problem-solving strategies and epistemological stances. We also point out that in their early attempts at using these expert-like tools, students were often unsuccessful in solving problems, and suggest that educators look for ways to encourage students' early attempts at using expert-like tools so that students can continue to develop their facility with them, even if their first uses are incorrect or incomplete.

In particular, researchers have singled out math as a particular sticking point in problem solving in introductory physics. Much of the existing research seeks to document student understanding, or misunderstanding, of particular mathematical tools, such as differentiation or coordinate systems. Our teaching experience shows that even when students appear to have mastered the appropriate tools in previous classes, they may still struggle to use those tools effectively in physics problems. In previous work, one of us [Redish and Kuo, 2015] laid out an argument that this is largely because the ways that physicists make meaning with mathematics are unfamiliar to students. Even if they are skilled with the manipulations of algebra and calculus, students' expectations about how to interpret variables may lead them astray. For example, many students, given a problem about test charges and electric fields, will say that changing the magnitude of a test charge changes the magnitude of the electric field it measures. They reason from the equation $E = F/q$ that if q increases, E decreases. The students understand the math involved well, but don't account for the way the force on a charge changes with the charge - there was a hidden functional dependence they did not see because physics culture assumes the reader will associate every symbol (in this case, F) to its physical meaning. That would make the functional dependence of F on q clear, but students don't yet expect to have to find this physical meaning when solving problems. The challenge for educators is to create problems and problem-solving environments that encourage students to search for physical meaning in mathematics.

In creating problems, educators often separate "qualitative" problems that test and build intuition from "quantitative" problems to develop mathematical skills [Hsu et al., 2004], indicating an implicit assumption that these are separate faculties that are used and developed individually. We believe that for experts, intuition and mathematics are not insulated from each other, or even cleanly separable. Instead, they reinforce each other; intuition is often connected to mathematics and mathematics is understood partially via intuition. While solving a problem, an expert will blend mathematical forms such as equations (or abstracted properties of equations), with intuitive conceptual schema to create richer mental spaces than those derived from formal mathematics alone.

For an example of this blending of intuition and mathematical form, we look at Sherin [2001]'s description of "symbolic forms", a class of blended intuitive-formal conceptual structures that experts (and in Sherin's case, second-year physics students) use to understand equations. To introduce symbolic forms, we'll take an example from Sherin, who describes two students thinking about a ball falling through the atmosphere at terminal velocity. The students intuitively understand that air drag and gravity are both acting on the ball, but balance each other out, leaving no net acceleration. In Sherin's account, the students activate a conceptual schema for "balancing" of competing influences. This balancing schema could potentially match many different physical scenarios, or even everyday scenarios, such as expenses balancing out income when breaking even financially, but here is it called to understand air drag and gravity. The students then associate the balancing schema with the abstracted symbol template for equations, $\square = \square$, where each square represents one of the two balancing influences. The students know that they are looking for an equation with an expression related to gravity on one side and an expression related to air drag on the other. The students' work on a specific equation is then informed by this pairing of the intuition behind balancing with the symbolic template. The combined intuition and formal structure are collectively a symbolic form. Sherin identified 21 symbolic forms in his data corpus; our purpose here is to use them as one example of blended intuitive and formal thinking that is found in experts and potentially in students as well.

Symbolic forms are not a complete account of how physicists make meaning with equations. The example of failed meaning-making in the equation $E = F/q$, cited earlier, involves the correct use of the symbolic form Sherin identified as "prop-", where a schema related to "if one goes up, the other goes down" is blended with the symbol template $\left[\frac{\cdots}{\ldots x \ldots}\right]$, but this symbolic form alone wasn't enough to lead students to the right answer.

Based on an exploratory analysis of problem-solving interviews, we suggest that students, in the right circumstances, use a large and diverse arsenal of productive, sophisticated, and creative ways to conceptualize physics problem-solving. They do not always access these resources when they would be productive, and many of the difficulties students experience with using math in physics are not so much difficulties of having the appropriate tools, but of applying them appropriately. While much of the hidden curriculum will need to be learned via years of enculturation in the physics community, there are entire swaths of it that don't need to be explicitly taught so much as activated. Small interventions that encourage students to use specific problem-solving strategies, can, in some cases, greatly enhance students' access to productive ways of thinking about mathematical tools that are rarely explicitly taught.

The strategies we're investigating are commonplace, well-known to physicists, and generally well-regarded components of effective problem-solving. They include examining special and extreme cases, dimensional analysis, and estimation. Our contribution to understanding these strategies is to suggest that their scope can be very broad. They can be used at different stages of problem-solving and in different ways. We also give examples of how students use these strategies construct meaning from mathematical expressions in ways similar to how experts do it.

2.2 Theoretical Framework: Resources, Framing, and Epistemic Games

Our analysis is situated in the resource model [Hammer, 2000, Redish, 2004]. In this framework, students don't have monolithic conceptual understandings; they have many small pieces of knowledge, or resources, that they can call on while solving a problem. When solving a problem, students will activate various resources and construct a solution based on them. If students don't solve a problem correctly, it may be that they don't have the appropriate resources, or that they do, but aren't activating them in that context. In the previous example of a test charge and the measured electric field, students did activate resources relating to understanding inverse mathematical relationships (including the prop- symbolic form), but did not activate resources related to the functional dependence of force. Whether or not students activate a resource can depend on how they associate it with other resources they are using, so in a future problem, students might improve their performance if they've learned to activate resources related to functional dependence when they see questions about forces in electromagnetism.

The issue is not so simple, though. The students in question were all able to recite the mantra "the electric field is independent of the test charge". In this sense, they knew the answer to the problem, but they didn't call on this knowledge, or if they did, didn't apply it. In addition to resources related to manipulating mathematical equations and resources related to intuitive understanding of physics, students also have "epistemological resources", resources related to how they seek to obtain and justify knowledge [Hammer and Elby, 2003].

A student who uses an equation because it makes intuitive sense may come to the same answer as a student who uses an equation they found in a textbook they consider authoritative, but the way they are thinking about knowledge is very different; they are using different epistemological resources. The students who answer the test charge problem incorrectly are probably not activating epistemological resources related to interpreting each variable physically, or resources related to finding concordance between memorized facts (such as the electric field being independent of the charge) and the results of reasoning based on equations.

To understand why students sometimes use one set of epistemological resources and sometimes another, we use the lens of epistemological framing [Bing and Redish, 2009b]. Because we could potentially use any resource at our disposal (i.e. every fact, technique, or type of reasoning we can conceive of) on a given problem, the space of problem-solving strategies we have to search through to find one effective approach is extremely large. We begin by narrowing the problem down to a certain type of problem, and then search through the resources we associate with that type. Calling on a physical principle to solve a problem requires activating different epistemological resources than using an equation does, and those resources often are associated with different epistemological framing [Gupta and Elby, 2011, Kuo et al., 2013]. Students who answered that changing the magnitude of a test charge changes the magnitude of the measured electric field may have entered a "calculation" frame, and didn't remember or pay attention to their knowledge that the electric field is independent of the test charge because they didn't frame the task as one in which physical principles are relevant.

Moving towards expertise in problem solving is as much about using what resources you have effectively as it is about picking up new resources. As students work physics problems, they need to learn not only new content, but new ways of relating to the content. They need to be able to effectively frame epistemologically and activate appropriate resources. All of these are difficult tasks that live mostly in the hidden curriculum.

Analyses of problem solving often break the task down into a series of steps.

Sometimes this is prescriptive, as when textbooks list a series of steps to make in solving a problem. For example, Redish et al. [2010] describes a textbook with the following scaffold for problem solving

> Model! - Make simplifying assumptions.
>
> Visualize! - Draw a pictorial representation.
>
> Solve! - Do the math.
>
> Assess! - Check your result has the correct units, is reasonable, and answers the question

and gives an example where the method failed. The textbook posed a question asking us to find the volume occupied by the water evaporated after sweating during exercise. The solution manual followed each step, finding that the volume was simply the volume of an ideal gas with the appropriate number of molecules, ignoring that the evaporated water will, by convection and diffusion, spread out over a very large volume. The textbook's solution manual follows each individual step, but nonetheless comes to a nonsensical answer to a problem by failing to "tell the story of the problem". From this example, Redish finds

> Tying the analysis to a rubric a formal set of mapped rules . . . does not help if it does not also activate an intuitive sense of meaning by tying the problem to all we know and recognize about a system

We also view problem-solving as a series of steps, but not as steps for students to follow, but as a framework for researchers to understand how students solve problems. This approach is common in physics education research. For example, in analyzing student difficulties using math in physics, Wilcox et al. [2013] proposed the ACER framework, which consists of Activation of the tool, Construction of the model, Execution of the mathematics, and Reflection on the result.

Whereas a prescriptive problem-solving script tells students to follow precise steps in a given order, Wilcox et. al. write, "...we are not suggesting that all physics problems are solved in some clearly organized fashion, but a well articulated, complete solution involves all components of the ACER framework." That is, having the framework allows the researchers to narrow their focus and identify specific tasks students are struggling with, rather than simply bemoaning that they can't apply math appropriately. In that paper, Wilcox et. al. found that students' resources for the technique of taking a Taylor expansion weren't activated by the appropriate signal, which was one variable of interest being very much smaller than another, and suggested that problems be written to focus on building this particular association for students between signal and mathematical technique.

Frameworks like ACER are effective at picking out specific technical steps that students don't take in problem-solving. Our interest here is broader, including student epistemologies, attitudes towards mathematics, conceptualization of the entities involved, and other aspects of the hidden curriculum. The framework of epistemic games is a flexible one that allows analysis of both problem-solving moves and the motivations behind them.

We have previously discussed epistemological frames in problem-solving. Framing is a general feature in psychology, and when we work in a particular frame it often cues a script for how that type of activity typically goes, which sets expectations for what will happen next and what sorts of actions are appropriate [Goffman, 1974].

An epistemic game is a script (with additional structure to be described below) that allows us to understand the moves students make in problem solving [Tuminaro and Redish, 2007]. As we watch students solving problems, we assign their problem-solving to some particular epistemic game, which we take to structure the types of resources they call on and the order in which they use them. An epistemic game will generally have a particular epistemological frame associated with it, but adds additional structure. The viability of epistemic games as an analysis framework stems from its psychological plausibility via the connection to psychological scripts and that, when Tuminaro and Redish [2007] analyzed student problem solving, they found that certain epistemic games were repeated many times on different problems and in different circumstances. The term "epistemic game" comes from Collins and Ferguson [1993], although the version we use here is that of Tuminaro and Redish [2007].

In an epistemic game, as in games like solitaire or chess, one or several players make moves. These moves might be mathematical moves, such as *add the same quantity to both sides of the equation*, conversational moves, such as *offer a reason supporting your position*, or physical moves, such as *draw a picture of the situation.* Because players can make various types of moves, analyzing the moves lets us focus on different aspects of the hidden curriculum in problem-solving.

As the players of an epistemic game make moves, they gradually fill out an epistemological form, a template for what the solution to the problem should look like, which may be physical or verbal. Finally, players either reach the e-game's stopping condition and decide they are done, or else switch to a different game or give up on the solution attempt.

Tuminaro and Redish identified six common games that students play during problem solving, such as *recursive plug-and-chug*, in which students identify a formula and put values into it without interpreting the results, and *mapping meaning to mathematics*, which describes the problem solving process in which students analyze the physics of a situation, turn their analysis into equations, manipulate the equations, and then turn the result into a new physical understanding.

Students use e-games to guide their inquiry, and their (generally unconscious) choices for what e-game to play have large effects on their problem-solving process. Different games have different rules about what sort of evidence is salient, what sort of moves are allowed, what type of arguments to give, and what it means to be done with a problem. When students get stuck on a problem or come to answers that don't make sense from the viewpoint of experts, they often have resources that would allow them to solve the problem, but never access them because they are not included in the current frame [Tuminaro and Redish, 2007, Bing and Redish, 2012].

We do not consider playing an epistemic game favorable or unfavorable; that depends on which epistemic game and how appropriate it is to the situation. Epistemic games also aren't confined to students; experts play them as well, and do it very effectively. For example, in his short paper "A Model of Leptons" Weinberg [1967], searches for an equation to describe leptons and their interactions. The method is to list various properties the equation should have—what symmetries it has, what types of solutions to avoid, etc. Each such consideration can be translated into a particular feature that the final equation should have, and by combining a sufficient number of features, only one equation is left that satisfies them all—the final equation derived for leptons and their interactions. Weinberg is playing an epistemic game we call "significant features". This is a game used to generate solutions to a given problem (as opposed to evaluating a proposed solution). To play, one lists relevant significant features a solution ought to have, such as a maximum at a certain place, or matching a certain symbolic form. Each feature is translated into a formal constraint or piece of the sought solution, such as the derivative being zero at the maximum or a symbolic template which matches the symbolic form appearing in the equation. As the player discovers more features and their associated forms, they gradually fill out the equation (or plot or other form) they are seeking. The game ends when they either decide they have completely specified the answer to the problem or decide that they don't know enough features to do so.

In Sherin [2001]'s work, two students decide that under constant acceleration the equation for velocity as a function of time is either $v(t) = v_0 + at$ or $v(t) = v_0 + \frac{1}{2}at^2$, but cannot decide between the two. Sherin analyzes this as using the "base plus change" symbolic form. Students conceptualize the situation as velocity starting at some given value, then changing to a new value, and realize that this maps onto the symbolic template $\square + \Delta$. The symbolic form doesn't distinguish between the terms at and $\frac{1}{2}at^2$ as "changes" to map onto Δ in the symbolic template. Both are positive (for positive acceleration and time) and indicate an object speeding up.

Sherin's analysis is that using only a symbolic form isn't enough for students to determine the correct equation. We agree, and add that the students are playing the same "significant features" game that Weinberg did in building a model of leptons. They begin with a feature they want to the solution to have - matching the conceptual schema of base + change, and translate that into a mathematical form - the $\square + \Delta$ symbol template. Although they ran out of features to finish constraining their answer to the one correct answer, they were nonetheless playing the same epistemic game, just with very different material and at different levels of expertise.

2.3 Data and Analysis

The students we interviewed were enrolled in an introductory physics for life science course at the University of Maryland. Most are juniors, with some sophomores and seniors. The course prerequisites include one semester each of calculus, probability, chemistry, and two semesters of biology. Students are mixed between having taken physics in high school and not.

This is a population of relative novices in physics, but who have taken from 5 - 12 college science courses before taking this one; they generally have strong expectations about how science courses and problem-solving in them work, which the instructor (Redish) routinely challenges.(See [Redish et al., 2014] for more details on the creation and principles behind the course.) All interviews used a think-aloud protocol, encouraging students to write and articulate their thoughts at all times as they solved problems. Some interviews were one-on-one with the interviewer (Eichenlaub) and other were group interviews in which the interviewer was present but participated minimally, with occasional small interventions designed to prompt use of specific problem-solving strategies. We conducted a total of 24 hour-long interviews with 23 different students enrolled in the first of two semesters of this course.

With these interviews, we were interested in the breadth of approaches and conceptualizations students take in problem solving, including whether and how they blend physical intuition with mathematical formalism and how they conceive of variables, parameters, and entire equations. We chose problems and problem-solving strategies that we hoped would elicit epistemic games with a strong interplay of intuition and formalism in hopes of bringing out a diversity of interesting conceptual systems in students' solution attempts. The strategies we investigated were *examine extreme or special cases*, *dimensional analysis*, and *estimation*, chosen especially because they are all familiar parts of an expert physicists' toolkit, but are not always taught explicitly at the introductory level. Please see appendix A for the problems given in interviews.

We wanted to make fine-grained analysis of small, interesting incidents in our interviews, so we took video of the interviews ensuring that the field of view captured all students (for group interviews) or student and interviewer (for one-on-one interviews) so that we could reference speech, gesture, and other expressions. Students wrote on a whiteboard, which we photographed at the interview's conclusion.

Our goal in analyzing these interviews was to generate hypotheses about cognitively-rich ways that students can interact with math and physics. This was exploratory analysis, not confirmatory, so the results we present here are case studies to be examined in more detail in the future. Our focus was on finding particularly interesting moments throughout the problem-solving sessions, including moments of blended mathematical/intuitive sensemaking and moments that show how students conceive of the mathematical entities they're working with. To that end, we reviewed the videos highlighting incidents that stood out to us, then discussed them together to generate hypotheses regarding student conceptualizations that interested us. Here we present those hypotheses along with descriptions of the incidents that we watched while generating them.

Below, we describe each strategy and report briefly on how students in our interviews took up the strategy before discussing, through the lens of epistemic games, specific cognitive aspects of problem-solving that these strategies elicited.

2.3.1 Extreme and Special Cases

Most physical systems we examine in problem solving have one or more free parameters that enter the problem. For example, in trying find the effective spring constant of two springs connected in series to form a single combined spring, the individual spring constants are such parameters. If we set one of these parameters to its largest or smallest possible value, we're looking at an extreme case. So for springs in series, we could set the second spring constant to be infinite, in which case it is completely rigid, does not contribute at all to the stretching of the combined spring, and the effective spring constant would simply be that of the other spring. Using this fact to try to understand something about the general situation is a strategy we call "extreme case" reasoning. We might also consider the case where the two spring constants are equal. Then each spring stretches the same amount, the total stretch is twice as much as the stretch of an individual spring, and the effective spring constant is half that of an individual spring. We call this "special case" reasoning. The two are almost the same, but extreme cases have been discussed independently in the literature, so we identify them as separate but closely-related reasoning strategies.

Clement and Stephens [2009] studied extreme cases in a grade school setting, finding that looking at the extreme case helps students build vivid, dynamic mental imagery, consistently leading to better intuitive understanding of physics scenarios. Used in quantitative problem solving, extreme cases not only boost our intuition, but also allow us to connect that intuition to equations we've generated or are considering. Our accuracy and intuition for thinking about extreme cases has led physicists to make their study a standard problem-solving tool [Morin, 2008]. Nearing et al. [2003] elaborated on why extreme cases lead to better intuition in his undergraduate textbook on mathematical physics

> How do you learn intuition?
>
> When you've finished a problem and your answer agrees with the back of the book or with your friends or even a teacher, you're not done. The way to get an intuitive understanding of the mathematics and of the physics

> is to analyze your solution thoroughly. Does it make sense? There are almost always several parameters that enter the problem, so what happens to your solution when you push these parameters to their limits? In a mechanics problem, what if one mass is much larger than another? Does your solution do the right thing? In electromagnetism, if you make a couple of parameters equal to each other does it reduce everything to a simple, special case? When you're doing a surface integral should the answer be positive or negative and does your answer agree?
>
> When you address these questions to every problem you ever solve, you do several things. First, you'll find your own mistakes before someone else does. Second, you acquire an intuition about how the equations ought to behave and how the world that they describe ought to behave. Third, It makes all your later efforts easier because you will then have some clue about why the equations work the way they do. It reifies the algebra.

Extreme cases, to Nearing, are not about the physics situation alone or the mathematical expression alone, but a way of bridging the two into a unified qualitative and quantitative understanding of physics.

In a prototypical use of the extreme or special case reasoning, students first derive an expression, in terms of parameters of the problem, that is a potential solution to the problem. For example, they might find the acceleration of a block in terms of various masses, angles, and coefficients of friction involved. They then use their physical intuition for extreme cases to evaluate this potential solution.

This evaluative use can be analyzed as a “sanity check” epistemic game. This game begins after students generate a candidate solution to a problem, and is used to test whether the solution makes sense. The prototypical moves of the game are

1. Identify a feature which the candidate solution intuitively ought to have.

2. Check whether the candidate solution has this feature.

3. If it does, identify a new feature the solution ought to have. If it does not, either reject the solution and start over, or enter a new epistemic game to determine why the solution and feature do not match.

4. Continue playing the game until you can’t think of any more features or are satisfied with your confidence in the candidate solution.

When playing the sanity check game with the extreme case strategy, these moves could look like this:

1. Identify a physical variable in the problem.

2. Imagine it becoming extremely large, extremely small, or some special value that stands out.

3. Intuitively identify the behavior of the system in this case.

4. Analyze the same limit of expression in the potential solution.

5. Compare the results of (3) and (4) for consistency. If they are consistent, confidence in the solution increases. If they are inconsistent, choose a new e-game to figure out whether it is your intuition or the mathematical expression that is incorrect.

6. Repeat for other variables in the problem.

This game encourages students to repeatedly compare a mathematical expression with a physical intuition, and so promises to be a good place to learn about how students use math to inform physical understanding and vice versa.

Although we've outlined a canonical version of the game above, physicists use extreme cases in many other ways. The snippets from physicists discussing the relation between the Yukawa and Coulomb potentials in section 1 discuss sending a parameter (α) to an extreme (zero), but instead of examining the physical behavior of a system in this limit, they discuss an equation itself simplifying to a different equation.

Further, in many cases beyond the introductory classroom, we can only find analytic solutions for the limiting cases of an equation, so studying the asymptotic behavior of otherwise intractable physical systems has become the most common analytical approach in modern mathematical physics [Bender and Orszag, 1999]. As a result, extreme cases and special cases lead to a host of useful tools, resources, and intuitions for physicists, including for example perturbation theory and the WKB method. The power of this game is one of the reasons that the predilection of introductory students to "put numbers in right away" (thereby reducing the problem

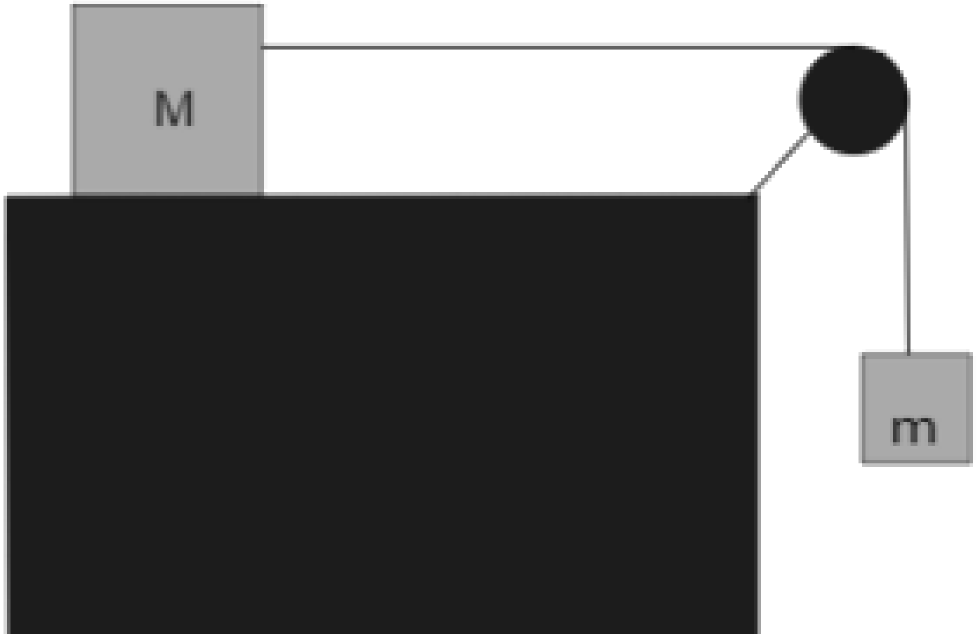
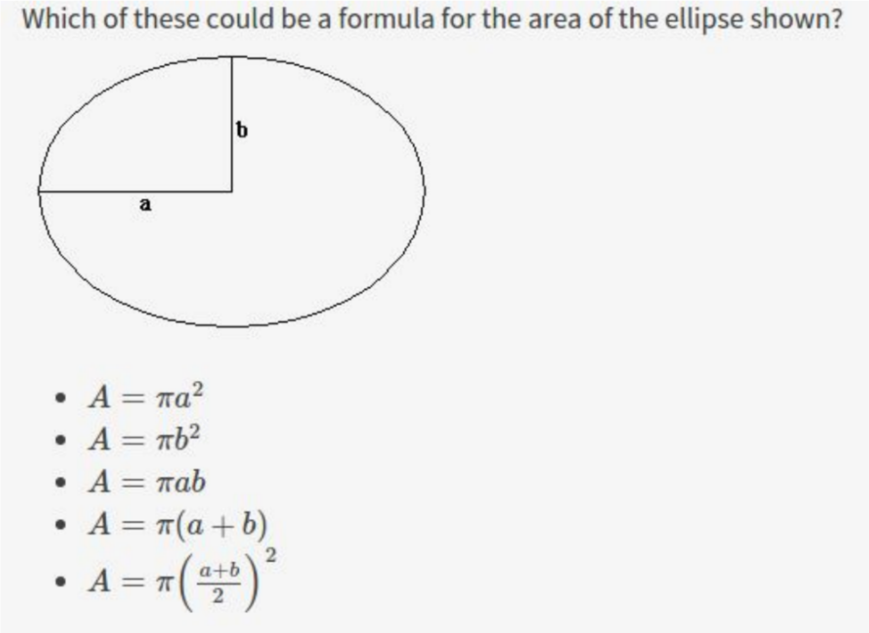


Figure 2.1: The half-Atwood problem: A block of mass M is attached to a block of mass m via a massless string strung over a pulley as shown. The setup is frictionless. What is the acceleration of the block m?

to one that looks more like "just math") is often counter-productive.

In interviews, we gave students several problems where we expected the extreme cases game to be useful: the half-Atwood machine (Figure 2.1), the electric field on the axis of a ring of charge, springs in series and parallel, and the area of an ellipse (all reproduced in appendix A).

In every case, we found that students have strong and accurate physical intuitions for the extreme or special cases. In some circumstances, students consistently spontaneously play the sanity check game using special case reasoning. For example, every student interviewed on the ellipse problem (Figure 2.2) considered the special case $a = b$, a circle, and used it to evaluate the given answers. No students,



Which of these could be a formula for the area of the ellipse shown?

- $A = \pi a^2$
- $A = \pi b^2$
- $A = \pi ab$
- $A = \pi(a + b)$
- $A = \pi\left(\frac{a+b}{2}\right)^2$

Figure 2.2: The ellipse problem

on the other hand, spontaneously checked the extreme case $b \to 0$, however, when prompted by the interviewer to consider "a long, skinny ellipse", most did use this extreme case to answer the question correctly.

Extreme/special case reasoning also proved consistently valuable to students answering the half-Atwood problem (Figure 2.1) and to students finding the electric field on the axis of a ring of charge.

The students in our interviews found this strategy less effective when asked to determine the effective spring constant of two springs connected in series. Asked to consider this problem without being prompted to think of extreme cases, Lizzie, Myra, and Lelia (pseudonyms) had the following discussion:

1. Lelia: What's Hooke's law again? Oh yeah, T is this. [writes an equation for Hooke's law] So in this. The length would technically be twice as long.
2. Lizzie: oh for the two
3. Lelia: technically this k coefficient would be twice as long as one of them.
4. Lizzie: yeah [erases board and writes $T = k\Delta L$]
5. Lelia: so I think k-series would be them added together. Cause I remember I remember from
6. Lizzie: the homework
7. Lelia: yeah there's two connected the new k coefficient is twice as much, I think.

8. Lizzie: we have two k's. [all writing equations involving k, T, and ΔL]
9. Lizzie: k-series would be k-one plus k-two
10. Lelia: yeah, that's what I'm thinking

Lizzie, Lelia, and Myra (did not speak above) associate higher spring constants with more length of the spring, leading them to conclude that springs in series have an a spring constant that adds. After working on other problems for twenty minutes, they returned to the springs, and the interviewer asked what would happen if one spring were much stiffer than the other

1. Lizzie: the stretch, the easy one would stretch a lot
2. Lelia: and the hard one would stretch a little bit, so the total stretch would be mostly due to the softer spring. so i mean again I guess k-constant would be the softer one.
3. Lizzie: but the hard one would still contribute a little bit
4. Lelia: yeah, but we don't know. I don't know how much, you know what percentage
5. Myra: can we like divide it by the number of springs?
6. Lelia: like k-one plus k-two divided by two or something?
7. Lizzie: or n?
8. Myra: cause I'm thinking because if one is way easier to stretch and the other one is not stretching at all, but each spring is still

> contributing some stretching, so then you divide it by the number of springs.

Their physical intuition is correct, but in the remaining time, they are unable to match their intuition to an equation, and ultimately revert to their original answer of $k_{eff} = k_1 + k_2$. Although their effort to play extreme cases didn't result in a correct equation, they did make correct conclusions about the mathematical form of the answer, specifically that the effective constant should be (very nearly) the same as that of the softer spring, and they consistently attempted to match physical intuition to equations. However, without a clear mapping from spring constants onto physical stiffness, it was difficult for them to find a correct equation.

2.3.2 The Dimensional Analysis Game

There are several strategies based on the idea that if two physical quantities are equal, they must have the same dimensions. We refer to these strategies collectively as "dimensional analysis", and they are taught extensively at the introductory level [Robinett, 2015], while also remaining of professional interest to physicists for more than a century [Bridgman, 1922]. A prototypical example of playing the sanity check epistemic game for evaluating a formula using dimensional analysis would be

1. Find an equation that may be a solution to a given problem.

2. Evaluate the physical dimensions of each term on the left side of the equation.

3. Multiply the dimensions of all terms on the left hand side together to get the dimensions of the entire left hand side.

4. Repeat (2) and (3) for the right hand side.

5. Compare the dimensions of each side of the equation. If they are the same, the equation may be correct. If they are not, the equation is incorrect.

This game allows students to catch some mistakes in their answers. Students in our sample played dimensional analysis readily on questions that specifically asked about dimensions, for example asking which of a set of four formulas could be the surface area of an object, but also occasionally used it productively in questions aimed understanding functional relationships. For example, when asked,

> Sixteen students are sharing N large cheese pizzas. Assuming that the students share the pizza evenly, which expression gives the number of students each pizza must feed?

many students had difficulty choosing between the expressions $N/16$ and $16/N$, among other distractors. Two interviewees noted that the number 16 had units of students, and because the answer they were looking for had units of students, the choice must be $16/N$

Our data set was not set up to investigate the more elaborate dimensional analysis game in which students are asked to use the dimensions of relevant variables to explicitly construct formulas, or pieces of formulas, in cases where the full analytical derivation is too long, complicated, or intractable to be useful [Robinett, 2015], although we believe this game would be interesting to research in the future. Constructing a formula from elemental pieces, as well as understanding an incomplete formula which contains scaling information but cannot be numerically evaluated, may lead to rich student cognition.

2.3.3 Estimation

By estimation, we mean integrating personal knowledge, a corpus of memorized numbers, and approximation heuristics to obtain order-of-magnitude estimates of interesting quantities, either in physics or in everyday scenarios. Like dimensional analysis and examining extreme cases, estimation is a highly valued in the physics community and in physics education, which have a culture of "Fermi estimates", "back of the envelope" calculations, and "order of magnitude" estimates. For example, *The Physics Teacher* publishes a "Fermi Question" in each issue [Weinstein, 2018], and several universities have undergraduate courses in estimation [Phinney, Chiang].

We chose to investigate estimation because performing estimates generally requires students to think about their everyday experience and find methods of quantifying it, often while building equations that multiply various such terms together. Thus, it forces students to use intuition and a formal understanding of mathematics simultaneously.

A case study by Modir et al. [2014] established an estimation epistemic game involving six moves,

1. Problematize

2. Propose method

3. What to remember

4. See if parts are enough

5. Pure Calculations

6. Evaluation

and documented how a student estimated the energy in a hurricane by going rapidly forward and backward between these moves.

In one of our interviews, a group of four students, Amelia, Zane, Jean, and Chris, attempt to estimate the time it would take a submersible submarine to sink to the bottom of the ocean. The group agreed to assume the ocean was 1000m deep, and Jean calculated a descent time of about fourteen seconds by assuming the sphere fell with ordinary gravitational acceleration. Several group members challenged the notion that the submersible would accelerate during its descent and proposed it would instead fall at terminal velocity, but never reached consensus before the following exchange

1. Amelia: Well if you think about it based on the previous situation that we said, we said it was at a thousand meters (Jean: mmhmm) the force was two thousand newtons. Fourteen seconds technically could be legible just because a thousand meters isn't really a lot. We have a really heavy (Zane: that's true) like submersible, so it kind of makes sense in that situation.

2. Zane: let's go with it

3. Jean: go with the...
4. Zane: fourteen seconds, yeah
5. Amelia: It all depends on like, all these variables. With these variables it would make sense that it would be dropping that fast.
6. Jean: And we're assuming there's no um, buoyant force, no viscous force

Although Zane called on counterintuitions several minutes before this exchange ("it's not going to hit, you know, a hundred thousand miles per hour at the bottom."), and repeatedly argued against the constant acceleration approach, the group decided that their calculation "kind of makes sense", ultimately accepting a highly unreasonable answer. Despite their incorrect conclusion, we see in this passage group members calling on a sense of whether numbers are reasonable for a given physical situation, questioning the relation between unknown parameters and quantities of interest, and examining the simplifying physical assumptions that go into their reasoning. At the conclusion of the interview, the interviewer mentioned that their conclusion had the submersible reaching the ocean floor at roughly 300 miles per hour, and the group burst out laughing. It may be that the group's considerable efforts at sense-making failed largely due to an unfamiliarity with the relevant units, as well as neglecting to convert them into more everyday terms.

In this incident, we see a group negotiating what physical effects to model mathematically and what to ignore. This skill is essential to all physical modeling. For example, in introductory physics we often model the flight of a thrown ball using only a uniform gravitational force, giving a parabolic trajectory. In doing so, we ignore aerodynamic drag, other aerodynamic effects (e.g. lift), nonuniformity of the gravitational field, inertial forces due to Earth's rotation, magnetization of the ball in Earth's magnetic field, the Yarkovsky effect (black-body radiation is red/blue shifted in the ball's reference frame due to its rotation, cause a net torque), momentum imparted by sunlight the ball absorbs or reflects, transfer of material in and out of the ball's surface, and many other effects. Some of these can be important or not for a ball, depending on the accuracy we want and the parameters of the situation. Others are effectively never important for a ball thrown on Earth, but are relevant for, e.g. dust particles in space. Physicists often estimate the sizes of such effects to see whether they belong in more complete and explicit model. By improving student estimation skills, we also empower them to build better-informed mathematical models, and to understand the extent of those models' applicability.

2.4 Blending and Sensemaking

In most frameworks to analyze student use of mathematics, there is a step in which the student manipulates the equations. For example, in ACER, this step is Execution of the mathematics, described as

> Transforming the math structures (e.g., unevaluated integrals) in the construction component into relevant mathematical expressions (e.g., evaluated integrals) is often necessary to uncover solutions. Each mathematical tool requires a specific set of steps and basic knowledge. For

> example, executing a Taylor approximation may require knowledge of common expansion templates (e.g., $\sin x \approx x + x^3/3! + \ldots$) and how to adapt these templates to the mathematical model developed previously. Alternatively, one might need to know how to compute derivatives of complex functions. The mathematical procedures performed in this component are not, at least to experts, context free. In addition to employing base mathematical skills, experts maintain awareness of the meaning of each symbol in the expression (e.g., which symbols are constants when taking derivatives).

Although this description indicates that the operations are not purely formal, and that the problem-solver needs to remember the context and meaning of the symbols, the steps on which we understand the equations' emergent meaning and match them to physical understanding are separate steps from the steps of symbolic manipulation under these frameworks.

Research on the manipulation step has mostly focused on the difficulties that students have in making manipulations or on the procedural resources they use while manipulating equations (for example, thinking of physically sliding a variable from the numerator of one side of an equation to the denominator of another)[Wittmann and Black, 2015].

Experts use individual mathematical manipulations as sources of physical sensemaking. Kustusch et al. [2014] studied physics professors solving a thermodynamics problem that involved taking partial derivatives. There were many choices for which derivatives to take, and experts used physical insight into the derivatives' meaning to guide their choices. In a review of the literature on mathematical sensemaking inside the mathematical manipulation steps of problem-solving, Kuo et al. [2013] found "no studies that focused upon the mathematical processing step in quantitative problem solving or described alternatives to using equations as computational tools." The same authors then contrasted two students, one who describes a kinematic formula in terms of its meaning via a symbolic form, another who saw the formula essentially as a black-box tool, and found that these students performed the mathematical manipulations in a problem using that kinematic concept differently. The student who understood the formula via a symbolic form was able to blend mathematical and physical reasoning to take a shortcut solution to the problem, while the other student was not.

If we value this sort of blended sensemaking, we should find ways to encourage it in students. We believe extreme-case reasoning is one way to do this. In order to use extreme case reasoning, students must think about formulas and physical systems simultaneously, and as a result, they find new and creative ways of conceptualizing and manipulating equations.

For example, Myra, while considering the "springs in series" problem, has written $\frac{T}{k_1} + \frac{T}{k_2} = \Delta L_{total} = \frac{T_{sum}}{k_{series}}$ and below it $\frac{T}{k_1} + \frac{T}{k_2} = \frac{T}{k_{series}}$ on her whiteboard, saying

> I'm thinking that if you apply a constant force, for k-one will give like this amount of length plus k-two will give like this amount of length,

> then that's like the total amount of length of the series, which equals to k over T-series. And that makes sense to me. I just don't know how you would like not put the T in the equation.

Although the group did not take up her method and she soon abandoned it, Myra's expression was correct, and a short algebra step away from the desired solution. In generating this expression, Myra didn't start with basic definitions and follow a purely formal procedure. Instead, she blended her conceptual understanding of stretching with the mathematical formalism while manipulating mathematical expressions.

Shortly before, Lelia stated, "and both would contribute just like one would contribute like one would have less change than the other. They'd still both probably be a part of the stretch." Myra's key insight was to translate this "both contributing" intuition into a symbolic form [Sherin, 2001], a basic template for an equation, along with a meaning used to understand entire classes of equations that build on that template. Here, Myra uses what Sherin identifies as the "parts of a whole" template, $[\square + \square + \ldots]$.

Myra fits Lelia's idea about both springs contributing stretch onto this template via the heuristic equation $\text{stretch}_1 + \text{stretch}_2 = \text{stretch}_{total}$. Then, using the definition of a spring constant, which contains a variable ΔL for the stretch of the spring, Myra substitutes in the stretch of each spring, making each term physically meaningful as she does, obtaining $\frac{T}{k_1} + \frac{T}{k_2} = \frac{T}{k_{series}}$.

In a separate instance, Bert was working on the half-Atwood problem. His solution had a sign error, $a = \frac{mg}{m-M}$ instead of the correct $a = \frac{mg}{m+M}$, due to an inconsistency in how he set up his coordinate system.

The interviewer introduced and scaffolded the extreme and special case game for Bert, who readily took it up, discovering that his solution had the blocks reversing direction based on their mass, which he rejected as intuitively incorrect. Instead of reworking the entire problem from scratch, Bert tried making small modifications to his answer to eliminate the problem, for example introducing an absolute value in the denominator to keep it from changing signs. As he continued introducing and testing new solutions, he looked at $\frac{M-m}{mg}$ as a potential solution, considered the extreme case where $M \gg m$, and said

> So then this is super big that's super small. [pauses, draws a minus sign on M in the numerator] Still doesn't make sense. Still not working. Cause one of these [the masses] are big then it's gonna be big acceleration. That's not what should happen. Should be as this one grows [points to M] it gets smaller, so like that has to be in the denominator.

In suggesting that M must go in the denominator, Bert has repurposed the extreme cases game. Instead of evaluating potential solutions, he is placing constraints on what the unknown correct solution must look like. Like Myra, he blends his physical intuition and symbolic forms to achieve this.

The symbol template Bert uses is a division template, $\frac{\square}{\square}$, along with a conceptual schema about inverse proportionality. It is a schema where as one quantity increases, another decreases, but in the extreme case it shows that as one quantity grows very large, another becomes very small.

In applying this symbolic form, Bert begins with his intuitive understanding that very large, heavy objects are difficult to move and blends in his formal understanding of inverse proportionality to creatively generate a new instance of the extreme case game.

Bert did not wind up solving the problem; he rejected the correct solution on the mistaken grounds that it was symmetric with respect to interchange of m and M, but despite not coming to a complete solution, he generated unique insights as well as a partial solution by renegotiating his relationships to the equation he was searching for while playing the extreme case game.

2.5 Implications for Instruction

It is common to see backsliding in surveys of student epistemologies over the course of introductory physics. For most courses, students on average exit their college physics course with less-favorable beliefs about how to learn physics than they had when they entered [Redish et al., 1998, Adams et al., 2006]. As epistemologies are tied to problem solving strategies [Ataide and Greca, in press], it's likely that students' conceptions of the role of mathematics and their approaches toward using it also deteriorate over most year-long introductory sequences. This means that although we observed surprising and expert-like strategies in our problem-solving interviews, we need to be wary of the possibility that our classes lead to students using these strategies less and less with time.

The reward and feedback structures in many introductory courses focus on evaluating whether a student can perform a certain calculation correctly. This includes grades on homework and exams, and in many circumstances, the verbal feedback students receive from instructors, for example that in ”initiate-response-evaluate” questioning [Mehan, 1979]. In most of the episodes we've cited in this chapter, students wouldn't have received positive feedback from such systems. Bert didn't get the correct answer when he found creative new applications of extreme case reasoning. Myra blended her physical intuitions with formal mathematics in a symbolic form to get an expression equivalent to the correct answer for how springs add, but her group didn't take it up, and they left the interview without have reached a consensus on the correct answer. Alma, when checking the special case of a circular ellipse, used a dynamic ontology of the equation (see chapter 3 for more on ontologies) to reinforce her understanding of the test she was performing, but wasn't able to distinguish two answers which both passed that test, and she wound up choosing the wrong answer. Each time, the students were displaying expert-like problem-solving behaviors that we might not expect to see in introductory courses, but because they didn't come to the correct final conclusion, in many classrooms they wouldn't have received points on a test, heard their teachers praise, reiterate, extend on, or dive more deeply into the reasoning, or seen their peers enthusiastically take up the same methods. Because the type of feedback students receive can significantly affect their attitude toward learning [Carlone et al., 2014, Russ et al., 2009], this lack of positive feedback when trying expert-like strategies could easily quench students' fledgling attempts at useful, general ways of solving problems and understanding physics.

It isn't surprising that the techniques that work for experts in problem solving are less effective for novices. Learning to use tools takes practice. Riding a bicycle is much faster and more efficient than walking once you know how to do it, but it can be wobbly, frightening, and even dangerous at first. If we want students not only to try out strategies such as testing special cases or blending intuition and formalism through symbolic forms, they need a freedom to fail, encouragement to try out new ways of thinking, and positive reinforcement when they do so. Spike and Finkelstein [Spike and Finkelstein, 2016], studying recitation sections, found that the extent to which TAs do these things depends on their beliefs about the goals of instruction. When instructors expand their goals beyond seeing students perform calculations correctly (whether quantitative or qualitative) and value the growth of new and useful ways of thinking, classrooms environments can take the seeds of expert-like thought we've observed here and nurture them.

In our own courses, these observations have led us to two ways of encouraging new problem-solving behaviors. The first is asking questions which focus on evaluating the meaning of formulas, as opposed to using them as black boxes. For example, a problem from the textbook by Serway and Jewett [2004] reads

> Consider a gas at a temperature of 3500 K whose atoms can occupy only two energy levels separated by 1.5 eV . . . Determine the ratio of the number of atoms in the higher energy level to the number in the lower energy level.

The solution involves using the formula for the Boltzmann factor as a black box tool. To encourage different ways of reasoning about the formula, in a class one of us (Redish) taught recently, a quiz question asked

> When a membrane allows one kind of ion to pass through and not another, a concentration difference can lead to an electric potential difference developing across the membrane. For example, if the concentration of NaCl on one side of a membrane is $c_1 = 10mM$ and $c_2 = 2mM$ on the other, letting only Na+ ions through (and not Cl-) will build up a potential difference across the membrane. This is controlled by the equation that says that the electric potential energy, $q\Delta V$, balances the concentration difference effects via the Boltzmann factor thus:
>
> $$\frac{c_1}{c_2} = e^{\frac{-q\Delta V}{k_B T}}$$
>
> For a given set of concentrations (c_1 and c_2 fixed) would you expect increasing the temperature to increase , decrease, or leave the Nernst potential, ΔV, unaffected?

This question encourages students to reason about the functional form of the Boltzmann factor, perhaps by imagining extreme cases or using symbolic forms. It also encourages students to think of T not as a fixed entity, but as a parameter that can be tuned to change both the physical behavior of a system and the numerical value in an equation.

In addition to asking questions that encourage students to reason about formulas instead of apply them in order to get the right answer, we also ask questions that encourage students to reflect on formulas without the need to extract a final correct or incorrect answer. For example, in one of our recitation exercises, students are asked to construct their own equation to describe when a worm will begin to suffocate as we scale up its size (reducing its surface area to volume ratio) [Redish and Cooke, 2013]. We then ask students,

> Our analysis in [the previous part] was a modeling analysis. An organism like an earthworm might grow in two ways: by just getting longer or isometrically – by scaling up all its dimensions. What can you say about the growth of an earthworm by these two methods as a result of your analysis in [the previous part]? Does a worm have a maximum size? If so, in what sense? If so, find it.

These more open-ended and reflective questions ask students to use formulas - formulas they have constructed - for interpretation and coming to new inferences, both about physical systems and about the mathematical properties of equations.

Throughout this chapter, we have searched for a number of creative ways students approach problems, including thinking about the extreme cases, conceptualizing parameters in different ways, and using equations for estimation. In interviews, students do all these things, but they can easily lead the student seemingly nowhere—no correct answer to a question, no encouragement from an instructor, no adoption by peers. To encourage students to try out useful but difficult-to-master new strategies, we continue refining the way we ask questions and attend to student thinking during instruction.

Chapter 3: How do students' ontological metaphors for equations change depending on what epistemic game they play?

3.1 Overview

In this chapter, I explore the cognitive resources students use to think about equations, especially the ontological resources - those that give students a grasp on the type of thing an equation is. Analyzing student utterances, I identify a small corpus of ontological metaphors used to describe equations and develop a codebook for the ontological metaphors. I also coded the same data for the epistemic games students were playing, and found several significant correlations between ontologies and epistemologies. I'll begin here with the idea of ontological metaphors.

Equations such as $y = x^2$ or $E_z = \frac{kQz}{(z^2+R^2)^{3/2}}$ are abstract. These equations are distinct from the ink on a page used to represent them, or even from the concept of ink in a certain pattern on a page. Suppose, hypothetically, that two people happen to read this thesis. The first sees the equations represented as ink on paper and the second as pixels on a screen. If they compare notes afterwards, they will probably believe they read the same equations, even though the physical representations were completely different. It's unlikely that they would even have considered the possibility that they read different equations had the topic not been discussed here (and assuming there are no printing errors or corrupted PDF files along the way)!

In *Metaphors We Live By*,Lakoff and Johnson [2008] describe how, in order to reason about such abstractions, we think of the abstraction as if it were a more familiar, more concrete thing. They cite examples including fairly specific metaphors such as *the mind is a machine*, evidenced by phrases such as "My mind just isn't operating today" and "Boy, the wheels are turning now" (used to refer to someone thinking). They also cite more generic metaphors such as *inflation is an entity*, used implicity in the sentence "Inflation is lowering our standard of living."

For Lakoff and Johnson, this sort of metaphor is inextricable from how we think about abstractions:

> Once we can identify our experiences as entities or substances, we can refer to them, categorize them, group them, and quantify them and, by this means, reason about them.

We began our project interested in how students reason about equations. Because equations are abstractions, we expected them to be dealt with metaphorically. And because metaphor structures the way we think about abstractions, understanding the metaphors we use to reason about equations is an important step towards understanding how we make physical sense with them.

In this chapter, we examine the metaphors used for equations. Working from a resources perspective, we show that students can access a variety of ontological metaphors; in fact the variety is considerably wider than we have seen in, for example, introductory physics texts. The metaphors students use depend on their needs in the problem-solving situation, and on the individual student or group of students. To understand something of the structure of how different metaphorical resources are accessed in different circumstances, we relate the equation ontologies (henceforth an alias for “ontological metaphors used for equations”) to the epistemic games students were playing when using that ontology.

Equation ontologies can be studied on a small grain size; each one is apparently independent, and students switch fluidly back and forth between different ontologies. This allows us to look for patterns within epistemic games, which play out over the course of minutes, as compared to seconds for a particular use of an equation ontology. These short units of coding allows us to detect lability in student cognition that we likely wouldn’t see on the timescale of epistemic games.

We also thought that equation ontologies were under-studied, and could potentially be of wide interest, as we discussed in section 3.5.

3.2 A note on ontologies

In philosophy, ontology is the study of existence. For example, a philosopher studying ontology might take it for granted that the ink-on-paper which represents an equation exists, and that the computer monitor on which an equation’s representation can be displayed exists, and then ask whether the equation itself truly exists [Effingham, 2013].

My meaning in this chapter is slightly different. In practice, students do not wonder whether an equation “really exists”. They reason about the equation in various ways, thinking about it as if it exists. It’s this thinking that we’re interested in.

In information science, there is another concept of “an ontology” (as distinct from “ontology”, the philosophical field). An ontology for a domain is a graph (usually a tree) of all the concepts in the field. Each node on the graph is something that exists; each edge a connection between the ideas, procedures, etc. that exist. For example, the Ontology for Biomedical Investigations [Bandrowski et al., 2016] is a large graph including specific nodes like “blood plasma” and general nodes like “material entity”. These nodes are connected (via intermediates) because blood plasma is a material entity. Other aspects of the ontology include data analysis procedures or lab techniques.

This concept of ontology is significantly more formal than ours, but provides a useful analogy. We might think of a human’s concept of what things exist as something like this formally-structured ontology, with nodes like “entity” and “machine” and “my car” connected in a chain from parent to child. Nodes may inherit properties from their parents. We already know a lot of things about machines generally (e.g. they have parts that can wear down), and when we categorize our car as a machine, we import the general properties of a machine when thinking about our car. When we think of a mind as a machine, we are similarly making the mind a child node of machine, and so can use the things we already know about machines to reason about minds (correctly or not). Lakoff and Johnson caution that this sort of inheritance of properties is only partial, so we don’t wish to state that cognitive ontological metaphors have the same cleanly-defined structure as the ontologies of information science.

3.3 Expert Ontologies of Equations and Teaching Introductory Physics

Here are a few examples of physicists writing about the relation between the Yukawa potential, $V(r) = \frac{qe^{-mr}}{r}$ and the Coulomb potential, $V(r) = \frac{q}{r}$.

> In the limit of $m \to 0$ the Yukawa potential becomes the Coulomb or gravitational potential. . . [Heile, 2015]

> ...if we choose . . . $m_0 = 0$, the potential reduces to the Coulomb potential energy. . . [Townsend, 2000] [source uses m_0 in place of m]

> We can take the limit $\alpha \to 0$ and recover the Coulomb potential.[Hassani, 2013][source uses α in place of m]

> The Coulomb potential of electromagnetism is an example of a Yukawa potential . . . [Wikipedia, 2016]

> We see . . . that if the mass m of the mediating particle vanishes, the force produced will obey the $1/r^2$ law. If you trace back over our derivation, you will see that this comes from the fact that the Lagrangian density for the simplest field theory involves two powers of the spacetime derivative . . . [Zee, 2010]

In some cases, physicists see themselves as enacting a change in the Yukawa potential. They or their reader actively “take the limit” or “choose $m = 0$”. Other times, the Yukawa potential changes, but there's no clear agent involved. It may "become" or "reduce to" the Coulomb potential and the mass may "vanish", but no entity is identified as enacting the change. In contrast to these dynamic descriptions, the relationship can also be described statically. Nothing in particular is happening when the Coulomb potential "is an example of" the Yukawa potential.

This is just a sampling of physicists' language on the topic. The details of how they describe the Yukawa potential-Coulomb potential relationship may depend on both the physicist and the context of what they're communicating in complicated ways. Our goal here is simply to illustrate that there is a significant diversity of ways to conceptualize of an equation.

These examples come from professional, graduate, and upper-division undergraduate material, where such a diversity of conceptualizations of equations is commonplace. By contrast, in introductory physics textbooks, equations are usually treated as static entities to be scrutinized.

> Outside the nucleus the nuclear force is negligible, and the potential is given by Coulomb's law, U(r) = +k(2e)(Ze)/r,... [Tipler and Mosca, 2007]

> Coulomb's law can be written in vector form ... as $\tilde{\mathbf{F}}_{12} = k\frac{Q_1Q_2}{r_{21}^2}\hat{\mathbf{r}}_{21}$ ... [Giancoli, 2000]

> The electric force acting on a point charge q_1 as a result of the presence of a second point charge q_2 is given by Coulomb's Law: $F = \frac{kq_1q_2}{r^2} = \frac{q_1q_2}{4\pi\epsilon_0 r^2}$ [Nave, 2017]

The main exception we have observed to this "equations as static entities" ontology is in descriptions of formal operations on equations. These descriptions come up during derivations (e.g. "differentiate with respect to t", "set them equal to each other", etc.). Also, equations are sometimes described as active entities, for example "Coulomb's law describes a force of infinite range which obeys the inverse square law [Nave, 2017]" in that they "describe" things, but this does not represent the same diversity of conceptions we saw with regard to the Coulomb and Yukawa potentials.

This mostly-static view of equations stands in contrast to introductory physics sources' descriptions of the physical quantities the equations represent

> We can divide up a charge distribution into infinitesimal charges . . . [Giancoli, 2000]

> The force exerted by one point charge on another acts along the line joining the charges. It varies inversely as the square of the distance separating the charges and is proportional to the product of the charges. [Tipler and Mosca, 2007]

In describing the force, field, or charges associated with Coulomb's law, introductory textbooks use both agentive language ("We can divide") and non-agentive ("The force . . . acts. . . "). The second quotation here also mixes dynamic ("varies inversely. . . ") with static ("is proportional to. . . ") language in the same sentence. So while a diversity of ontological viewpoints are generally considered acceptable for thinking about physics in introductory settings, this seems to apply much more to physical quantities than to equations. As we move to more expert settings, the equations themselves take on the same diversity of ontologies.

Sfard [1991]'s notion of conceiving of functions as either objects or processes is similar to ours, but here we consider "process" views where the equation itself is changing, as opposed to Sfard's notion of a static function which describes change when inputs transform into outputs. Our point here is simply to illustrate one more small piece of the diversity in expert conceptual systems used to make mathematics physically meaningful. This piece, like symbolic forms, is never explicitly taught. It is a part of the hidden curriculum, and something we can try to find evolving in students as they progress towards expertise.

3.4 Novice ontologies of equations

In physics education, there has been considerable effort to understand the ways the different ways that students view equations epistemologically [Airey and Linder, 2009], e.g. whether they ought to map closely to phenomena or be treated formally, be accepted as given by authority or derived from fundamental principle, and their relationship to modeling. Here, we are interested in a different type of view of equations: their ontology, or what types of object they're considered to be.

Earlier, using the example of physicists discussing the Yukawa and Coulomb potentials, we suggested that there is a variety of ways that physicists conceive of the equations they're working with. Physicists in different contexts speaking to different audiences sometimes thought of equations as dynamic objects, with one equation transforming into another, and other times thought of them as static, with one equation being a special case of another. Additionally, when equations changed, sometimes it was the speaker or the audience actively making the change, and sometimes the equation changed without a specific agent being identified.

The three problem-solving strategies introduced so far all call on students to think about equations in new ways—to hold them accountable to common sense (estimation) and to check various features of them (dimensions and special cases). We might wonder whether interacting with equations in certain ways changes the conceptualization that students have of equations.

In watching students play epistemic games with mathematics, we saw a diversity of conceptualizations of equations emerge. For example, Alma, in working the ellipse problem, checks the special case $a = b$ with reference to the formula $A = \pi \left(\frac{a+b}{2}\right)^2$

> ...so a plus b squared over two squared times two is four plus b that would be 2 ab. b squared plus yeah. okay. yeah. okay. so then you would have r squared plus two r squared plus r squared which equals pi four r squared over four, so I guess it's a plus b over two cause you're taking the average. Oh, it's like you're turning into a circle. that's cool. yeah.

By checking the special case, the ellipse is "turning into a circle", but Alma makes this reference not while working with the geometric object, but with the equation and substitutions on it that she was making. In other words, the ellipse is "turning into a circle" in that it becomes the formula for the area of a circle when $a = b = r$. This dynamic picture of an equation mirrors that a Yukawa potential that "becomes the Coulomb" potential in an extreme case. She is working with the formula, but instead of saying that the formula turns into a formula for a circle, she says "you're turning it into a circle", referencing a geometric object (the circle) while working with a non-geometric object (the formula). We suggest that for Alma, in this moment, there is no significant distinction between the formula and the object it describes, which, if correct, shows a very strong example of binding meaning to an equation.

Similarly, Amelia was examining the equation $N(t) = N_0 e^{-t/\tau}$ for the number N of particles remaining when they decay over time t with a time constant τ. (The interview protocol for this series of interviews is not available because the interviewer asked the questions verbally, writing equations out by hand on a whiteboard. Videos of these interviews for scholarly review may be available on request.) In examining the special case where half of the original number of particles remain, Amelia described actively changing equations via procedural language, such as "I divide each side by the initial amount. I el-en [take the natural logarithm of] each side", but she also described changing equations not according to any fixed procedural rules, "I changed the equation, if I'm doing this logic, because I don't remember what the half life equation is off the top of my head. So I rewrote the equation to say that $Q(t)$ is equal to one half times the initial amount times e to the negative t. t referring to just time..."

In both cases, the agency in changing the equation lies in Amelia herself. In the first case, she follows formal manipulations. In the second, she is "doing this by logic", presumably a reference to some mix of common sense, intuition, and mathematical reasoning, as a contrast to memorization. She created an entirely new equation based off a template from the old one, assigning specific physical meaning to each term she created.

Students can take varied stances towards the types of objects that equations are while manipulating, creating, and interpreting them in many contexts, not simply in the context of the strategies we investigated. We believe this menagerie of conceptualizations of equations and interactions with them is especially rich in these epistemic games that play out with these strategies due to their requirements to blend symbology and physical meaning.

3.5 The usefulness of studying ontologies in Physics Education Research

Ontological metaphors, or simply "ontologies", are widely studied in PER. Most of this work studies ontological metaphors related to specific physics concepts, for example, "current is a fluid", but in section 3.6 I'll look at some literature which studies ontological metaphors in equations.

There is significant disagreement within the field on how we should think about ontologies, both from research and instructional perspectives. Researchers such as DiSessa [1993] and Gupta et al. [2010a] describe a dynamic, knowledge-in-pieces view of ontologies. Here, students don't necessarily have a single ontological metaphor for a physics concept. Based on the context and their prior experience, they activate different ontological metaphors at different times for different purposes. Other researchers such as Chi and Slotta [1993] hold that students reason from a largely-coherent ontological perspective.

Those are differences in research perspectives, but they lead to significant differences in instruction as well. From a perspective of ontological coherence, one of the biggest difficulties students face in learning a new subject is that they are using the wrong ontological metaphor. For example, we might conceive of some scientific concept using either a "process/interaction" metaphor (e.g. "gravity pulls objects together") or a "substance" metaphor (e.g. "there is a lot of gravitational influence here"). In a world where ontological metaphors are coherent and at best slowly-changing, one of these metaphors is productive for students and the other is disruptive. Instruction should attempt to discover what metaphor students are using, what metaphor is best, and, if they aren't the same, what types of interventions will help students transition to the more-productive metaphor.

Meanwhile, if students activate different ontological metaphors based on context, either metaphor might be helpful or harmful, depending on what the student is trying to use it for. For example, Gupta et al. [2014] describe how a metaphor for gravity as an object led a group of teachers to insightful reasoning about why objects of different mass accelerate in the same way, even though some might consider substance models for gravity incorrect or non-canonical.

Although researchers do not agree on how ontological metaphor works or what it means for instruction, Lakoff and Johnson's fundamental assumption - that we reason about the abstract via metaphor - seems largely unchallenged within PER.

That debate and interest in ontology continues so strongly today is evidence that studying the ontological metaphors behind equations may lead to results of interest and value to the community.

3.6 Existing work on ontological metaphors in equations

3.6.1 Algebraic terms are objects and parts of an equation are locations to which they can be moved

Although we did not find any work specifically analyzing students' ontologies of equations in the same way we intend to, our work builds on a variety of perspectives on ontologies.

For example, Wittmann et al. [2013] studied how students make algebraic manipulations of equations, uncovering an important ontological metaphor in the process. They write that students, "treat the terms of the equation as physical objects in a landscape, capable of being moved around." They cite evidence from both the students' language and their physical gestures to argue that students were invoking their concrete experience of motion to understand mathematics. Students already had an understanding of moving things around from one place to another, including for example knowledge that the thing is the same after this operation, but exists in a new place. Metaphorically thinking of terms in an expression as objects being moved, they made sense of algebraic manipulation rules.

Specifically, in separating variables in the equation

$$mv\frac{\mathrm{d}v}{\mathrm{d}x} = mg - c_2v^2$$

students would think of grabbing the "dx" and dragging it from the denominator of the left hand side to the numerator of the right. They did the same with $mg - c_2v^2$. In doing so, they grouped all these symbols into a single entity to be dragged.

Wittman et. al.'s findings in PER are closely aligned with Lakoff and Johnson [2008]'s more general arguments. Lakoff and Johnson believe that metaphors are "embodied". By this they mean that that the ultimate root of how we understand a metaphor is our physical experience with the world. Many metaphors, for example, are based on the "up/down" distinction. An example is speaking about "high expectations" which can then be "raised" or "lowered". This metaphor is based on our experience with gravity.

Wittman et. al.'s finding of algebraic metaphors being based on objects moving from place to place is exactly the sort of thing Lakoff and Johnson would expect, lending credence to our key assumption that their view of ontological metaphors will be productive in understanding how students think about equations.

It's interesting to note that in addition to being productive, these metaphors are informal. Most instruction in algebraic manipulation includes procedures like "add the same quantity to both sides of the equation" or "divide both sides of the equation by two", but in the heat of discussion during a problem-solving session, students use a much less-formal mode of speaking and thinking. (Wittman et. al. argues that the speech and gesture seen aren't just verbal analogies standing in as a shorthand for more formal ways of understanding the manipulation. Instead, the dragging-objects metaphor is the manipulation in the students' conceptual systems.)

Wittman et. al. also raised an important open question. We won't try to answer it here, but discussing it sets some background for us. Students metaphorically moved terms from one part of an equation to another, and usually did this in a way that yielded mathematically valid results. Simple movement doesn't place much constraint on where the moved term goes; one could easily have moved it to the denominator when it should have moved to the numerator, or moved it within the right hand side of the equation when it should have crossed the equal sign to the left hand side, etc. Evidently, the metaphorical structure of movement isn't sufficient to enforce correctness. But if the metaphor itself isn't encoding what manipulations are correct, allowed, or productive, why even create the metaphor? If the rules of algebra must be layered on top of the metaphor as extra constraints, why not just use the rules without this additional metaphorical structure?

Perhaps the metaphor provides a substrate on which students can more reliably learn and apply the rules of algebra, but as of right now, this isn't known. Alternatively, it may be that the metaphor has other functions entirely, such as being mostly about facilitating communication between two or more people working an algebra problem, rather than ensuring that an individual makes legal algebra moves. The metaphor might not do any work in helping do algebra correctly, but it might be enormously useful in telling someone about what you did.

This question about why students often use the motion metaphor accurately and what they gain from the motion metaphor sets the stage for our own study of ontological metaphors for equations. Do the metaphors we use to think about entire equations exist to improve our accuracy? Do they instead facilitate developing our symbol sense [Arcavi, 1994] (e.g. the goal-directedness of the manipulations we choose)? It isn't enough to make legal algebra moves, since one can make scores of legal algebra moves only to wind up proving $X = X$ or some such.

Or, might metaphors result as natural outcroppings of the way we frame a mathematical task? If a student is thinking of a mathematical task as an act of creation, will they think about equations differently than if they see the task as an act of uncovering a known answer that was simply hidden from them?

3.6.2 Mathematical ontological metaphors evolve in a predictable way on the path to expertise

I have a hobby of writing answers to questions online, and one of the most surprisingly-popular answers I've written is about why mathematicians invented complex numbers [Eichenlaub, 2012]. It recounts a story from Needham [1998]s *Visual Complex Analysis* about what caused people to take the idea of complex numbers seriously.

No one would have given complex numbers much consideration if they were merely a "way for $\sqrt{-1}$ to exist". Instead, mathematicians would just have held the seemingly-obvious position that $\sqrt{-1}$ doesnt exist. It was only when complex numbers became crucial to a procedure for solving cubic equations - a major open problem at the time - that mathematicians took them seriously.

This historical progression in ontology, from dubious phantom to useful computational tool to complete kosher entity, seems to be replayed again and again in miniature with students as they learn. After learning the basics of complex algebra, reflective students ask questions such as

> Sometimes a quadratic equation gives an imaginary number. What is one and how can they exist?

It's a questions we ask over and over after learning the basics of manipulating abstract concepts."Yes, but what actually is this thing?!" I've heard people ask it about electrons, spacetime, quantum states, infinite sets, etc. Dissatisfaction about not knowing what a thing is has probably contributed more than a few cohorts to the legions of scientific and mathematical cranks.

Sfard [1991] proposes that the development of mathematical ideas always follows the same path. In both the mathematics community and in individuals, mathematics begins as a process, a set of rules or actions that act on things we already understand, and eventually wind up as objects themselves, ready to be acted on by the next idea. This provides a detailed theory about mathematical ontological metaphors, how they develop, and how they're related to understanding and creating new mathematics. Because the theory is so general, it surely makes the strongest claims about the roles of ontological metaphors in mathematical thought that I've encountered.

For an example of how Sfard outlines ontological metaphors evolving, a mathematical function can be thought of as sucking in x and spitting out $f(x)$. This is how almost every student will think of it, and indeed how it is generally taught.

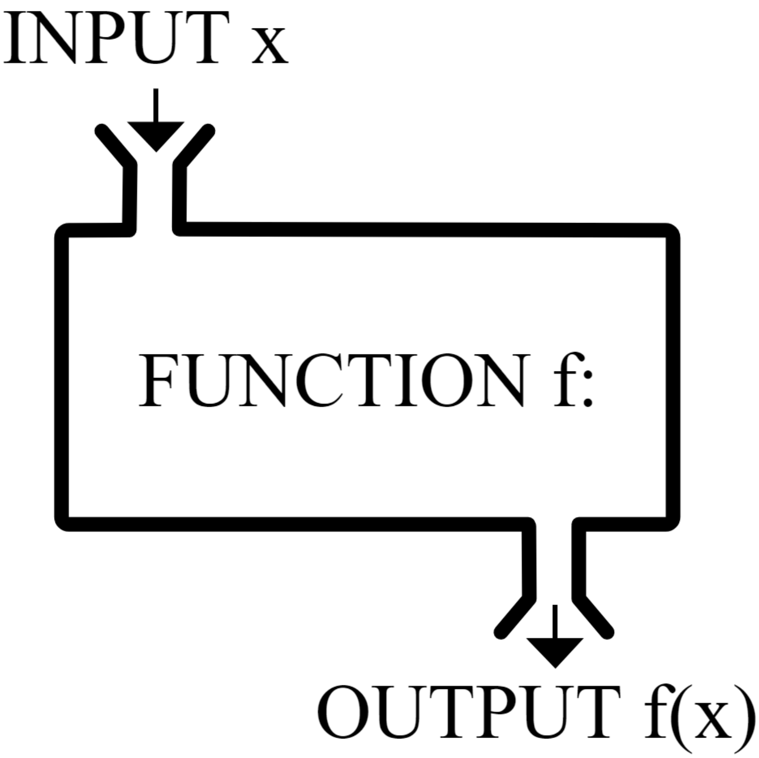


Figure 3.1: The function as a process - a machine that takes in x and spits out y

Figure 3.1 shows an example of this conception of a function from Wikipedia contributors [2018], and a Google search found many other pages using the same basic idea.

Crack a set theory book, though, and you'll learn that a function is a set [Halmos, 2017]. Specifically, it's a set of ordered pairs (x, y), with x from the domain and y from the range. Nothing is "happening" in a function. It's not turning into y, or associating with y, or mapping onto y, or any other verby-thing. Its a static entity, not a description of "what youre supposed to do to x".

It took a long time for us to get this static sort of function. Euler struggled with the idea of what a function really is, and wound up with a vague, process-based definition. Sfard attributes the modern definition to Bourbaki in the early 20th century, and even then mathematicians resisted it for a while.

I remember reading this definition in Halmos [2017] and immediately liking it. By that point I'd done a few years of college and used functions daily for long time. I was already used to thinking of functions as objects from such exercises as considering the action of a particular particle trajectory (i.e. "take this entire function mapping time onto position and feed the thing into a machine that spits out a number for the action"). Since I was feeding the entire trajectory into an "action-calculating machine", I needed to think of the trajectory as a single object. Even outside such contexts, I had plotted functions thousands of times, which encouraged me to view the function as a single static entity, since that's what a graph is. I was ready to hear about a rigorous way to treat a function as an object because I had come to terms with that long ago.

Even after we're capable of thinking about functions as objects, we still use the process picture in everyday life when it's convenient. Sfard gives plenty of examples of mathematicians moving fluidly back and forth between them. Here, I'll speculate that a lack of this fluidity could contribute to student difficulty regarding position, velocity, and acceleration.

In a typical classroom exercise in 131, we used a sonic ranger to track the motion of a flat board as it moved back and forth, then made position, velocity, and acceleration vs. time plots and asked questions about them. We would point to some part of the plot and ask what's going on there, ask what it means about what the other plots will look like at the same time, etc. Answering such questions means flipping back and forth repeatedly between thinking of the graph as an object, thinking of the motion as a process, and even thinking of the graph as a process when we trace it out. It might be the process/object duality that gives students as much trouble as the derivative/integral relationship we're studying.

Sfard proposes three stages to learning a mathematical concept:

1. interiorization: getting familiar with the details of the process (e.g. learning to count)

2. condensation: thinking of the process as a single step (e.g. understanding the idea of counting every grain of sand on a beach)

3. reification: suddenly understanding the entity behind the process (e.g. realizing that numbers exist independently of counting)

This is a bold thing to do. We know that experts have "chunked" their knowledge, with the classic example being chess players described in Miller [1956]. Sfard attempts to give a description of how chunking occurs, going so far as to diagram out expert reified knowledge versus a novices unorganized view in a graph-like schema.

Our methodology in this study is to investigate these sorts of claims about the structure of knowledge with interviews. We'd call in students at various stages of development, have them solve problems and explain their thinking aloud as they go, and examine the evidence, or lack thereof, for going through each of the three stages Sfard proposes. Sfard's epistemology is fundamentally different, building a case mostly around a consistently-repeated pattern in the historical development of mathematical ideas, combined with informal field observations of students. This suggests a method for follow-up work to the results we find here. If we believe we've gained insights into student ontological metaphors for equations via interviews, later work could investigate the historical development of those same metaphors in the mathematical community at large, mirroring Sfard's approach.

Sfard acknowledges that experts can often skip straight to reification; the full three steps are not necessary for them, but she posits they are for novices. But if the ontological journey Sfard describes explains so much of what it means to come to terms with a new field, why do so many people want a good explanation of why a negative numbers times another negative nummber is a positive number (for example, such a question is among the top 0.04% most popular questions on the highly-active math question site Math Stackexchange [Sev, 2010])? I suspect that many of the people interested in this have reified negative numbers long ago; it's the process of multiplication they dont feel they understand. Sfard, I think, goes so far as to identify reification and understanding, which I think sells understanding short. Nonetheless, even if reification is only one piece of understanding, and then only usually a part of that process, it still demonstrates the strong role that ontological metaphor plays in how we think about math, and justifies our interest in the subject.

Many of the things Sfard says about math seem to apply to physics as well. For example, in an interview for the PBS program *Closer to Truth* [Tho], Kip Thorne gave a clear description of how physicists reify the idea of spacetime to view the entire past and future of the universe as a single, extant, four-dimensional object, rather than as a process unfolding in time.

Thorne: There is a unified space and time all unified together. Einstein, when he saw this reformulation of his laws, he didn't think much of it... until several years late he discovered he had to have the unification of space and time in order to formulate his laws of general relativity. ... That means spacetime after unification is absolute. There's nothing personal about it. Nothing relative about it. We make it personal by the manner in which we observe. The manner in which we move. Now on the other hand it is so difficult to think about this unified spacetime that even I as a physicist a large fraction of my work make a choice, I will take my personal point of view and I will talk about space and I will talk about time. ... I slip back into the personal point of view because I have deeper intuitions. Or it's easier. It's easier to think about space and time from the personal universe ...

Interviewer: How then does the future exist in the block universe? Is there some deep philosophical...

Thorne: Well the future is here [points up above his head] and the past is there [points around waist level]. I don't think there's any deep philosophy. There's mathematics. ... That interpretation of the mathematics has a certain philosophical content, but philosophers often take this beyond where we want to go. We want to go only far enough that it is useful to us.

Thorne validates many of Sfard's points about the development of ontological metaphors from process to object. His "personal" view of space and time is the one he describes using to talk to layman, and the one Einstein originally used to formulate relativity. This is a process-like view of the universe because in it, time unfolds actively, and the events of the universe's history play out like a movie. The "block universe" view is an object-like metaphor, and in Thorne's account it develops in individual physicists according to the same path it did in the community of physicists. Physicists also switch fluidly back and forth between the metaphors, choosing them based on their utility, not philosophical considerations of their ultimate truth. This example illustrates how the reification process Sfard describes isn't limited to pure mathematics, and is also important to the mathematics used in physics.

Richard Feynman also described the reification process from an introspective point of view in *Surely Youre Joking, Mr. Feynman!*[Feynman, 2010]

> I had a scheme, which I still use today when somebody is explaining something that I'm trying to understand: I keep making up examples. For instance, the mathematicians would come in with a terrific theorem, and they're all excited. As they're telling me the conditions of the theorem, I construct something which fits all the conditions. You know, you have a set (one ball) - disjoint (two balls). Then the balls turn colors, grow hairs, or whatever, in my head as they put more conditions on. Finally they state the theorem, which is some dumb thing about the ball which isn't true for my hairy green ball thing, so I say, "False!"

Feynman is describing being an expert and quickly reifying objects appropriately. It does seem that this process is rooted strongly in analogy, but I think its details, especially among students (who are unlikely to be able to reflect on and articulate their thoughts as well as experts sometimes do), are still not known.

From an instructional perspective, Sfard suggests that a period of rote manipulation may be essential because you need to go through each of the stages. You need to spend time "interiorizing" and "condensing" a concept before you can "reify" it. Its not quite clear why this should be true for students and not for experts, though.

Sfard's work leads us to believe that studying ontological metaphors is important because they develop over time and play a crucial role in students' facility with and comfort with new ideas. It seems likely that the blended sensemaking we described in chapter 2 is made possible only by a process of reification, or that it helps students to reify the ideas they're working with.

3.7 Perspective, framework, assumptions

3.7.1 The cognitive perspective

Our analysis of equation ontologies takes a cognitive perspective - it looks for how these metaphors work in the minds of individuals.

There are good reasons to consider other perspectives in researcher on ontological metaphors and their development. In section 3.6.1, we considered the possibility that certain ontological metaphors' main function isn't necessarily to help students think more clearly about mathematics for themselves, but to have a method to communicate it with other students. It could make sense to take a more social perspective in studying metaphor.

Partially, our choice is driven by the data we collected. Most of our data is from individual interviews; we conducted the interviews before defining research questions around ontological metaphors for equations.

More importantly, it's psychologically implausible that, even if metaphors are playing important roles in interpersonal communication, they would not structure individuals' thoughts as well. The causality could run purely from cognitive to communication, i.e., we develop metaphors to think about equations for ourselves, and then because we have those metaphors internally, we use them for communication. It could run from communication to cognitive as well. Maybe communication with teachers, and the need to communicate with other students, drives students to develop ontological metaphors for the abstract, and once they've adopted those metaphors, the metaphors start to shape their thought. Another possibility is that metaphors for structuring thought and metaphors for communication co-develop with a complicated interplay.

Regardless of the dynamics, the result in any of these cases would be that the same metaphors are key in both communication and in internal cognition. While not logically necessary, this coherence between cognitive and verbal metaphors is the main point argued for by Lakoff and Johnson [2008], and here we take as a hypothesis.

3.7.2 The resource framework

Our fundamental theoretical framework for thinking about cognition is the resource framework [Hammer et al., 2005], a way of thinking about physics reasoning as "knowledge in pieces" [Disessa, 1988] or a "society of mind" [Minsky, 1991]. Human knowledge is built out of many different pieces of understanding, and figuring out complicated problems is often more about coordinating all the bits of knowledge we have and applying them appropriately than it is about simply gaining knowledge.

For example, most physics instructors know that before (and often after) instruction in an introductory mechanics class, many students will get basic questions about Newton's third law wrong. If asked about a small car pushing on a truck to get it up to speed, they'll say the car exerts more force on the truck than the truck does on the car.

But what are students really getting wrong? There's a lot of things they could say about the car pushing the truck:

- if a car slams into a truck, the car comes out worse and takes more damage
- same thing if the truck slams into the car; car still comes out worse
- the truck has a more powerful engine than the car
- it's harder to move the truck than it is to move the car
- the truck takes more gas to drive a certain distance
- if you slam into a concrete wall with the truck at 40 mph, youll do more

damage than if you slam into the wall with the car at 40 mph

Researchers such as Elby [2001] have pointed out that there's very little that the students are actually wrong about; in fact they understand the underlying situation pretty well. It might even seem the only thing they're getting wrong is the word "force", and not any physical phenomenon at all. Why would students who understand that physical situation still have so much trouble answering physics questions about it correctly?

Looking at a successful instructional approach might help. Elby [2001] describes an effective approach to car-truck interaction problems in a 2001 paper. The approach asks students to call up their raw intuition, and points out that they do know a lot about what's going on. This might get mapped to something like "The car reacts twice as much during a collision with the heavier truck". Next, students consider refining this intuition into more physically precise statements such as "the car feels twice as much force during the collision" and "the car has twice as much acceleration during the collision". By mapping out the consequences of these statements and holding them up against intuitions, students improve dramatically in thinking about precise physical concepts.

In this sort of approach, the instructor wasn't trying to teach students something completely new or from scratch. The goal was to take individual bits of student knowledge and refine and restructure the way students interpreted, called on, and related those bits to each other. This exemplifies the resource framework's viewpoint. It's a constructivist [Council et al., 2000] view in which students build new understanding for themselves, guided by the instructional and peer environment. They do this by reorganizing their little bits of knowledge to more-useful structures for thinking about physics.

3.7.3 Ontological resources

The resources I wrote about in section 3.7.2 were resources for thinking about a specific physical scenario, but the society of mind view holds that all of cognition is structured in a similar way. That means that the resource framework can help us understand other types of cognition besides that of "what is the force here?"

For example, Hammer and Elby [2009] describe "epistemological resources" - resources about how we can know things. Students don't have a single, monolithic epistemology for physics knowledge, such as "the textbook contains everything we need" or "you need to derive everything from first principles". Instead, in a given scenario they'll draw on all sorts of different ways to gain knowledge, each its own epistemological resource. An epistemologically-sophisticated learner is one who organizes and activates epistemological resources productively given what they're working on.

We take the view that ontological metaphors can be studied in the same way. In the extended quote from Kip Thorne in section 3.6.2, we saw a description of using different ontological resources to think about space, time, and spacetime. We might view the future as already existing or not, for example. When viewing it as existing, times in the future are metaphorically mapped to positions in space, specifically vertical position, leading to Thorne pointing to a spot in the air saying "the future is here", then pointing to a spot below it, "and the past is there." But the metaphors are fluid; he switches between them, and their goal is to be functional and help solve problems.

Gupta et al. [2010b] argued that the same is true and valuable for students. Instead of it being best to use a single, coherent ontological metaphor in all circumstances, they find different types of metaphors, and indeed different pieces of different metaphors, constructive because each is a resource which affords a certain type of thinking. Gupta et. al. focus on "substance" vs "process" metaphors for thinking about all sorts of physical quantities, for example blood flow and the flow of electrons, chemical equilibrium, heat transfer, force, etc. They show that people can construct novel categories of things on the spot depending on context, that the metaphors we use combine pieces of various ontological categories, and that we can switch between metaphors.

This viewpoint strongly informs what we look for in studying the ontological metaphors of equations. We don't expect students to have a single metaphor for what an equation is. Instead, we expect various ontological resources to be activated depending on what type of action the student is performing or how they are framing the mathematical activity. This suggests moving to smaller grain sizes as we study and code our data.

3.7.4 Epistemic games

We reviewed epistemic games in chapter 2. In brief, people need some way of organizing the resources they use. They tend to frame situations, asking, "What kind of activity is going on here?" Making that decision dramatically narrows the search space of resources are appropriate at what time. We refer to framing a physics problem-solving activity, calling on a picture of what sorts of things are likely or allowed to happen while solving the problem, as playing an epistemic game.

In chapter 2, we introduced several specific epistemic games associated with problem-solving strategies:

- examining extreme of special cases
- dimensional analysis
- estimation

We also introduced the concept of ontological metaphors for equations, and pointed out that students, like experts, use a variety of ontological metaphors for equations.

Now, we consider whether equation ontologies and epistemic games are related.

3.7.5 Fitting equation ontologies into a bigger picture of mathematical sensemaking

In chapter 2, we introduced epistemic games as important to how students combine physical intuition and mathematics, especially in the interpretation of algebraic expressions, during problem solving. In chapter 4 we will describe the development of a concept inventory to measure this sort of sensemaking in a large class environment, and question the interpretation of the test's results in chapter 5. Here, I want to make the case that studying equation ontologies is aligned with the broader research goal of learning about mathematical sensemaking in IPLS students.

Students working with equations need much more than a set of rules for legal moves in manipulating equations. Arcavi [1994] cataloged a large set of instincts, intuitions, and similar behaviors that are missing in someone whose only algebraic knowledge is a perfect understanding of algebra's rules. This includes things like realizing when switching to a different representation (e.g. a geometric picture of a circle instead of an equation $(x-a)^2+(y-b)^2=R^2$) would be productive, or recognizing when only the largest-power or smallest-power terms of a polynomial will matter for a given type of argument.

Moving to physics, students still need this symbol sense to guide their work, but it can be enhanced further by a physics sense - an understanding of what sorts of algebraic forms of expressions are likely to yield physical insights, what types of thought experiments might tell us things that can be fed back into limiting cases of equations, etc.

All of these high-order types of cognition surrounding equations are very abstract, but they often revolve around doing certain things with equations - making manipulations, rewriting in a new way, mapping individual terms to physical meanings, etc. Conceptual metaphor is what grounds the abstract processes. We suspect that ontological metaphors, and ontological resources associated with them, are important because the metaphors help students to recognize and carry out productive actions on the equations.

For example, if we want to take the limit of an equation with a mass as a parameter, as Bert did in chapter 2, we are implicitly thinking of equations are mutable. We gave several other examples of students assigning equations, and the mathematical objects they relate to, different ontological statuses in that chapter. We also related them to what the student was accomplishing with the equation at the time.

To study this more systematically, we ask about the relationship between equation ontologies and epistemic games. Epistemic games are interesting because while playing one, certain resources, especially epistemic resources, are more likely to be called on. There are two reasons we can see that some ontological resources might be more closely associated with certain epistemic games.

First, the types of mathematical actions performed may depend on the epistemic game being played, and different ontological metaphors afford different mathematical actions.

Second, students playing a particular game may envision their relationship to the equation differently depending on which game they're playing. For example, in a game where students are testing whether a given equation makes sense, they may be more likely to treat equations as immutable objects than they would in a game where they are trying to determine what equation would match their intuition for how a given system should behave.

One reason that studying ontological metaphors may have value is that they're significantly more specific, and work on a smaller grain size, than epistemic games. It may be difficult to structure an instructional intervention around playing an epistemic game. Asking students specifically play a certain epistemic game, with a series of steps, would likely be counterproductive. It would encourage following the steps of an e-game for an unknown purpose, and would be unlikely to create the same cognitive framing and activate resources in the same way as an e-game students played organically. Instructors could model playing an e-game themselves, but catching onto a broad re-framing requires a fairly holistic view of the activity. By contrast, instructors or writers could intentionally include appropriate ontological metaphors, with an expectation that students adopting those metaphors would bring new resources to bear on a problem. Instructors can only do that they have confidence in how the ontological metaphors are connected to resources used, and connecting ontological metaphors to epistemic games is one example of that.

3.8 Study design

In this chapter, I report on analyzing the data from problem-solving interviews described in chapter 2. I look at the ontological metaphors students used to talk about equations. I also look at what epistemic games students played. Finally, I look for patterns in what ontological metaphors are used most when when playing different epistemic games.

3.8.1 Why we chose semi-structured, think-aloud interviews

When conducting interviews, we weren't initially narrowed down to a research question on ontologies of equations, but we knew we were interested in how students adapt their mathematical knowledge to solve physics problems. We also wanted to know what challenges they face when trying to do that.

We chose our interviews to be semi-structured so that the interviewer could pursue any bit of mathematical reasoning by the student or group that seemed especially interesting. By "semi-structured", we mean that the interviewer had a script with a problem to ask. Before beginning the interviews, we also discussed what sorts of questions the interviewer would ask. Still, the interviewer was allowed to stray from the script to ask clarifying questions or modify a problem slightly. Also, if students went on a tangent that still seemed potentially informative about mathematical sensemaking, we were happy to explore whatever avenue they were heading down.

In practice, the interviewer intervened and improvised to a varying amount in interviews; this is reflected in the detail of the interview protocols show in appendix A.

Additionally, we sometimes planned for the interviewer to suggest a particular epistemic game intervention, especially suggesting looking at extreme or special cases of a formula. We didn't expect this behavior to arise spontaneously during problem solving, but we wanted to study it. Having the interviewer intervene rather than writing it as a prompt allowed us to present the technique as a suggestion to a pre-existing problem the student had been working on, rather than as “one more section” of a multi-part problem. Considering the “script-like” use of checks when prompted by writing described by Sikorski et al. [2017], we thought interviewer prompts would promote richer epistemological responses from the students than written prompts.

3.8.2 Both individual and group interviews

We present results from both individual and group interviews. Both types of interview had advantages for our study.

Individual interviews allowed the interviewer to watch what the student was doing in real time closely and interact with the student easily at any given time. The interviewer sat next to the student, and the student anticipated holding a conversation with the interviewer during the interview. This let the interviewer find areas where they judged that the student could explain in more detail and ask pointed questions without too much interruption to the flow of the interview.

Group interviews were also important. During group interviews, three or four students, generally not previously acquainted, would solve problems together. The interviewer sat farther away from the group, not at the table, and the students generally ignored the interviewer except for occasional interventions, usually 1 - 3 per hour-long interview.

In the past, group interviews have yielded some of the most-insightful results on how epistemic games are played, especially in moments where students playing different epistemic games try to collaborate on the same problem. For example, Bing and Redish [2009a] found they could generate insights into the differing epistemic games students were playing by analyzing the "warrants" they used when discussing their solution methods. These warrants, a tool from Toulmin [2003]'s analysis of arguments, make a connection between a claim and the evidence for the claim. That is, they spell out why a fact you've stated should be considered evidence in favor of a position you support.

When arguing or negotiating with other students on a problem, students may be explicit about these warrants, or they may be inferred from the context, but either way, the types of warrants they used were key indicators of their epistemic framing of the situation. For example, Bing and Redish describe a pair of students discussing the meaning of the work-energy theorem applied to a problem about moving an object along one of two possible paths. They document the students attempting to play different epistemic games - one attempting to play a "mapping symbols onto meaning" game in which mathematical expressions are expected to translate directly back into statements about physical experience, and the other attempting to play a "match to authoritative source" game in which it's expected to invoke known rules to answer a question. The students' clashing frames lead them to find a third epistemic game - one of making explicit mathematical calculations and interpreting their meaning.

Each time, it was by finding and highlighting the warrants behind the students' statements that Bing and Redish could identify these epistemic games and build their understanding of the problem-solving session. Without having two students trying to communicate to each other and reach agreement, it's doubtful they would have been able to pick out the warrants and gain the same sorts of insights into the students' epistemologies. We included group interviews based on this and other research's findings that group discussion elicits clarity in student epistemologies (e.g. [Hammer, 1995, Hogan, 1999]).

3.8.3 Problem design and interview protocol

We used a variety of problems throughout our interviews. For a complete copy of all the interview protocols, please see appendix A.

3.8.3.1 Ellipse and half-Atwood machine

We wrote this interview protocol with a goal of writing problems that were best-solved by looking at extreme cases, without explicitly asking students to look at extreme cases. We wanted extreme cases to be a useful tool for the problems, but only to come in as a useful tool after the difficulty of solving the problem itself was made apparent to the students by working on the problem for a while without the limiting cases tool.

The ellipse problem 2.2 is written to have two compelling answers until the limiting case of a very flat, thin ellipse is examined.

The half-Atwood machine problem 2.1 is a very standard physics problem; it has no special physical context, is somewhat removed from everyday experience due to the lack of friction, and generally would be expected to elicit the most straightforward, textbook-style problem solving in students. This would allow for for a larger contrast in epistemic-game playing if examining the extreme cases does encourage students into epistemic frames strongly-associated with mathematical sensemaking. Seeing rich student thinking on this problem would be evidence that this sort of rich problem-solving behavior can really happen in any physics classroom, since the problem is so traditional.

3.8.3.2 Terminal velocity

This interview examines the terminal velocity of the bathysphere, an early submersible, falling through the ocean. Students are first asked to build a complicated mathematical model for terminal velocity involving both the viscous resistance (proportional to v) and the drag force (proportional to v^2) on a sphere. This leads to applying the quadratic formula to solve a complicated equation with many different terms. We expected students to approach this task very formally, since the first part of the prompt leaves little room for interpretation or physical insight and asks for explicit mathematical calculation.

We were then interested in how switching to an estimation epistemic game would affect the students' problem-solving. The next part of the problem gave a picture of the bathysphere with only partial information on its physical makeup - students would have to use the picture (which features two men standing next to the bathysphere) to estimate parameters. Then, they would have the option of using their very long, complicated formula from earlier in the problem, or realizing that some terms in their formula were unnecessary and making a much simpler estimate while sacrificing very little accuracy.

Finally, we asked about falling volcanic ash, which is very low Reynolds number (compare to the bathysphere, which is high Reynolds number), to see if this dramatic switch might prompt extreme-case reasoning.

3.8.3.3 The ring of charge and systems of springs

For group interviews, we wrote two very challenging problems which could be helped by both dimensional analysis and extreme case reasoning.

Like the ellipse and half-Atwood interview, this interview consists of two problems which physically and mathematically bear little relation to each other, except that there is some possibility of "far transfer", in that techniques of dimensional analysis and examining limiting cases are potentially useful in both problems, so if the interviewer were to suggest them for one problem, there's a possibility that the students would adopt them for the other problem as well.

The ring of charge problem (see A.2) is, like the half-Atwood problem, a fairly-standard textbook problem which is somewhat beyond the sophistication of what PHYS131/132 students are usually expected to handle. Solving it generally requires drawing a useful diagram of a three-dimensional setup, doing some trigonometry and applying the Pythagorean theorem, breaking vectors into components, and recognizing symmetry and its consequences. It also requires understanding the idea of a fixed amount of charge spread around in a ring. Students need an intuitive understanding of integrating over that ring of charge and need to rely on the superposition principle. All this is to say that the analytical aspects of the problem are very challenging, so that examining extreme case might be especially helpful.

The problem is difficult enough that we didn't expect groups to solve it, but we were interested in their intuitive graphs of what they expected the behavior to be like, especially whether they would recognize that the special case of the center of the ring, and the extreme case of far from the ring, can be analyzed intuitively without calculation.

The system of springs (see problem A.2) asks for the effective spring constant of two springs attached to a box in parallel (next to each other) and of two springs attached to a box in series. We expected these problems to call up symbolic forms. The problem of springs in series is still quite challenging. We expected that students could correctly interpret the physical behavior of the system with the limit of very stiff or very loose springs, and wanted to see how easily those physical interpretations could inform their mathematical work on this challenging problem.

3.8.4 The interview environment

We recruited students by making a brief announcement about our research project during the PHYS131/132 recitation periods and passing around a sign-up sheet. We then used the sheets to recruit students throughout the course of the study, paying a moderate compensation for their time.

We interviewed students in a small office in the physics building at the University of Maryland. Students were not acquainted with the interviewer on their first interview, but some returned for several interviews (including individual and group interviews, as well as validation interviews for the MEGS, as discussed in chapter 4).

For individual interview, the interviewer sat either side-by-side with the students (Eichenlaub) or across from the student (Hemingway) at a table so as to see what the student was writing and pointing at while talking aloud. We read the student a short introduction to the interview before beginning, explaining think-aloud interviews.

We recorded the interviews with a wide-angle, high-resolution sports-style camera, generally above the interviewer and student, looking down at the table. The camera usually captured the faces and upper bodies of the student and interviewer, as well as a whiteboard which we placed in front of the student on the table.

After several interviews, we realized the camera didn't always catch everything being written, and replaced the single whiteboard with a small stack of whiteboards; the student would set one aside after filling it up, and we photographed it after the interview.

Our goal was to collect as many sources of evidence on student cognition as possible - gesture, speech, and written artifacts. However, in this analysis, we rely primarily on speech to examine the ontological metaphors students use. We relied on gesture in certain places analyzing these same interviews in chapter 2, where it was useful especially when students pointed to particular terms within an equation they wrote. However, we found that it was more practicable to construct a coding scheme relying on linguistic cues.

3.8.5 Early analysis, selection of equation ontologies as a research focus

In doing the analysis presented in chapter 2, we noticed a variety of "rich moments" in the interviews. These were moments when, before doing any careful or systematic analysis, we nonetheless noticed conceptually or epistemologically-unexpected statements from the students. These are the moments that would catch our attention while rewatching the data. For that analysis we organized the rich moments into a theme of diverse and often expert-like behaviors that students would try out as problem-solving techniques. Students often made interesting insights while failing to reach a correct final answer to the problem, or failed to have their group recognize that answer as correct.

We decided to focus this follow-up study on equation ontologies in relation to epistemic games.

3.9 Research question

The question we will attempt to answer, within our data set, is:

> Do students systematically tend to use some equation ontologies more frequently when playing sensemaking-focused epistemic games, and other equation ontologies more frequently when playing other epistemic games

such as “plug-and-chug” or “formal manipulations”?

To answer this question, we need to

- recognize ontological metaphors for equations
- categorize them appropriately into types of ontologies
- develop a catalog of epistemic games students play
- determine what epistemic game students are playing at different junctures
- record the instances of ontologies used in each epistemic game
- analyze the resulting data to find any patterns

3.10 Data analysis

In this section I'll describe how we decided to use coding to analyze data from problem-solving interviews to answer the research question.

3.10.1 Focus on linguistic evidence

My primary source on ontological metaphors is Lakoff and Johnson [2008], which is based on primarily linguistic evidence to uncover metaphors. Lakoff and Johnson use examples of everyday speech as their data, extracting the metaphorical structure from the words themselves. For example, they use the existence and coherence of phrases like “I *demolished* his argument”, “I've never *won* an argument”, and “He *attacked* every weak point in my argument” to give evidence that we use war or fighting as a metaphor for argument, and that this structure the way we think about argument and how we act in an argument.

Using linguistic evidence to establish cognitive metaphors does require a sort of leap of faith. The only evidence being presented in the previous example is that we use the same words to talk about war and arguments; that in itself isn't proof that we use our cognitive resources built to understand war for the alternative purpose of understanding arguments.

Lakoff and Johnson rest their case on three main points. The first is that the connections they are making are intuitively plausible. The second is that they are generally part of a larger, coherent structure. For example, throughout most cognitive metaphor, we can find the metaphor "up is good". Some examples based on those of Lakoff and Johnson might be "Things are looking up", "I'm at the top of the rankings", "Ali at his peak", "She raised her status", etc. If we were simply substituting words from one domain into another in a meaningless form of overloading, it would be surprising to see the extent of coherence in metaphors that we do when examining linguistic evidence.

Finally, Lakoff and Johnson connect their cognitive metaphors to embodied cognition, the idea that we build abstract concepts up from simpler ones, ultimately reducing to the way that we understand the physical world and how our bodies interact with it. So, up is good, according to Lakoff and Johnson, because if you're healthy and strong, you're more likely to stand up straight. A more straightforward example might be "waking up" (as opposed to being "put down") which uses the metaphor "conscious is up". We assign consciousness to up and unconsciousness to down because conscious people stand up erect whereas unconscious people fall to the ground.

This embodiment hypothesis suggests that gesture may also be important in understanding conceptual metaphor. Gesture has an important role in PER research because students, especially early on and especially when sensemaking, may have productive and correct ideas that they can't yet fluidly articulate in the language of physics, a language they're still learning. But we might see these ideas, and students may partially communicate them to each other, in gesture [Scherr, 2008].

Gesture has even played an important role in the study of ontological metaphors in PER. For example, Dreyfus et al. [2015] used gesture as a key part of their evidence for how students and physicists combine two different metaphors for energy (a position-based metaphor and a fluid-based metaphor) in a process of "conceptual blending". Gesture was also important in Wittmann et al. [2013]'s analysis of separation of variables, discussed in section 3.6.

Analyzing gesture in a coding process, though, presents challenges. One of the prerequisites of effective coding is defining the codes specifically and unambiguously enough to achieve high reliability when different people (who have trained to do the coding) code the same data. Gestures may be more difficult to categorize in this way, as people spontaneously invent new gestures as they speak. The gestures often require significant context to understand, as any game of charades demonstrates. The same is true of language, but to a lesser extent, as individual sentences are often understandable when taken out of context.

To investigate further what advantages might be lost by ignoring gesture when searching for ontological metaphors, I asked the PERG group to join me in analyzing some data from a problem solving interview for ontological metaphors using a transcript of the data, then again while watching the video source of the same transcript. PERG group members often felt that they understood the data better when watching the video, but didn't point out any cases where a conceptual metaphor was clearly present in the student's gesture, but not in speech. The exception was indexical gesture - e.g. pointing to or underlying an equation while saying "this" or "that equation", etc.

Analyzing several further video clips in the same way independently, I came to the same conclusion. So for analysis of this project, I decided to transcribe words and only those gesture necessary to understand what the student was referencing in their speech. I then coded the transcripts.

3.10.2 Quantity of data to analyze

I decided to analyze all the data available from the problem-solving interviews that place a strong emphasis on the "special or extreme cases" epistemic game. This amounted to ten interviews between 45 minutes and one hour long. These were seven individual interviews and three group interviews.

3.10.3 Identifying epistemic games

To identify epistemic games, I used the codebook developed by Tuminaro [2004]. This includes the e-games

- mapping meaning to mathematics
- mapping mathematics to meaning
- physical mechanism
- pictorial analysis
- recursive plug-and-chug
- transliteration to mathematics

The "physical mechanism" game and "pictorial analysis" game do not include any mathematical expressions, by definition. When students play those games, I left them uncoded because they could not correlate to ontological metaphors for equations. I found no instances of the "transliteration to mathematics" game, likely because students didn't have outside materials available during the interviews.

To Tuminaro's e-games, I've added

- extreme/special cases
- dimensional analysis

Please see appendix B for a description of these e-games in the style of Tuminaro. (I have added the word "from" to the meaning-mapping e-games of Tuminaro because I saw their titles as grammatically ambiguous.)

This makes my list of epistemic games

- mapping from meaning to mathematics

- mapping from mathematics to meaning
- recursive plug-and-chug
- extreme / special cases
- dimensional analysis

3.10.4 Descriptions of epistemic games

Because I am not producing an entire codebook for epistemic games, I will briefly summarize the e-games I am using from Tuminaro's codebook [Tuminaro, 2004].

3.10.4.1 Mapping from meaning to mathematics

In the "mapping from meaning to mathematics" game, students start with a story or set of ideas about what happens physically in a situation. For example, they might know that a gravitational force and a wind resistance force "balance" when an object falls at terminal velocity. Then they turn this physical understanding into a mathematical expression, for example, $F_{gravity} = F_{wind}$

Here is a short example of Lelia, Lizze, and Myra (does not speak) playing this e-game while solving the springs in series and parallel problem, described in appendix A. In this problem, they are asked to find the effective spring constant of two springs with spring constants k_1 and k_2 both when the springs are in series and when they are in parallel.

Lelia's first sentence in the third speaking turn refers to details of the instructions, which specified that the effective spring constant k_{series} should be in terms of

k_1 and k_2 when asking about the springs in series, but didn't explicitly ask for the solution for $k_{parallel}$ in terms of k_1 and k_2 when asking about the springs in parallel.

1. Lelia: We said that basically [in the series case] the k value would be smaller because this one would be easier to stretch, so dividing by something would make sense.
2. Lizzie: and the parallel one would be two times k one plus k two
3. Lelia: We don't have to use k one and k two for that one. It doesn't say. So k one plus k two divided by four?
4. Lizzie: That's the same thing as k divided by two.
5. Lelia: I'm putting that down. Our final answer.

In this passage, Lelia starts with the physical understanding that two springs in series will stretch more than a single spring, and so the formula for k_{series} should show that it is smaller than the k values for an individual spring. They use this requirement to determine that they should add a denominator to their answer with a value large enough to make the effective spring constant small. Similarly, Lizzie's suggestion that $k_{parallel} = 2(k_1 + k_2)$ stems from earlier discussion that the springs in parallel would be hard to stretch, so the spring constant should be large. The factors of two and four that they choose come partially from their counting the springs in the problem (at one point they suggest dividing by "n", the number of springs), and from a similar problem they remembered from homework.

The "mapping from meaning to mathematics" epistemic game is played commonly when students are asked to find an unknown expression. I had an expectation that this e-game might be associated with the “equation as mutable - agentive” ontology, because during this process, students are creating an equation. Their epistemic stance is that equations are made by assembling various pieces of physical reasoning, and they are filling out an epistemic form of an unfinished equation. As they fill out the epistemic form, the equation they are building changes in front of them, becoming increasingly complete, so if equation ontologies are closely associated with epistemic frames and my above account is largely correct, I might see many instances of students talking about equations as if they personally change the equations.

3.10.4.2 Mapping from mathematics to meaning

In this epistemic game, students have a mathematical expression at hand, and are using the expression to understand something more about the physical scenario.

Here is a short example of Dorothy in the midst of playing this e-game while working on the half-Atwood problem shown in figure 2.1. In this setting, she is trying to determine the acceleration of the blocks, assuming that they are stationary. The text transcribes Dorothy’s words, while the parentheses show my interpretation of the equations she was referencing.

> Acceleration is equal to v over t ($\frac{v}{t}$), so really if you wanted to find out its acceleration, you would have to record some times, and then find out its v f minus v i ($v_f - v_i$) over its delta t (Δt; $\frac{v_f - v_i}{\Delta t}$). That’s only way that you can go about like looking like, if you’re doing a test. But

> you can't say, oh, I'm looking at a block and I know it's ten newtons per kilogram. Because you're not measuring anything, and you have no Newtons available.

Dorothy is working towards building her understanding of the physical situation of stationary blocks, and does it starting from the equation $a = \frac{\Delta v}{\Delta t}$. She rejects the idea that it can be done without a mathematical analysis, instead claiming that only by measuring certain variables can the acceleration be found.

For an example of how I coded equation ontologies, I coded two utterances in this passage with equation ontology tags before coding it as the mapping from mathematics to meaning e-game. These were "acceleration is equal to v over t", coded as "equation as immutable", because it takes the stance that this is "the" equation for acceleration, implying it is the only equation for acceleration. I coded "find out its v f minus v i over its delta t" as "equation as parts" because it acknowledges that each term in an equation corresponds to a separate measurement. The expressions $v_f - v_i$ and Δt are linguistically treated as separate entities.

This epistemic game involves making sense of mathematics - taking individual expressions and turning them into statements about the physical world. I had some expectation that "equation as parts" would be used frequently when playing this e-game. In order to make sense of an equation, students often look at symbols or groups of symbols individually, because it's often individual symbols that correspond to specific quantities in physical world. I also thought that grouping might be a common ontology in this e-game, because we might, for example, have a group of symbols all of which determine one force (e.g. mg is two symbols which determine the gravitational force on a mass m in a gravitational field g), so those symbols might be grouped together frequently when interpreting the equation.

I also thought this epistemic game might be associated with the ontological metaphor "equation as one form of a relationship" because in order to read meaning off from equations, students might have to solve them for different variables, cancel terms, expand expressions out, etc. to get a more readable form. In doing so, they might have a greater tendency to speak about equations as if the same equation can take different forms. The concept of the same equation having different forms might be a prerequisite for the epistemic stance that finding appropriate forms of a given equation is a good way to generate new knowledge from it.

3.10.4.3 Recursive plug and chug

I only rarely observed this e-game. In it, students realize there's a particular quantity they want to find: acceleration, pressure, electric field, etc. Then they find an equation for that quantity, and attempt to plug numbers into the equation as a complete solution to the problem.

For example, Jean played this game when working on the terminal velocity task described in appendix A.

> Jean: So, I don't know if we're right, but I was thinking, since we know the acceleration is ten, and then if we take the integral of that we get the velocity, which is ten t, you take the integral of the velocity one

> more time and you get the equation for the position, which is position equals five t-squared, and then I plugged in, [looking at Zane] we said a thousand meters, for the depth [Zane nods], and then like plugged it in to five t-squared and solved for t. I got that it takes about fourteen seconds for the thing to fall, which doesn't really make sense because I feel like that's pretty short.

As is typical for the plug-and-chug game, Jean doesn't use any sort of physical story behind her calculation, she simply takes some given values (in this case, the incorrect value that the acceleration of an object falling through a fluid is about $10m/s^2$) and uses some equations (e.g. $v(t) = \int_0^t a(t)\mathrm{d}t$) to determine an answer.

What's not completely typical is that Jean stops to reflect on the sensibility of the answer.

Zane challenged the physical assumptions behind this calculation ("Cause it's not [gestures with a large up-and-down hand motion] um, you know, it's not x squared falling because it's, it doesn't, it wouldn't accelerate as it goes down.") but Jean didn't acknowledge this sort of argument as epistemologically valid, replying, "but that is how I did it on the homework".

This illustrates the effectiveness of e-games as an analytical tool. Jean's response sounds like a non-sequitur until we realize that her epistemic frame is different from Zane's. Zane's physical story falls outside the boundaries of the moves allowed in the plug-and-chug game. His description of the physical mechanism is not entirely clear, and it's likely that the only point Jean got from it was that Zane was challenging her solution. Needing to back up her solution, but not playing an e-game that allowed for physical storytelling, she brought in the sort of evidence that is relevant in recursive plug-and-chug: that this solution method worked on another problem. It wouldn't be until Jean played a different e-game that she would be likely to engage Zane's point directly.

I expected the "equation as whole" ontological metaphor to play an important part in the recursive plug-and-chug e-game because the player is looking for "the equation for X", with X being some particular phenomena.

3.10.5 Coding order and details

In hopes of avoiding choosing equation ontology tags based on epistemic games I had already identified, I did all equation ontology tagging first, then separately coded for epistemic games before putting the two together.

I transcribed the interviews in Otranscribe, then coped the transcripts to Dynalist. There, I used Dynalist's tagging system to place tags for each ontology code next to the corresponding utterance.

I used Dynalist's color highlighting to code for epistemic games, with a different color corresponding to each e-game. I did not code regions of transcript in which there were no ontology tags.

To count co-occurrences of tags and e-games, I used Dynalist's search feature to display all tags of a certain type, and manually counted the number of occurrences in each color.

3.10.6 Frameworks for identifying ontological metaphors

In PER, two methods of identifying student metaphors from speech are predicate analysis and grammatical metaphors [Gupta et al., 2014].

According to Gupta et al. [2014], predicate analysis was developed by Sommers [1969]. Predicates are the parts of speech where the subject does something, i.e. what the subject "has, does, or is". The idea behind predicate analysis is that there is a mapping between predicates used in speech and ontological categories used to think about objects.

For example, if we say that something "took an hour", the thing is a process of some kind, since we generally would not say that an object lasted an hour, as in "table took an hour". If someone did say "the table took an hour", they are probably referring to a process, such as the construction of the table, and using "the table" as a metaphor for that process.

By constructing a list of predicates and classifying them according to the ontological metaphor they go with, we can construct a codebook for ontological metaphors.

Brookes and Etkina [2009], in their analysis of the ontology of forces, used a different method of classifying ontologies based on grammatical metaphors. When a student makes some utterance in reference to a concept like force, they looked at whether it functioned as a noun or a verb grammatically. Nouns can then be assigned object-like ontological status and verbs process-like ontological status in the students' metaphorical systems.

I wanted to explore a broad variety of ontologies beyond the object/process divide often studied in PER, so I don't found grammatical metaphors an appropriate technique for my study of equation ontologies.

No dictionary of predicates for equation ontologies yet exists, but in future work, examining the predicates carefully may allow building such a dictionary to function as a codebook for equation ontologies. As yet, it remains unknown whether, when discussing equations, predicates and ontological metaphors have the sort of one-to-one mapping assumed by predicate analysis, so it remains a future challenge to determine whether predicate analysis can be applied here.

3.10.7 Unit of coding for analysis of equation ontologies

One reason I'm interested in equation ontologies is that they happen fast. Epistemic games are "sticky". When a student is playing one, even a TA directly strongly and repeatedly suggesting changing tactics to a different e-game can be completely ineffective [Tuminaro and Redish, 2007]. Students tend to stay in e-games for minutes at a time.

By contrast, a single mention of an equation only takes about one to three seconds, so the unit of coding for equation ontologies is one to several words, or one phrase.

Although the linguistic unit that indicates an equation ontology is very short, it's not usually possible to determine the code from a one-second snippet. I usually needed some information from the surrounding context in order to interpret the particular utterance I coded. In this sense, the unit of coding could be as long as about a minute. Usually, only about five to ten seconds of context was enough to understand what the utterance was referencing.

For example, I coded Alma's utterance "it was two different r's" as "equation / phenomena not differentiated". Without context, this utterance might be difficult to understand. Alma said while working on the ellipse problem (2.2). The surrounding speech was,

> Alma: the regular area of a circle is pi r squared. So that would be a circle with two radiuses like this, radii, I guess. So is pi times r times r, so it would have to be A equals pi times a times b, I guess.
>
> Interviewer: and how are you getting from one to the other?
>
> Alma: so I just drew the picture and pretended it was two different r's, so like, this was a [points to a horizontally-oriented radius of a circle she has drawn, and write label "a"] and this was b [same action for a vertically-oriented radius]. But they equal each other. It would be like if a equals b equals r, then the equation would be pi a b equals pi r squared.

This exchange takes 42 seconds. In this context, Alma has just discussed both a picture of an ellipse and the equation for a circle. She uses "r" to refer to a variable in an equation ("area of a circle is pi r squared") but also uses a and b to refer to radii drawn as line segments on a picture of a circle. Her use of "two different r's" may apply to the equation for an ellipse or to the axes of an ellipse in a picture.

I didn't use a code for every utterance that referenced an equation. When none of the codes clearly applied, I left that utterance uncoded.

3.10.8 Generating codes for equation ontologies

To generate codes for equation ontologies, I worked both independently and with the research group. My general plan followed the recommendations of DeCuir-Gunby et al. [2011]'s advice on developing a codebook.

I read through several transcripts creating codes very freely with no specific targets in mind, then looked at the large list of codes generated and worked to consolidate them to a few codes.

I also used our experience analyzing interviews in chapter 2 and teaching and TA experience to suggest possible codes. My advisor contributed in the same way during discussions of potential codes.

After writing definitions of these codes, I discussed them with my advisor and changed the definitions based on perceived ambiguities. I then and applied them to the transcripts, pulling examples as I went, and making minor revisions to the codebook's code definitions to match the data more closely.

I eventually had ten codes:

- equation as whole
- equation as parts
- equation as mutable - agentive

- equation as mutable - non-agentive
- equation as immutable
- grouping
- equation / phenomenon not differentiated
- symbol as parameter
- symbol as variable
- equation as one form of a relationship

The codes may overlap. For example, when Alma mentioned "so it's like you're turning into a circle" while working on the ellipse problem in chapter 2, she was talking about an equation as a whole, and the equation was mutable (agentive) because it there was a specific entity changing the equation. It was also an example of equation / phenomena not differentiated, because it was not possible to tell whether she was referring to the equation for an ellipse turning into the equation for a circle, the picture she had drawn of an ellipse turning into a picture of a circle, or some internal concept of an ellipse that she had abstracted out from those representations turning into an abstract concept of a circle. So in this case three different codes applied to the same utterance.

For definitions and examples of each code, please see appendix C.

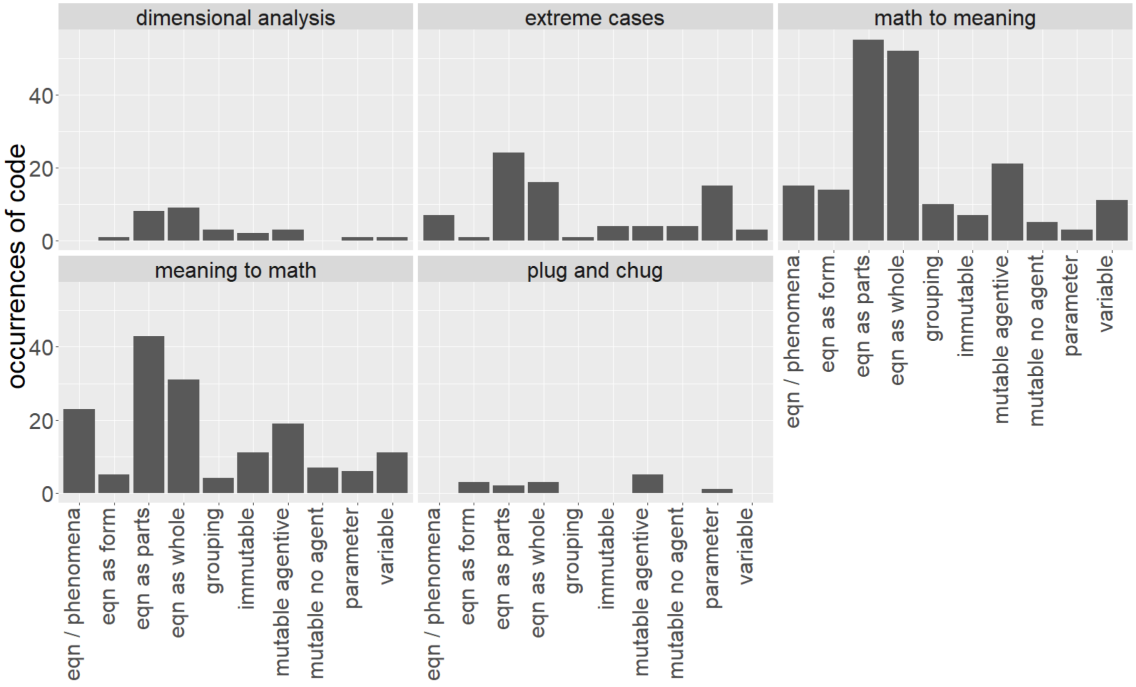


Figure 3.2: Occurrences of each ontological metaphor in each epistemic game.

3.11 Analysis and results

3.11.1 Ontological metaphors broken down by epistemic game

I counted the number of occurrences of each ontological metaphor in each epistemic game. The results are shown in figures 3.2 and 3.3.

3.11.2 Measuring significance of associations

Eying these plots shows that some tags are more closely associated with some epistemic games. For example, it seems clear that in this sample size and with my coding scheme, students spoke about symbols as parameters much more often when they were playing the extreme cases e-game.

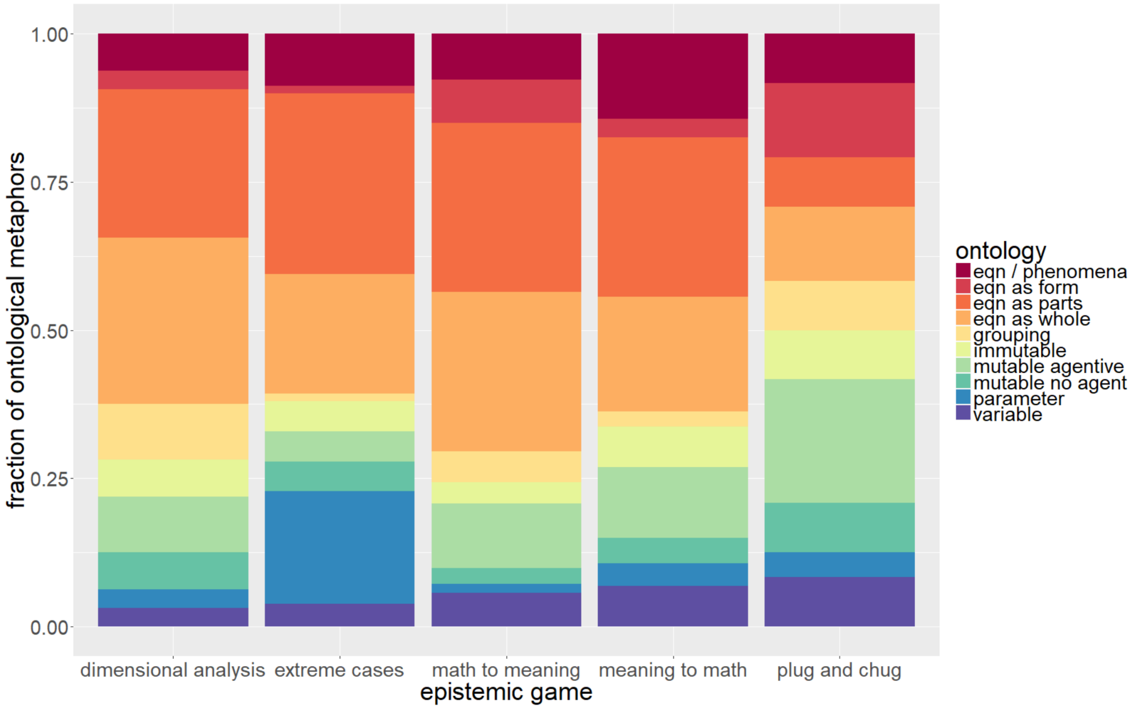


Figure 3.3: Fraction of each ontological metaphor counted in each epistemic game.

I would like to test whether these differences were likely to have arisen by chance, or whether they represent a real signal within this data set.

Here is my procedure for doing this. We can imagine a "base distribution" of how often ontological metaphors are used, which I take to be their distribution across all e-games. Then, looking at the number of metaphors used in each e-game in total, we can find the expected number of metaphors of each type in each e-game.

For example, there were 475 total ontological metaphors in my data set. "equation as a whole" accounted for 111 of them. There were 160 total metaphors in the "mapping from meaning to mathematics e-game". So the expected number of "equation as a whole" metaphors in the "mapping from meaning to mathematics" e-game is $160 \times \frac{111}{475} \approx 38$.

Before continuing, I'll make one refinement. I'll change the base distribution from what I described above. Instead of using the distribution of all ontological metaphors as the base distribution, I will use the distribution of all ontological metaphors in all the other e-games. So in this example, that's the distribution of ontological metaphors from all e-games except "mapping from meaning to mathematics".

The reason for this is that if a particular ontological metaphor is very common in a particular e-game, especially if it's a common e-game, that drags the base distribution away from where it should be and makes the bar for significance inaccurately high. This is a concern in my data set because "mapping from mathematics to meaning" comprised 40% of all ontological metaphors, so it could have a significant impact on the over all distribution, as could the "mapping from meaning to mathematics game", which was also common.

This sets an expectation for the number of metaphors of a certain type in a certain e-game. The real value will be different. For example, the expected number of "equation as whole" metaphors in the "mapping from meaning to mathematics" e-game is, under my refined procedure, about 41. The observed number is 31. Is this a significant result?

If I imagine as a null model ontological metaphors popping up at random an independently, their number would be Poisson-distributed, so the standard deviation would be the square root of the number of expected occurrences. In this case, the leads to a standard deviation of $\sqrt{40.6} \approx 6.4$, and the real answer is about 1.5 standard deviations from the expected number.

This is a bit wrong as a procedure for a few reasons. First, ontological metaphors are probably not independent. Metaphors are likely to be the same as the previous metaphor, even if they are equally likely in all e-games. This raises the standard deviation, so my estimate is a bit too low.

Next, reporting the number of standard deviations for a Poisson variable is somewhat of a bad statistic because the distribution is not symmetric, but this error is minimized if the mean is large.

I'm also not accounting for the error in the base distribution. Even if all ontological metaphors were equally likely in all e-games, I wouldn't get a distribution that matches the true base distribution, but would get some error on that. Including this uncertainty would decrease the significance of all my results.

If I were to add one "equation as whole" metaphor to the "mapping meaning to mathematics" e-game, I'd not only increase the count there from 31 to 32, I'd increase the total count in the the "mapping meaning to mathematics" e-game from 160 to 161, increasing the expected number of "equation as a whole" metaphors in that e-game. My statistical technique doesn't account for this, either.

Finally, the Poisson distribution I've used here may not be ideal. I used the base distribution to set a probability, p, for the fraction of occurrences of a particular metaphor type in general. Then I used the total number n, of metaphors in a particular e-game to estimate the number of instances of a particular metaphor in that e-game. A better distribution for this problem might be the binomial distribution, with the probability for k instances of the particular metaphor occurring being $\binom{n}{k}p^k(1-p)^{n-k}$. However, the binomial and Poisson distributions shouldn't be very different, as long as I have enough data to be likely to find significant signals.

A more-detailed method than the one I'm using would be to write a simulation that takes into account all the above effects to calculate p-values. However, this would ultimately just change the significance levels of my statistics a small amount, and I don't need to make fine-grained conclusions based on those numbers, so for this study, I will stick with the procedure I've already described.

Also, some research on coding uses hypothesis testing such as logistic regression to answer questions such as mine, and a Bayesian approach of assigning priors and updating based on my observations might be valuable as well, but I'm less familiar with how to use these approaches on the type of data I've collected, so I will leave their potential exploration to future work.

I applied my significance-testing procedure to the count of ontological metaphors in each e-game, finding a number of standard deviations for each metaphor-e-game pair, which I've visualized in figure 3.4.

Because there are 50 ontology-e-game pairings, I would expect some random deviations of between 2 and 3 standard deviations, but beyond 3 standard deviations would be unusual to show up, even in a sample of 50 drawings from a base distribution. This is again a fairly rough criterion, and the credence we give to an association being significant is a sliding scale, not a strict cutoff. Still, below I'll mention those associations that are stronger or close to three standard deviations.

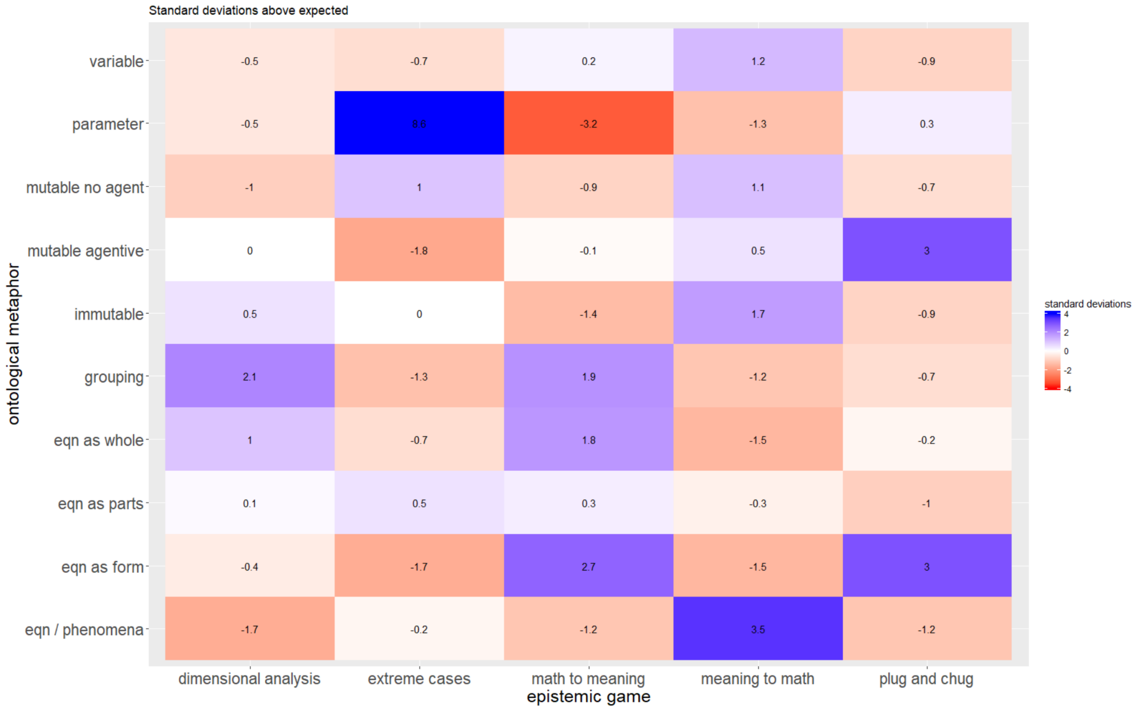


Figure 3.4: Number of standard deviations of ontology counts away from the expected number for each epistemic game.

3.11.3 Significant associations

Below, I detail each of the significant associations I found, and give a brief interpretation of each.

3.11.3.1 Extreme and special cases / symbol as parameter

The strongest association I found, with 8.6 standard deviations above the expected, is that students use the “symbol as parameter” ontological metaphor extensively when playing the “extreme or special cases” e-game.

The “symbol as parameter” ontological metaphor entails thinking of a parameter as something that specifies the general setup to a problem, rather than a having a different value at each specific point in space or time. For example, the mass of the boxes is a parameter in the half-Atwood problem, and the spring constants are parameters in the springs in parallel and springs in series problems.

When students use “symbol as parameter”, they are thinking about the system in a more abstract way than when simply try to understand the system’s behavior with a fixed set up parameters. In the half-Atwood machine, for example, it would be possible to set the system up and watch its evolution. That would let us see the change in variables, such as position of a block, happening continuously in front of us. When thinking about how changes to the mass of the block affect the system, students must instead think in some more abstract space where system acceleration is a function of the mass. When considering changing the value of the mass parameter, they are thinking about motion through this more abstract space.

(Also, from linguistic cues, they sometimes think about a series of individual trials, each with different values of the mass parameter, and make comparisons across these trials, another form of abstraction from simply watching a single run of the experiment.)

The reason for this association between the extreme cases e-game and symbols as parameters is fairly straightforward. While playing the extreme cases game, students imagine changing some symbol in their equation to an extreme. This is often done with a parameter.

The extreme cases e-game carries with it certain epistemological associations. For example, the e-game is usually used as a way of evaluating potential formulae for plausibility. Unlike directly solving a problem, it doesn't affirm a solution as correct, but instead simply lends it credence. However, the extreme cases game can be used in other ways, as when Bert used it to constrain the potential form of a solution to the half-Atwood problem (see chapter 2 for a discussion of this case).

That student utterances often fit the "symbol as parameter" tag while playing this e-game, and not nearly as much at other times, shows that it was mostly when taking a certain stance towards equations, one in which they varied parameters in an abstract space to test plausibility, that students spoke about symbols in a certain measurable way.

3.11.3.2 Mapping from meaning to math / equation and phenomena not differentiated

The "equation and phenomena not differentiated" equation ontology is when a student utterance could be taken to refer to an algebraic expression, a physical or geometric object, or both, suggesting that the student may not be differentiating equation and the phenomena they represent in their cognition.

This ontological stance is positively associated with playing the "mapping from meaning to math" e-game. In other words, when students begin with some sort of physical story about what's happening in a scenario, then turn that story into mathematical expressions, they often speak as if there is no clear difference between the expressions and the phenomena they represent.

While this association is plausible, I am surprised that the same association doesn't appear in "mapping from mathematics to meaning".

As an example of how the association between this tag and these e-games appears, here is Lizzie using this ontological metaphor while playing "mapping from mathematics to meaning"

> Myra: k over two is k series. Cause I think that makes sense because it would be like easier to pull if they're in series, and k is smaller. When k is small that just means that it's easier to pull I think.
>
> Lizzie: yeah, it is. Higher k is harder to pull. [boxes an expression $2k$ on her board].

The utterance is “Higher k is harder to pull.” One cannot physically pull on the symbol k, but “higher k” is nonetheless a reference to the symbol. It appears as if the symbol is standing in for the physical stiffness of the spring in Lizzie’s utterance, as opposed to “higher k means the spring is harder to pull”, which differentiates them.

This example is the “mapping from math to meaning” e-game because Myra and Lizzie were starting with an expression $k_{series} = k/2$, which they had hypothesized earlier in the interview, and evaluating it for physical sensibility.

An example from the “mapping meaning to mathematics” e-game comes from Christopher:

> Christopher: Tension force pulling it that way so it has to move that way at a certain speed of tension with the force. This has 9.8. I don’t think that matters, right now. Gravity. Gravity acting on you. Also has a tension force pulling it up. Is it right to say that this tension force has to be equal to this tension force?
>
> Interviewer: yeah
>
> Christopher: Okay. So to pull we need to pull m g.

The utterance I tagged is “we need to pull m g”, a reference to pulling, a physical action, on mg, a symbol that stands in for a force. The character of this utterance is almost exactly the same as that of Lizzie in the previous example.

I don’t have strong hypotheses on why the tag is associated only with one of these e-games and not the other, since their epistemological frame is very similar.

3.11.3.3 Mapping from math to meaning / symbol as parameter

I've already discussed both the "mapping from math to meaning" e-game and the "symbol as parameter" ontological metaphor. These are negatively associated in my data set. "Mapping from meaning to math" has a much smaller negative association with parameters.

My interpretation of this anti-association is that when students are mapping from math to meaning, the are focused on the object-level, direct behavior of the system, not thinking about its more abstract parameter space. This suggests that the extreme cases e-game is valuable because it promotes a type of thinking about systems and equations that students rarely use, even when using mathematical sense-making epistemic frames.

3.11.3.4 Mapping from math to meaning / equation as one form of a relationship

When students play the "mapping from mathematics to meaning game", they often speak about equations as if the symbols on the page are only one representation of the equation. Moving the symbols around by solving for a new variable, expanding an expression, etc. results in the same equation expressed in a different form, in this ontological metaphor.

This association shows that when students take an epistemological stance in which they believe they ought to be able to take an equation and determine its implications for a physical scenario, they also realize that this is often best done by changing how the equation looks to be easier to interpret (e.g. by solving for the variable they're interested in).

Christopher exemplified this association while working on the half-Atwood problem:

> so tension of mass large m is equal to gravity plus little m. And we care about acceleration. Gravity is equal to the tension of little m. Well tensions, mass. Big m minus tension little m. Can we common denominator that?

Christopher is beginning with an expression he's previously derived from first principles for the relationship between various forces in the problem. He wants to turn that equation into some physical understanding of why the system accelerates at different rates for different parameters, and to do that, he recognizes that different forms of his equation will help. Hence "we care about acceleration" shows his interest in understanding a particular variable, while "Can we common denominator that", the utterance I tagged with the "equation as form of a relationship" code, shows that he anticipates turning the form into a new equation to help him get insights. (This is also an example of "equation as mutable".)

3.11.3.5 Dimensional analysis / grouping

The dimensional analysis / grouping association is weak, only 2.1 standard deviations, but students often refer to the dimensions of a group of variables taken together, so dimensional analysis and grouping are linked. Dimensional analysis might therefore promote a level of fluidity in how we handle variables, at least on the axis of thinking of them as many individual entities, or at times as conglomerations that we want to reason about jointly.

3.11.3.6 Recursive plug and chug

Two ontological metaphors ("equation as mutable - agentive" and "equation as one form of a relationship between variables") were significantly more common in the recursive plug and chug e-game. However, the total number of tags in the plug and chug game was very low because the game was rarely played. The result that these tag association are significant might be a result of the various approximations I made in setting criteria for significance. These associations are something to keep an eye on in larger data sets, or data sets in which students play the plug-and-chug game more often.

3.12 Conclusions and implications

People use many different types of reasoning, both in everyday life and in physics. In physics problem solving, one particularly important way students can vary is in their epistemologies. The epistemological framing students take towards solving a problem affects what types of evidence they bring to bear, what sorts of arguments they accept, and how they communicate with partners or instructors.

Epistemological framing influences more than just problem solving, as well.

Students listening to a lecture will interpret it differently if they frame it differently, for example. They will take very different attitudes towards labs if they view labs as leading them to create a model than they would if they viewed labs as trying to get the expected results in order to verify a known theory.

PER researchers have also studied the variety of ontological metaphors both students and experts use to think about physics. Most work on ontologies has been interested in ontologies for their own right, asking about how productive different ontological metaphors are for various physics concepts. They may map out a trajectory for how ontologies develop in students, for example.

Here, I've taken the view that students use a variety of ontologies for mathematics, but not that any one ontological metaphor is better than any other. I haven't been interested in these metaphors exclusively for their own sake, either. Instead, I wanted to know whether using certain ontological metaphors was linked to certain epistemological frames.

Sometimes in problem solving, we want students to find an equation that they haven't seen before. To do that, an epistemological frame in which your own experience and understanding of physics are valid resources to draw on is very productive. In this frame, students can view themselves as creators of equations, not just receivers of knowledge. But if students are viewing themselves as creators of equations, wouldn't they also have to view equations as belonging in the category "things that can be built"? An alternative view might be that the equation already exists, and students are "finding" it. A cursory search for "is mathematics created or discovered" shows that many people care deeply about this distinction, artificial as it might sound when presented as an ontological metaphor. Why would they care so much if ontological metaphors weren't setting their entire framing of mathematical experience?

As it turns out, I didn't find strong evidence for the particular epistemology / ontology link I just described, but I found several others.

I began this thesis in chapter 2 looking especially at extreme / special cases. I found they were valuable tools that students took up readily. When students thought about extreme cases, they activated their physical intuitions and sought coherence between them and mathematical statements. Here, I found that they also adopt a new ontological metaphor in this process - that of thinking about symbols as parameters.

Fortunately, many of the students in PHYS 131 already had a lot of facility with thinking of symbols as parameters, and only needed to activate it to be successful in using the extreme cases game. But ontological metaphors, like other thinking tools, must be learned at some point (though not necessarily explicitly). If an instructional environment is seeking to help students use this tool but failing, it might be helpful to look at designing interventions around looking at symbols as parameters and visualizing systems in an abstract parameter space before working on the extreme/special cases tool.

Within the sensemaking e-games of "mapping from meaning to mathematics" and "mapping from mathematics to meaning", both important parts of the general instructional goals espoused in this thesis, two other ontological metaphors played an important role: "equation as one form of a relationship" and "equation and phenomena not differentiated". This suggests that instructors looking to create opportunities for sensemaking could look at activities that encourage students to use or develop these ontologies. For example, to reflect on how different forms of the same equation might be useful for solving different problems. This was in fact an essay question on a PHYS 131 exam.

3.13 Future work

My codebook for equation ontologies is not yet completely validated. Future work will need to involve multiple coders and compute inter-rater reliability before this work can be considered complete.

This work should also be validated with other data sets, and possible with larger data sets. All the data here comes from a set of problem-solving interviews I conducted. While this presents a controlled environment, it also presents an artificial one.

We have collected various artifacts from PHYS 131/132, including test essay questions (such as the one mentioned in the previous question) and free-response validation of the MEGS. We also have some data of students solving problems together in lab and tutorial sessions. All of these settings to continue testing the ontology / epistemology link in settings more natural to the course.

I also hope that other researchers will find this work interesting and pursue the ontology / epistemology link in other settings: different courses, different demographics of students, different instructional environments, etc. There are probably a variety of different research methodologies that would also be useful here.

There are many more ontological metaphors than I've been able to catalog here. For example, I saw some students extensively use an "equation as process" metaphor. In this metaphor, an equation is viewed as a set of instructions for what to do to the values of the variables. Marion use this metaphor extensively when working on the ellipse problem:

> the last one pi a plus b over two squared. um. So a plus b. So then I said before that was just equal to a length and then we have a length divided by pi over two. Take that out. And then um, and then we square that so divide by two the length divided by two that's still a length and square it.

The equation is seen as a series of steps or instructions:

- add a to b
- divide the sum by 2
- square that

and this view seems to affect the way Marion makes sense of the equation, in this case by finding its dimensions. There are many types of reasoning still to explore, and many potential connections between them.

Chapter 4: The Math Epistemic Games Survey (MEGS)

4.1 Overview

In solving physics problems, we value student sensemaking: finding meaning in the quantities and terms of a problem and fitting these meanings into a larger conceptual structure. Mathematics, in particular, provides many rich opportunities for physics sensemaking, but these opportunities are often ignored. Previous work has characterized student problem-solving behavior as belonging to one of several "epistemic games" - cognitive frameworks for understanding "what's going on" in problem solving [Tuminaro and Redish, 2007].

Here, we introduce the Math Epistemic Games Survey (MEGS), a 30-question, multiple-choice concept inventory of mathematical questions set in the context of sensemaking, especially for physics for the life sciences. The MEGS focuses on four core epistemic games: examining extreme cases, dimensional analysis and scaling, estimation, and mapping symbols to physical meanings. We describe the creation and validation of this survey and its early results in a large introductory physics for life sciences course over three years. We show that even in reformed-pedagogy, active-learning classrooms which obtain good results on standard physics concept inventories, students may show negligible change in how they use epistemic games to solve challenging, context-rich problems. However, with specific targeted instruction in key mathematical sensemaking strategies, we have observed modest gains in MEGS scores.

4.2 Motivation and Need for a Test

We created the MEGS as part of the project “Understanding and Overcoming Barriers to Using Math in Science” (Under/Over project), an NSF grant won by Professor Redish. The overall goal of the project was to better understand why using math to reason about physics is different and challenging for students who have taken math courses, especially the introductory calculus sequence for biology majors offered at UMD.

Previous work by Redish and Kuo [2015] combined various qualitative observations that provided an outline of the sorts of adaptations that are expected in this transition, while Tuminaro and Redish [2007] provided a framework for categorizing them. A multiple-choice test focused on problem solving strategies plays an important part, alongside qualitative analysis of problem-solving interviews, in measuring and disseminating this work, and creates the method for measuring future classroom interventions’ effectiveness on a particular set of problem-solving skills.

4.2.1 The need for a concept inventory

PHYS131 and 132 enroll about 400 students per year at UMD, a much larger number than we could hope to collect data from by classroom observation or interviews. A multiple-choice assessment of problem-solving strategies interpreted as epistemic games would allow us to gather a great deal more data relevant to improving these classes than interviews alone.

Multiple-choice assessments also afford different types of analysis. They can be graded as correct/incorrect, allowing us to compute scores, which provide quantitative measures of the change in students' problem-solving behavior over the course of a semester or year (when the test is applied twice, at the beginning and end of the period studied). There are many caveats to the interpretation of this sort of data, but with a carefully-constructed test they provide valuable feedback on the effectiveness of instruction, as the example of many previous concept inventories shows [Madsen et al., 2017].

4.2.2 Pre-existing tests

After deciding to use a multiple-choice assessment on epistemic games in UMD's PHYS131/132, we had to decide whether to use a pre-existing assessment or design a new one.

There was no previously-existing test explicitly written to measure student use of epistemic games. This doesn't mean our test would necessarily fill a void, because the problem-solving strategies we identify as epistemic games go by other names in other work. For example, White et al. [2017] identify some of the same strategies as "metacognitive gimmicks". They assessed student facility with metacognitive gimmicks by analyzing student responses to prompts which asked them to check their answers after solving problems in an electromagnetism class [Sikorski et al., 2017]. This was in place of using a concept inventory. Our search did not show any concept inventories aimed specifically at dimensional analysis, extreme cases, and estimation.

Most existing assessments focus on either a specific content area of physics (e.g. FCI focuses on Newton's laws), or are "belief/attitude" assessments (e.g. MPEX). Physport [Phy] has only one entry in its "problem-solving" assessment category, the Assessment of Textbook Problem-Solving Ability (ATPSA) [Marx and Cummings, 2010]. This assessment uses content from mechanics, including Newton's laws, momentum, and energy. It would therefore be inappropriate to use this test with students who have no prior experience in physics. Many 131 students fall into this category, and we wanted our concept inventory to be effective for them. ATPSA identifies the problem-solving subskills it evaluates as, "classify the single content area relevant to the problem and identify equations germane to that area; and mathematically manipulate symbols and numbers to arrive at a correct numerical solution" [Marx and Cummings, 2010]. The focus of our research was on supplemental strategies that would scaffold student sensemaking, both to build physical and mathematical understanding and intuition, and to serve as a backup and metacognitive aid to the process described by the ATPSA.

A second relevant pre-existing assessment is the Mathematical Modeling Conceptual Evaluation [Thornton]. This assessment studies student understanding of the functional forms of equations and how they relate to modeling via various representations, especially graphs and diagrams of vector arithmetic. While closely aligned with our sense-making goals, this assessment was not appropriate for our research project because it required physics-specific background knowledge our students aren't expected to have, and because it has not been validated by student interviews or expert evaluation.

Because the focus of our study was on mathematics, we also investigated several assessments developed in the mathematics community. Marshall [1988] writes in support of tests of "higher-order skills" and constructing tests to measure not how much a student knows, but how the student applies the knowledge; a key goal of the MEGS. Although Marshall didn't present such a test, she did suggest that paradigms of statistical analysis based on graphs are more relevant for this type of assessment than classical test theory, a topic I'll explore in Chapter 5.

Another related test was Lawson [1978]'s "Test of Formal Reasoning", which didn't use equations, an important part of our assessment. Tobin and Capie [1981]'s "Test of Logical Thinking" had questions on proportional reasoning were somewhat similar to ours. These were rarely amenable to strategies like extreme case reasoning or dimensional analysis.

Epstein [1993]'s "Basic Skills Test" asked several questions aligned with our goals, although the test was not designed to measure use of problem-solving skills, but as a more direct measure of accuracy in applying fundamental techniques.

We decided that none of the existing assessments were very close to our goals, and so we would create our own.

4.2.3 Objectives of the MEGS

After deciding to create our own assessment, we identified the specific goals of the assessment. These were based on previous work, especially Redish and Kuo [2015], as well as our previous classroom and problem-solving experience. We set these goals over brainstorming sessions between myself, Professor Redish, and Deborah Hemingway, and received feedback on them via a meeting of the advisory board for the Under/Over project.

We identified four problem-solving strategies we held to be helpful to students solving physics problems:

1. Examining extreme or special cases of formulas (special cases)
2. Dimensional analysis
3. Breaking estimation problems into pieces (estimation)
4. Mapping symbols onto their associated physical meaning (mapping meaning)

Our goal was for these strategies to be specific enough to be easily identifiable and codable in a problem-solving process, to be understandable at the level of introductory physics, to be broadly applicable to a wide range of physics area and different types of physics problems, and to be well-accepted as important problem-solving strategies in the physics community. If we met all these objectives, we believed that the results skills would be useful to students and capture pieces of what is generally meant by “problem-solving ability” and “thinking like a physicist”.

In constructing the test, we had to decide whether to explicitly or implicitly evaluate the problem-solving strategies that interested us. Explicitly testing them would, for example, ask students to perform a specific dimensional analysis or compute the limiting case of a certain formula. An implicit strategy would present a problem in which the dimensional analysis or limiting case would be useful for solving the problem, but the problem wouldn't tell students to use the appropriate strategy; they would (hopefully) figure that out for themself.

We learned of several efforts to teach and assess some of these strategies, especially extreme cases and dimensional analysis. These efforts would often explicitly ask students to examine the extreme case of a given equation, or to find the dimensions of a given formula, and assess students on the accuracy with which they carried out the task. For example, Sikorski et al. [2017] taught students about the "usual three ways" of checking solutions (dimensional analysis, extreme or special cases, whether the answer is numerically sensible) in upper division electromagnetism. They noted,

> Perhaps counter to our hopes, we observed "script-like" implementation of the checks. For example, though not instructed to do so, students labeled their checks (e.g., Fig. 2, 3, 6) and most conducted the checks in a fixed order (units, limiting cases, reasonable values). We also observed a decrease in "other" kinds of checks that students used; even when asked to check solutions "in as many ways as they could think of", students checked for the three ways. The script-like implementation

> stands in contrast to more fluid application that we might expect of expert physicists, and hope to develop in our students.

Although they noted other evidence of students using the three checks for sensemaking, we sought to avoid the sort of script-like behavior noted here. We believe that decision is reinforced by other work showing that prompting for procedures that instructors intend for sensemaking can often reduce fluid, story-like problem solving behavior in students. Instead, students shift towards simply fulfilling the requirement to use that procedure (for example, Heckler [2010] and Kuo et al. [2015] showed this for prompts for free body diagrams and Holmes et al. [2017] found epistemological frame shifts towards procedural frames when exposing students to similar cues in critical-thinking-focused lab courses.)

Our goal in promoting and studying problem-solving strategies is for these strategies to be useful tools when students solve problems that arise in organic contexts. We want students to be able to estimate things in real life, not just on a test asking them to perform a piece of a curated estimation problem.

An important part of a making a strategy work for you is recognizing when it is useful and applying it in context. An assessment that gives an equation and asked about a certain extreme case does most of this work for the students. Scoring well on this sort of assessment wouldn't mean that students were using the problem solving strategies to good effect to bolster their physical understanding; it could instead indicate a technical mastery of the skills without realizing its importance to more general problem-solving.

We developed a goal to write problems that would challenge students and invite them to apply an appropriate problem-solving strategy, but would still have students do most of the work. That work includes choosing to use a strategy, choosing what strategy to use, and applying it appropriately to the problem.

We developed this instrument specifically for UMD's Phys 131/132 IPLS courses, with a goal to later disseminate the assessment to a broader community, but especially in IPLS. We wanted many of the questions to fit into a biological context and convey biological authenticity. This would contribute to the courses' overall goals of promoting the value of physics and mathematical models to understanding biological phenomena [Redish et al., 2013].

4.2.4 Initial Development of Test Items

We developed test items in three ways:

1. Borrowing and adapting items from other sources
2. Soliciting contributions from the advisory committee to the Under/Over project
3. Developing new items ourselves based on the goal epistemic games

4.2.5 Attributions

James Alexander, a member of the project advisory board, contributed drafts of questions that became MEGS questions 1, 10, 13-15, and 17. Question 6 was known to us from Robert Beichner in the SCALE-UP project. Questions 22 and 23 are based closely on questions in Brahmia et al. [2016] with question 23's original source being Clement et al. [1981] . Question 5 uses an expression from Phillips et al. [2012].

4.2.6 Internally-Developed Items

After identifying the categories of problem-solving strategies we intended to assess, we wrote test items targeted specifically at each strategy. Our goal was to write items that were best-solved with these strategies, although the problems could be solved in other ways. Our conjecture is that when students develop strong skills in applying problem-solving strategies, their accuracy will increase on the problems we wrote corresponding to those strategies.

For source material, we did not use physics content aside from motion, because we wanted the test to be applicable at the beginning of an IPLS sequence, when students are not expected to have any formal physics instruction.

We set many of the questions in a biological context, where math was being used to model a scenario in health, physiology, or cell biology. These topics are closely related to those studied in Phys 131/132, and one of our principal goals was to study the usefulness of mathematics in physics for biology. However, other problems are set in a more general context, such as everyday life contexts.

Additionally, as Question 25, we included a "validation" question asking students to select option "d". This was to eliminate the responses of students who were guessing randomly or answering the same way to every question without reading the questions.

We set these problems over the course of two months of weekly meetings. We brainstormed problems in and outside of meetings, then solved each others' problems, identifying the epistemic games we played in doing so. We discussed the problems' appropriateness in terms of level of difficulty, contextual suitability, and the importance of the identified epistemic games to the problem-solving process. We then proof-read the questions in an attempt to identify and fix small errors and ambiguities.

We rejected many early question drafts and revised others before assembling our first draft, which consists of 26 questions.

As described below, we both added and dropped questions on later revisions, eventually reaching 30 content questions (including the validation question). We then added two more questions. Question 31 asks students how much effort they put into the test, and we added it after concerns that some students were putting little effort into the test, and specifically might put different amounts of effort at the beginning and end of the semester. Question 32 asks what percentage of their answers students' believe to be correct, to allow a comparison between student confidence and actual accuracy.

For the current version of the MEGS survey, please see appendix D.

4.3 Revision and Validation

4.3.1 Test administrations of the MEGS

Various versions of the MEGS have been administered as pretests and post-tests in PHYS131 and 132 at UMD since fall 2015. The MEGS has also been administered at NEXUS/Physics-based courses at Montgomery College and Swarthmore College. For a complete listing of dates and versions of the MEGS administered, please see appendix E.

We tried delivering the MEGS with different methods. The MEGS was originally implemented as a Scantron test. Students received a printed paper question booklet, a blank Scantron sheet, blank white scratch paper, and, if needed, a pencil. Students were allowed to use a calculator if they had one available, but were asked not to use programming features of the calculator or to look things up online or on their smart phones. Students were also allowed to use smart phones as calculators, but were asked to keep the phone in the calculator app. Students worked alone on the MEGS.

In later administrations, we switched to an online version of the MEGS implemented in Qualtrics. This switch saved on paper, made processing of the results easier and faster, and avoided errors where, for example, students would skip one question on the scantron and answer a long string of questions one spot off (e.g. record their response to question 19 in the spot for question 20, then record their response to question 20 in the spot for question 21, etc.) I estimated these sorts of errors to contaminate about 2% of early MEGS responses, while more minor errors, such as not erasing an answer fully enough to be ignored by the scantron, were present on about 15% of Scantron responses, and required significant effort to correct by hand. Additionally, the online MEGS could be administered at remote institutions, and data shared back with the research team, far more easily than for a paper-based version. Finally, the online MEGS recorded the time students spent on each question, which allowed some analysis of student effort, and to exclude results in which students responded so quickly that we believed it unlikely that they were reading, considering, and working the questions before responding.

Participation in the research study was optional; students were free to refuse to share their responses with the research team, or to not take the test. (In most cases, students would have been sacrificing a small number of participation points by not taking the test at all. However, they would receive participation credit for taking the test but not sharing their results with the research team.) Students were not otherwise compensated for taking the MEGS.

At UMD, students usually received participation credit in Phys 131/132 for taking the MEGS and the MAX, but did not receive additional credit based on their score. As a default, students did not expect to see a score on the test or have it affect their class performance in any way other than the participation score they received. The MEGS specifies that no instructional staff will learn what student responses were to MEGS questions while the students were enrolled in Phys 131/132. However, after early concerns that students were not putting consistent-enough effort into the test to result in reliable quantitative comparisons between administrations, some instructors changed their policy. One instructor offered to use the MEGS results to provide students individualized feedback on their strengths and weaknesses in math. Additionally, in some instances, instructors changed their course policy to treat the MEGS as a graded assignment, which significantly affected student accuracy. For a complete summary of the policies in effect for each test administration, see appendix E.

Students took the MEGS during scheduled class time during the first or second week of classes (pretests) or last or second-to-last week of classes (post-tests). At UMD, students had time scheduled for labs. However, in the first week of the course, there was no scheduled lab, so this time was used to take surveys. The MEGS was given alongside the MAX attitude survey, and in some cases one or two additional surveys created by other research projects or by UMD for course evaluations. In these cases, students were asked to take the MEGS first.

4.3.2 Administration concerns

At UMD, the MEGS was in some cases administered (or the beginning stages of it were administered) by a member of the research team, either myself or Deborah Hemingway. In other cases, especially at the end of the semester, TAs were trained in how to administer the MEGS and handled the administration based on written instructions and verbal instructions given at a group meeting fur TA training. Sometimes a member of the research team attended these training meetings, and in some cases they were held exclusively by a head lab TA who had been briefed by the research team.

We know that in several cases, TAs gave students the wrong URL to take the online MEGS (for example, students in Phys 131 received the URL for Phys 132). When administering Scantron tests, TAs often did not follow written instructions on e.g. where to collect the tests. This raises questions about how closely MEGS administrations adhered to instructions, and whether TAs received conflicting written and verbal instructions. It is not guaranteed that students always had access to calculators and scratch paper and did not have access to the internet or other tools while taking the MEGS. Similar concerns apply during the first administration of the MEGS because the research team had not completely defined and disambiguated instructions in its first administrations of the assessment.

While we attempted to administer the MEGS the same way in every test administration, this was logistically not practical. In one case, the class time set aside for the MEGS was canceled due to weather. In other cases, students missed the class or added the class after the date when the MEGS was delivered. In these instances, the MEGS was set up in the course center, a collaborative study area reserved for use by PHYS 131/132 students. Students in that environment may have had others working on homework problems nearby while they were taking the MEGS.

Unlike exams and homeworks in PHYS 131/132, we never received reports of potential cheating on the MEGS from students or TAs, although it would not be difficult for students to cheat.

4.3.3 Validation Interviews

After drafting the MEGS, we began a series of validation interviews. Our goals were to understand how students interpreted the questions on the MEGS, to discover ambiguities, inconsistencies, or errors in the test, to observe the problem-solving behavior students used while working on MEGS items, and to generate appropriate distractors for the test items. Overall, we conducted 24 hour-long interviews. I conducted 20 of these and Deborah Hemingway conducted 4.

4.3.4 Procedure for validation interviews

Participants for MEGS validation interviews were drawn from volunteers from PHYS131 who had taken the MEGS several weeks to months earlier as a beginning-of-semester pretest. Subjects were recruited by an email sent out to the course, and were compensated for their time.

During interviews, the interviewee had a copy of the paper MEGS in front of them. The interviewer would suggest certain problems from the MEGS, either working in order through the test, or selecting problems of particular interest based on previous interviews. The students would read the problem and work it out on a whiteboard, explaining their thought process aloud while they worked. The interviewer would interject occasionally to prompt the student to continue thinking aloud. If the student was stuck on a problem, the interviewer might move on to the next problem, offer a metacognitive prompt (e.g. "can you tell me what your basic strategy is right now"), suggest a problem-solving strategy, (e.g. "can you try using a specific example number for each of these variables") or supply a crucial missing piece of information (e.g. "a normal weight for a six-foot tall person is 170 pounds"). When a student completed a problem, the interviewer often asked how confident the student was and whether they believed the answer made sense. The interviewer would then move on to another problem, or suggest applying a certain problem solving strategy, depending on how they judged the situation.

Interviews were video and audio-recorded. The interviewee sat either next to or across from the interviewer, with a white board on a table and whiteboard markers and erasers. The camera was positioned in an attempt to see the whiteboard, what students were writing, and the students and interviewers' faces and bodies, but this was not completely successful in all cases. After the first few interviews, students were given a stack of whiteboards and asked to move on to the next whiteboard when they had filled one up, rather than erase it (but students often still erased by instinct), and photographs were taken afterwards of student work.

Validation interviews followed a think-aloud protocol, in which the interviewer asked students to explain their reasoning aloud as they worked on problems [Ericsson and Simon, 1980]. In some interviews, students were given MEGS questions without answers so that their responses could be used to formulate appropriate distractors. In other cases, they were given MEGS problems with answers to more closely simulate the test-taking experience.

Generally, students worked much more slowly during validation interviews than when taking the MEGS. Median time to take the entire MEGS was about 30 minutes, but in hour-long interview sessions, interviewees typically worked on only

10 MEGS problems.

4.3.5 Revisions based on validation interviews

While conducting validation interviews, we continually revised questions, usually with the goal of student interpretation of the question being closer to the question-writers' intention. We also revised the questions to make the answer less ambiguous. For example, we made the values in question 10 depart from a straight line more severely to avoid ambiguity over whether the data differed "significantly" from a linear trend; some students correctly observed the size of the original deviation from a straight line, but were unable to decide whether it was significant. In some cases we caught typos or other errors in the questions, such as a misplaced minus sign.

We did not hold the standard that every student should understand what was being asked in each question, because for some questions, we expect that student understanding of the question will improve with better understanding of how language is used in physics and better ability to map words into symbols and build mental representations.

For example, some students have claimed that question 27, which asks about "running speed" while running opposite the travel direction of a moving sidewalk, is ambiguous, depending on ambiguities around the words "fast", "speed", and "velocity" as signed or unsigned quantities. We think it is unlikely that expert validation will confirm this ambiguity, and that as students become more adept at the “mapping symbols onto meaning” epistemic game, they will be more likely to see this question as clear and unambiguous. There are similar issues in question 26, asking where hundred-dollar bills fit on a chart of value by dollar, and 23, asking to map a verbal statement into an equation.

4.3.6 Observations from validation interviews

4.3.6.1 Most students have the requisite knowledge to solve most MEGS problems

The most striking result from our interviews was that although MEGS scores are usually only around 50 percent, almost all students interviewed were able to solve almost every MEGS problem they attempted.

We identified a number of possible contributors to this apparent discrepancy:

- In an interview environment, with a researcher present, students may be more motivated to solve problems correctly
- Students spend more time per problem during the interview than while taking the MEGS
- The interviewer occasionally helped the students in various ways.

Our goal in interviews was to observe student thinking as they worked on MEGS problems, so in many cases, after a student had looked at a problem, the interviewer prompted the student in some way. Sometimes this was simply encouragement to “explain your thinking”.

Other times, the interviewer was providing a sort of metacognitive scaffolding, asking the student whether they were confident in their answer, whether the answer made sense, whether it worked with a specific example, whether they had seen a similar problem before, etc. These sorts of metacognitive prompts were usually enough that students who heard them found mistakes in their solutions, became unstuck, and solved the problems correctly.

This suggests to us that mastery of the MEGS really is about calling up the appropriate resources at the appropriate times, something strongly facilitated by playing a productive epistemic game. If we had found that students getting MEGS questions wrong usually simply did not have the requisite knowledge (for example, did not know what dimensional analysis is, or did not understand how to evaluate the expression $\lim_{a\to\infty}\frac{ab}{a+b}$), it would contradict many of the hypotheses and assumptions we began the project with, and which undergirded work preceding this project [Redish and Kuo, 2015].

In a small number of questions, the interviewer gave a more specific suggestion, but in no case did the interviewer give information which is not common knowledge or covered extensively in prerequisite courses, and no student ever expressed that the interviewer told them something they didn't already know. Instead, they expressed that they knew the information, but hadn't previously considered it relevant.

Thus, our validation interviews discount the hypothesis that MEGS measures student knowledge; students already have the fundamental knowledge, and instruction around MEGS-type questions in physics class will realize gains by helping students access the appropriate knowledge at the appropriate time. Because accessing knowledge is strongly mediated by playing epistemic games, the MEGS is a survey of epistemic games, despite not asking about them directly.

4.3.7 MEGS questions are now mostly well-understood

After revising MEGS questions based on our observations, re-readings, and statistics from early MEGS administrations, we have now found that students in interviews apparently interpret most MEGS questions correctly most of the time. Even with a significant number of interviews, the sample size on any particular question is fairly small because there are 30 questions, and typical interviews covered less than a third of the test. Still, with the latest version of MEGS, every question has been read in an interview setting by at least three students, with no obvious errors in interpretation of the question which we were able to attribute to ambiguities in the question (as opposed to, e.g. a student misreading or ignoring a word in the question).

4.4 Test Reliability

In this section, we review some results from our administrations of the MEGS and show that, according to common standards in classical test theory, the MEGS may mostly be considered a reliable, valid test with good discrimination, with some caveats, which we discuss in the appropriate sections. A general rule for performing the analyses below is that the test should be given to at least 5-10 students per question [Crocker and Algina, 1986]. Below, we accumulate data from all administrations of the MEGS, giving $n > 1500$, well in excess of this minimum for a

30-question test.

For a summary of all the administrations of the MEGS, please see Appendix E.

4.4.1 Difficulty

In classical test theory, the "difficulty" of a test or question refers to how often it is answered correctly. This is sometimes treated as a property of the question, but is dependent on both the question and the population taking the test.

The over-all difficulty of the MEGS, across all our administrations pre and post, is 0.58. This is a good difficulty for a concept inventory on its target population. Engelhardt [2009] suggests that an average difficulty of 0.5 is ideal. If the difficulty is much lower, students are unable to solve the questions, the test is too hard, and a lot of what you see in the data will be guessing. If the test is easier, it doesn't allow measuring students at the higher end performance.

The individual test questions on the MEGS vary in difficulty from 0.09 to 0.90 (excluding the validation question, whose difficulty is 0.97). The MEGS question difficulties are show in table 4.1.

While it may seem that a question with a difficulty below 0.1 is too hard, the question with the greatest difficulty is question 16, discussed in section 4.8. We believe this question is within the realm of difficulty that IPLS students can handle, and indeed most students are able to solve it correctly in an interview setting, with minor prompting from the interviewer. We believe the case is similar for other

Table 4.1: Difficulties of MEGS questions based on existing data

MEGS question	Difficulty
1	0.85
2	0.64
3	0.77
4	0.55
5	0.81
6	0.51
7	0.19
8	0.53
9	0.60
10	0.70
11	0.67
12	0.61
13	0.87
14	0.11
15	0.78
16	0.09
17	0.67
18	0.47
19	0.54
20	0.34
21	0.90
22	0.52
23	0.37
24	0.82
25	0.97
26	0.42
27	0.71
28	0.52
29	0.45
30	0.42

difficult MEGS questions.

Questions with difficulties above 0.8 (questions 1, 5, 13, 21, and 24) are candidates for deletion on future revisions to MEGS, see section 4.9.3 for more.

4.4.2 Summary statistics

The mean score on the MEGS is 17.7 with a standard deviation of 4.2. The median and mode scores are both 18.

The distribution of MEGS scores is moderately left-skewed (long left tail). The skewness (third moment about the mean) of the MEGS is -0.22 over 1439 samples. A normal distribution on a sample of 500 will have skewness between -.18 and 0.18 more than 90 percent of the time [Doane and Seward, 2011], so the skewness observed in the MEGS is likely genuine. Although conventional knowledge is that left-skew results from a fairly easy test, in this case, it may also result from a small number of insincere test-takers scoring very low scores.

4.4.3 Standard Error of Measurement

When a student earns a certain score on the MEGS, that score represents partially a measure of the constructs underlying the MEGS questions, and partially the students' luck in guessing, or that they randomly happened to know a little more or less about MEGS questions than they would if the MEGS were re-written on the same topic but with different questions. Thus, the score is “wrong” by a certain amount. The typical amount a student score is wrong is the standard error of measurement of the MEGS.

Engelhardt [2009] presents the following formula to estimate the standard error of measurement:

$$SEM = \sigma_t \sqrt{1 - r_{tt}}$$

where σ_t is the standard deviation of the test scores and r_{tt} is the reliability of the test (either the split-halves reliability or the Kuder-Richardson score, both described below).

When the test is very reliable, the reliability r_{tt} is close to 1, and the standard error of measurement is very low because the test's high reliability makes us confident that the student has tested near their true score. When the test is highly unreliable, the student could lie almost anywhere in the distribution of all students; their score doesn't tell us much, and the standard error is as large as the spread in student scores itself.

Using the split-halves reliability test, the standard error in MEGS scores is 2.5. Using the Kuder-Richardson test, it is 1.8. This means that we can measure student MEGS scores roughly to an accuracy of two or three questions. A student who improves significantly more than 3 questions over the course of a semester made significant gains on the constructs underlying the MEGS, assuming similar testing conditions and effort.

Because individual administrations of the MEGS have typical sizes of more than 100 students, we can detect entire-class shifts in MEGS scores a factor of $\sqrt{100} = 10$ times smaller than for individual students; so a shift of 0.3 in the average score between pre and post-test, or a shift of above 1 percent, should be considered meaningful.

4.4.4 Test-retest reliability

The test-retest reliability determines whether a test is reliable by giving it twice, separated by a short time. If the test is reliable, students should get the same score on both times taking the test. However, if students learned new things between the test administrations, or answer the second time using what they remembered from the first time, the test-retest reliability is not a valid indicator. Based on comments from students, both conditions are true for the MEGS, so we will not include test-retest reliability estimates here.

4.4.5 Split-halves reliability

Another method to determine if a test is reliable comes from comparing scores on different parts of the same test. If the test is highly reliable, we expect students who score well on the first half to score well on the second half as well, for example. The split-halves reliability test divides a test into two halves and measures the correlation (Pearson product-moment correlation coefficient) of the student scores between them.

The correlation obtained depends on which questions are chosen for each half, so I conducted the test repeatedly, each time choosing the two halves uniformly at random from between all $\frac{1}{2}\binom{30}{15}$ possible two-half partitions of the questions. The correlation coefficient is usually transformed according to the equation

$$r_{tt} = \frac{2r_{hh}}{1 + r_{hh}}$$

with r_{tt} the Spearmon-Brown prophecy formula and r_{hh} the correlation coefficient between two halves of the test. This transformation leaves correlations of 0 and 1 alone and moderately increases all other correlations.

Using this transformation, I found the following distribution of r_{tt} to vary between 0.62 and 0.64 for the MEGS.

According the Engelhardt, reliability below 0.70 indicates “Low, useful only for group averages and surveys”, so it is worth asking why the MEGS has low reliability by this measure.

The reliability will depend on the length of the test. In a short test, the reliability will be low simply because the random variations in score are more significant compared to the total scores; there isn’t enough length for random fluctuations to cancel out. At 30 questions, the MEGS is not especially short.

Second, tests of “reliability” might also be called tests of internal consistency. They operate on the assumption that a test is testing a single underlying construct that students have to varying degrees. The MEGS was explicitly constructed to test four different constructs, potentially all related via things like metacognition, or correlated in the way that many cognitive constructs are. Because there are several different constructs underlying the MEGS, we should expect the same sort of consistency we might for a test of a single construct.

4.4.6 The Kuder-Richardson test

The Kuder-Richardson test is another method of estimating the reliability of test via internal consistency.

If every question on a test were unrelated to each other, so that students answered each question randomly and completely independently (knowing whether a student did well on a few questions told you nothing about how they would do on the remaining questions), then the variance in the score of each question would sum to the variance of the whole test; this is a property of independent random variables.

However, if the questions all measured some single underlying construct, students who score well on one question are likely to score well on the next one, etc. The variance in the scores on the entire test will be greater than the sum of the variances on individual questions.

The Kuder-Richardson test exploits this, defining

$$r_{tt} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k} p_i(1-p_i)}{\sigma_t^2}\right)$$

where r_{tt} is the Kuder-Richardson reliability, k is the number of questions on the test, p_i is the difficulty of item i, and σ_t is the standard deviation of scores on the test. In the case of independent questions, the Kuder-Richardson reliability is zero (in the limit of a large number of test-takers). It is always less than 1, but can be fairly close to 1, especially on longer tests.

The Kuder-Richardson reliability of the MEGS is 0.81 on the population tested so far, which Engelhardt identifies as “Fairly high, possible for measurement of individuals”. This contrasts with the split-halves reliability of the MEGS, although the tests are intended to measure the same thing and be roughly equivalent. The same considerations that qualified the split-halves reliability of the MEGS apply to the Kuder-Richardson reliability.

4.5 Item reliability and validity

In the previous section, we discussed statistics that apply to the entire MEGS test, taken as a whole. In this section, we review statistics that analyze each question of the MEGS individually.

4.5.1 Discrimination index

A question might be considered bad, or at least worthy of careful individualized analysis, if high-performing students tend to get it wrong and low-performing students get it right. On the other hand, if a single question could cleanly divide students who will do well from those who will do poorly on a test, that question would be contributing a lot of information. This is the idea behind the discrimination index of a question.

The students taking a test are divided into two groups - those with high scores (at or above the median) and those with low scores (below the median). The difficulty of a question is computed for each group, and the discrimination index is

$$D_i = \frac{H_i - L_i}{N/2}$$

where N is the number of students taking the test, L_i is the difficulty of question i in the students in the low-scoring group, H_i is the difficulty of question i in the students in the high-scoring group, and D_i is the discrimination index of question i. The discrimination index varies from -1 to 1, with -1 meaning a question which all low-scoring students answered correctly and no high-scoring students; 0 meaning low and high-scoring students answered the question equally, and 1 meaning that all high-scoring students answered it correctly and no low-scoring students did.

The discrimination indices of MEGS questions are shown in table 4.2

All MEGS questions have positive discrimination. Engelhardt writes that 0.30 is an acceptable discrimination index. All MEGS questions except 14, 16, and 25 (validation question) pass this benchmark.

Those with low discrimination, 14 and 16, are simply very hard; they can't have high discrimination because even in the high-scoring group, few answer correctly. We expect that in a future class which shows significant MEGS gains, these questions will have much higher discrimination.

Question 8, on unit conversions and the density of water, has unusually high discrimination, indicating that top-scoring students almost always answer it correctly, and low-scoring students rarely do. Only questions with difficulty near 0.5 can have such high discrimination. (Question 8 has difficulty 0.53).

Discrimination may be helpful in potentially shortening the MEGS in the

Table 4.2: Discrimination indices of MEGS questions based on existing data

MEGS question	Discrimination index
1	0.38
2	0.69
3	0.39
4	0.73
5	0.59
6	0.52
7	0.58
8	0.97
9	0.64
10	0.69
11	0.61
12	0.5
13	0.48
14	0.21
15	0.56
16	0.08
17	0.81
18	0.54
19	0.76
20	0.42
21	0.32
22	0.74
23	0.59
24	0.58
25	0.11
26	0.92
27	0.67
28	0.88
29	0.85
30	0.3

future; see section 4.9.3.

4.5.2 Point-biserial coefficients

The point-biserial coefficient tries to accomplish about the same thing as the discrimination index, but uses a different calculation which doesn't have all the same shortcomings; it can be as high as 1 even for questions with low overall accuracy.

The point-biserial coefficient is

$$r_{pbs;i} = \left(\frac{\bar{x}_{correct;i} - \bar{x}_{whole}}{\sigma_{whole}} \right) \sqrt{\frac{p_i}{1 - p_i}}$$

where $\bar{x}_{correct;i}$ is the average score on the whole test among students who answered question i correctly, $\bar{x}_{whole}$ is the average score on the entire test by all students, σ_{whole} is the standard deviation of the average score on the test by all students, and p_i is the difficulty of question i.

Engelhardt writes that point-biserial coefficients should be higher than 0.20. The point-biserial coefficients of MEGS questions are show in table 4.3

The point-biserial coefficient picks out the same two questions, 14 and 16, as the discrimination index did for having low discrimination. This is slightly more worrying, as it would be possible for the point-biserial index of these questions to be high, even though the number of students who got them right is low. Future validation will need to continue examining these questions carefully.

Table 4.3: Point-biserial coefficients of MEGS data based on existing data

MEGS question	point-biserial coefficient
1	0.37
2	0.45
3	0.34
4	0.4
5	0.48
6	0.31
7	0.45
8	0.54
9	0.4
10	0.46
11	0.39
12	0.29
13	0.52
14	0.17
15	0.46
16	0.11
17	0.5
18	0.35
19	0.45
20	0.29
21	0.4
22	0.43
23	0.37
24	0.5
25	0.3
26	0.48
27	0.44
28	0.46
29	0.46
30	0.2

Table 4.4: MEGS effort level interpretation

effort level	interpretation
1	I gave it my best effort
2	A lot
3	A medium amount
4	Only a little
5	No effort

4.6 Special concerns when using the MEGS

4.6.1 Difficulty of the test and student effort

In test theory, a concept inventory like the MEGS is designed to measure students' "optimal performance"[Crocker and Algina, 1986], their best effort given in an environment free from distractions. However, several pieces of evidence show that student effort is generally well below their best, especially at the end of the semester.

First, students report only a medium level of effort, and report lower effort at the end of the semester. After we examined the first results from the MEGS, we suspected that students weren't putting in a full effort at the end of the semester, and so added MEGS question 31, which asks for student effort on a five point scale defined in table 4.4

Students generally report lower effort on the post-test. In some administrations of the test, reported effort increased over the semester, but these were instances where students received participation grades on the pretest and accuracy-adjusted grades on the post-test. Under these circumstances, accuracy on the MEGS also improved substantially, indicating that students are capable of significantly higher scores than their usual, baseline performance when completing the survey for a participation grade.

The spread of MEGS scores generally increased over the semester, even when the average score stayed approximately the same. There are both more low scores and more high scores. One interpretation is that the new higher scores come from students who improve on the constructs underlying the MEGS over the semester, while the new low scores come from students who put very low effort into the post-test. Additionally, scatterplots of pre-post test performance show a number of implausible declines in MEGS performance. For example, on an early test, a student declined from 20/26 to 5/26 (early versions of MEGS had 26 questions, later expanded to 30). It is not unusual to see ten or so such large declines in a class of 120 students. These sorts of declines can significantly affect whole-class averages on the test.

To detect these sorts of low-effort responses to the MEGS, we added the validation question, question 25, which asks the student to select choice D. While a few students do answer the validation question incorrectly, the percent is small enough (about 3 percent) that it isn't able to adjust for low-effort by throwing out results that respond to the validation question incorrectly.

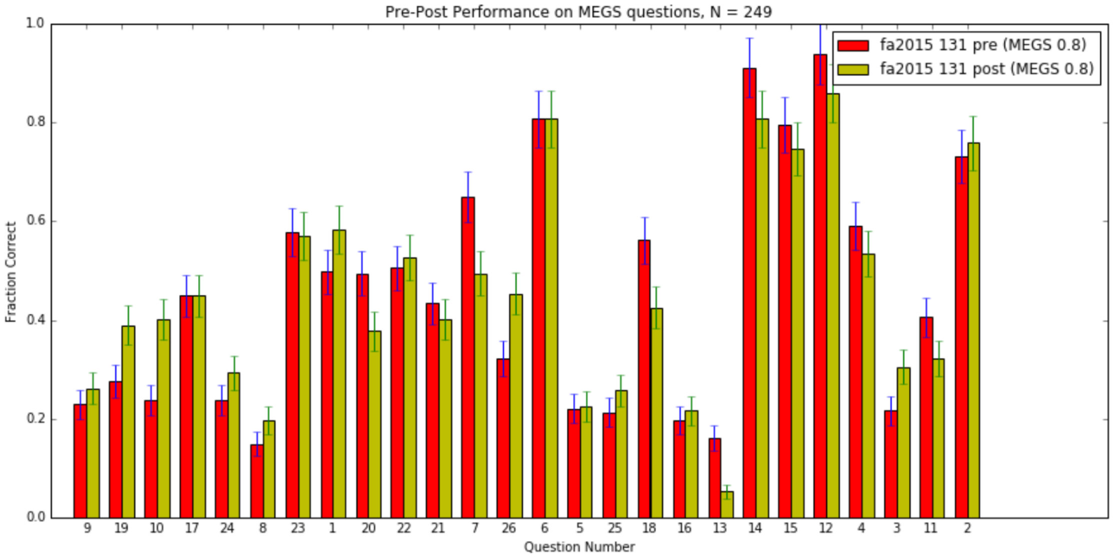


Figure 4.1: Pre and post performance on the MEGS, PHYS 131 Fall 2015, sorted by number of words in MEGS question

Another line of evidence suggests that MEGS results that are low effort, but not no effort, affecting the gain over a semester. Short MEGS questions, which can be read with less effort, generally show gain, whereas long questions show a decline. We noticed this on early sets of MEGS data, so I plotted pre and post performance on questions ordered by length, shown in figure 4.1

Among the shortest third of questions, students improved on 6 out of 8 and held steady on a seventh, declining slightly on the 7th-longest question, which came near the end of the test. Among the longest third of questions, students declined on 6 out of 8, improving only on questions 2 and 3, which came at the beginning of the test and which share the same introductory text. We interpreted these results to mean that students were less likely to perform optimally on long questions, and rewrote most of the long questions to make them more concise, and reduced the number of distractor answers.

We also switched the MEGS from a Scantron system to an online system on Qualtrics. This allowed us to measure the time students spent on each question on the MEGS. However, we have not yet filtered student responses based on the time they spent answering questions.

4.7 Results

Below, I'll review the things we've learned from early implementations of the MEGS.

4.7.1 Very little gain over one or two semesters of reformed IPLS instruction

Gains are often negative, and usually small. In some cases, they are positive because students were graded for accuracy only at the end of the semester.

This result of small gains shows that the MEGS is exposing a gap in the current instructional environment. The FCI, when applied to traditional lecture-based courses, showed gains of only 0.1 to 0.3 [Hake, 1998b], whereas active-engagement reformed courses show FCI gains of 0.3 to 0.7 [Hake, 1998b]. The early, low gains on FCI helped stimulate and justify the need for reforms [Hestenes et al., 1992b].

The MEGS does something similar. Although older implementations of Phys131/132 show good FMCE gains, later similar implementations show very low MEGS gains. (We never gave both MEGS and FMCE as pre-post tests to the same group of students.) Like early FCI results, early MEGS results show a set of skills that students aren't currently developing in our IPLS courses. Using the MEGS in new courses and new adjustments to current courses has the potential to justify and motivate modifications around mathematical sensemaking and epistemic games.

4.7.2 The MEGS is difficult and requires substantial effort compared to most concept inventories

As reviewed in section 4.6.1, there is substantial evidence that student performance on the MEGS is not optimal. This has rarely been a major topic of investigation in concept inventories. For example, Hake [1998b], in a major, many-institution study of the FCI, simply asked instructors, "Do you think that your students exerted serious effort on the FCI pretest", to which every instructor replied, "Yes". Hake continues, summarizing, "published reports of the courses not surveyed and my own knowledge of courses at Indiana suggests that students did take the pretest seriously.", and shows no other concern over student effort affecting the validity of gains. Physport's expert recommendations on concept inventories [Madsen et al., 2017, McKagan et al., 2017] do not discuss student effort. In fact, when they briefly discuss decline in student scores, they consider only the possibilities that students actually declined on the underlying constructs and that students got more questions right on the pretest via lucky guesses; effort is not mentioned. In a review of concept inventories, [Madsen et al., 2014] write, "Research shows that a majority of students take concept inventories seriously when they are given in class and not graded. Henderson compared students whose FCI score counted toward their final grade and those whose FCI score did not. He found evidence for no more than 2.8% of the students not taking the test seriously as a result of it not being graded".

This suggests that the MEGS is somehow different from most existing concept inventories. Typical concept inventory questions are short and don't require calculations, unlike MEGS questions. The multi-step process involved in solving MEGS questions appears significantly more effortful, as is shown in interviews where it was not unusual for a student to spend ten minutes working on a single MEGS question.

4.7.3 Students are mildly overconfident

Understanding whether or not your answer is likely correct is one small component of metacognition. Metacognition allows for a continual monitoring of the accuracy of your work, the prospects of your plan, and how much progress you've made while problem-solving process.

In many MEGS questions, for example question 16 on scaling and weight or question 23 on an equation from an English sentence, students in interviews generally applied an incorrect solution method and stopped, apparently confident and satisfied with their work. With small prompts from the interviewer, which (perhaps inadvertently) indicated there might be more to think about, most students were able to revise their answers and answer the questions correctly.

One goal of the MEGS is to measure to what extent students learn to provide their own metacognitive prompts, employing various methods to test the reasonableness of their answers. It seems likely that an intuition that the answer may not be correct is a good starting point for this. If students generally believe their answers are correct when they aren't, they might never get this important prompt and revisit their work. To test the relationship between confidence and MEGS accuracy, we measured student confidence at the end of the MEGS.

MEGS question 32 asks students what percentage of their answers they believe are correct. A few students do not understand the question. For example, one student explained in comments that they thought the question served no purpose; they believed all their answers were correct or else they would not have chosen them, and so the responded with 100 percent. About 5 percent of students respond to this question with a confidence of 0 percent or 100 percent, both presumably inaccurate representations of their beliefs, but we will proceed with these responses included because we cannot categorically decide which students interpret the question incorrectly, and which sincerely believe they answered all questions correctly (so far, one student has done this) or all questions incorrectly (no students have done this so far).

We can compare student confidence in their answers to their actual performance to determine whether students are "well calibrated", meaning they have accurate beliefs about how accurately they answer questions. When students predict they got more answers right than they actually did, they are "overconfident", if they predict they got fewer answers right than they actually did, they are "underconfident".

Calibration has been studied in fields such as prediction (e.g. of weather, sporting events, or elections)[Silver, 2012] and estimation (especially in business)[Hubbard, 2014]. Much of the work on calibration assumes that being well-calibrated is a unitary trait - a person is well- or poorly-calibrated generally. However, some research has found that people can be very well-calibrated in one domain (especially a domain of their expertise) while being poorly-calibrated on more general tasks, or tasks they are less familiar with.

We believe that studying student under and over-confidence may synergize well with work on metacognition and development of expertise, and potentially with other fields. Here, we briefly report on calibration in Phys 131/132 students.

Being well-calibrated is not a celebrated goal in PER instruction. Measuring student confidence is not yet a common practice in administering PER concept inventories, but has been done research. Lindsey and Nagel [2015] measured student confidence in introductory physics on both exam and FCI questions. They confirmed the "Dunning Kruger" effect that students with lower scores were generally overconfident, whereas students with higher scores were accurate or perhaps slightly underconfidence in their beliefs.

In general, 131/132 students are overconfident on the MEGS, but only mildly so. For example, in my class, on the pretest, students on aggregate believed they had answered 60.2 percent of questions correctly when their actual accuracy was 52.1. On the postest, their confidence was 57.4 compared to 48.5 percent accuracy.

Over all, students have been overconfident in 14 test administrations and underconfident in 1. Student overconfidence on the MEGS is minor. This provides a baseline of comparison. In mechanics, for example, a great deal of work has focused on "misconceptions", "preconceptions", etc. that students intuitively believe, but are incorrect. Indeed, Lindsey and Nagel [2015] found that students were more confident in FCI answers than answers to final exam questions.

It may be worth revisiting how well-calibrated students are on the MEGS if, in the future, a course has shown to consistently improve student performance on it. Would such a course also improve calibration?

4.8 MEGS impact on classroom practices

One chief finding from early administrations of the MEGS is that, at least in NEXUS Physics, students make very little improvement on the constructs tested by MEGS over the course of a semester. Since then, some Phys 131/132 instructors at UMD have used the MEGS results to inform classroom practices and homework assignments while others have not. So far the results are ambiguous or moderately-successful.

When I taught Phys 131 in fall 2017, I kept the MEGS results in mind, and used them as both general and specific guidance in constructing classroom exercises. For example, I and the other instructor knew that very few students answered MEGS question 16, on scaling up a person to a new height and finding their new weight, correctly. The first homework assignment included and equivalent problem on scaling up a statue of Testudo, the UMD mascot, and I returned to the topic of how mass scales with linear dimension on several classroom exercises, including two quiz problems, and it re-appeared on future homework problems. This was guidance on a very specific topic, but I also included more generally significant work on understanding scaling relationships, for example including sections on homework problems on how the acceleration we measure of an object in a video played in slow motion scales with the playback rate, and relating the result to the units acceleration. This sort of guidance from MEGS was more general - I knew that students could easily perform the correct computations on the scaling problem if they recognized their necessity. Problems like the acceleration problem were intended to prime students to be aware that scaling relationships are not always linear, and that examining units is a reliable and effective way to understand them.

MEGS question 16 did show improvement after the semester, with students in my class moving from 3.6 to 9.2 percent correct on that question, despite overall accuracy and effort on the MEGS going down over the semester. Students in the other section improved from 4.2 to 16.7 percent on the same question, although at the beginning of the semester, they were not graded based on accuracy, whereas at the end of the semester they were. Meanwhile, students in 132 in the same semester went from 6.0 to 6.8 percent accuracy on the question, and in Spring 2016 Phys 131 students went from 4.0 to 4.1 percent on the same question while working with an instructor who has not seen specific MEGS questions.

Overall performance in question 16 is still low. In general, classrooms informed by MEGS results are not yet achieving large gains. My class declined from 52.1 to 48.5 percent accuracy over the semester. This indicates that MEGS can serve a purpose in the physics education community; if a new class (or a new revision or administration of NEXUS Physics) shows significant gains on MEGS, that finding is more significant because we now know that this is difficult to accomplish.

Some classes have registered gains. For example, in spring 2017, MEGS scores improved from 61.1 to 67.9 percent in a class taught by an instructor who examined MEGS performance (normalized gain of 0.17). However, we have experienced some difficulties in communicating between researchers, instructors who meet students in class, lab instructors who directly meet with TAs, and TAs who give instructions for an supervise taking the MEGS. Especially because there are usually three sections of Phys 131/132, all of which may have different policies surrounding participation in and credit for the MEGS, it is difficult to guarantee that students were taking the test under similar conditions. Students reported similar, but slightly higher effort at the end of the semester in the case of gain described above (from 1.9 to 1.8 on a scale of 1-5 with 1 being "I gave it my best effort" and 5 "No effort".) In another case with gain of 0.21 (same instructor, Fall 2017), student effort increased from 2.8 to 2.2.

4.9 Future Work

The MEGS has been administered several times, and met a number of basic standards used in PER. In my opinion, it serves a purpose in providing useful information to instructors interested in mathematical problem solving in the IPLS context, but it is not yet completely validated, and there are outstanding issues with the test. This section describes our plans for future work on the MEGS.

4.9.1 Expert Validation

With a new concept inventory, it's important that the questions are clear, the answers are correct, and the questions are well-aligned with the test's goals. One way of verifying this sort of conceptual validity is by recruiting a panel of experts. The experts read the questions, highlight any ambiguities or flaws in the questions, identify the correct answers, and judge how well the questions align to the test's objectives. In the future, we intend to assemble such a panel of experts from former instructors of IPLS courses, researchers and course designers of IPLS, and other experts in use of mathematics in physics problem solving. After obtaining expert feedback, we will revise the MEGS to fix any content validity deficiencies this study discovers.

4.9.2 Re-evaluation of distractors

For the most part, we developed distractors for the MEGS informally. We brainstormed what we expected to be common distractors, and in some cases found important distractors from interviews. However, we have not taken a completely systematic approach to identifying the best distractors to use in the test. Not all MEGS questions are amenable to direct research on what distractors to create (for example, MEGS question 4 is intended to be solved by determining which of several formulas has the correct units for a surface area; there are infinitely-many possible correct and incorrect answers), but even for these questions, we can learn more about how the distractors we choose and working and search for principles to use in designing distractors.

We conducted interviews in which some students answered MEGS questions without the answer choices available. When we noticed students choosing answers different from those in our list, we sometimes included these as new distractors. However, we didn't analyze this entire body of interviews in a systematic way. Additionally, we conducted only six such interviews, covering only about a quarter of the test each time. This doesn't provide enough data to find all common distractors.

In the future, we plan to administer the MEGS again to large classes, but without answer choices. We'll give each student only a subset of questions on the full MEGS to avoid question fatigue, and ask students to show their work in detail, a task required on all midterm and final exams in Phys 131/132. Analyzing these responses will allow us to build a much fuller picture of common incorrect answers on MEGS questions, while also providing data on the process of answers MEGS questions outside of an interview context.

4.9.3 Elimination of questions

In test theory, tests like the MEGS attempt to measure students "optimal performance"[Crocker and Algina, 1986]. The test results will be most meaningful and informative if students give their best effort, and take it in an environment conducive to that. If students are giving only partial effort, especially different effort between the first and second administration of the test, the conclusions drawn from the MEGS are suspect.

This may be more of a concern with MEGS than most concept inventories because the MEGS is a hard test, in the sense of requiring considerable effort (as opposed to the sense of having a low number of questions answered correctly). Most questions on common concept inventories, such as the FCI, don't require detailed calculations or in-depth problem-solving, including planning and metacognitive monitoring of the problem-solving process. (This is not a denial that metacognition is important in solving FCI problems, where it likely plays a roll in recognizing that the student is relying on pre-instruction intuitions to arrive at an answer, and revising that to rely on the principles of mechanics instead.)

As discussed in section 4.6.1, students often perform worse on longer MEGS questions, an indication that they may not be motivated to read and put together all the pieces of these questions. Additionally, when instructors gave students more course credit the more MEGS questions they answered correctly, student performance on the MEGS shot up. Typical MEGS performance is about 50% in 131/132 when receiving participation credit and about 60% when receiving credit based on score; an effect size (Cohen's-d) of approximately 2.

Additionally, many students self-report lower effort on the MEGS at the end of the semester than at the beginning, and comment that they didn't solve all problems on the post-test MEGS because they remembered the question from earlier. Although MEGS scores generally do not improve significantly over the course of a semester, the standard deviation of MEGS scores reliably increases. This is consistent with there being some gain in the constructs underlying MEGS, but also some students who put in very little effort at the end of the semester, leading to both more high scores (from students who learned to apply epistemic games effectively) and more low scores (from students who answered questions at random or gave only partial effort). This suggests that the MEGS would be a more effective instrument if it elicited higher and more consistent student effort.

In the comment section at the end of the MEGS, students often say the test is very long. Fatigue appears to be a major issue, suggesting that shortening MEGS questions and eliminating questions could improve its validity as a measure of student use of math problem-solving epistemic games.

We plan to study the effect of using a shortened version of the MEGS, using statistical techniques to determine, for example, which questions contribute the least information to the MEGS results (i.e. we can find the entropy of the MEGS responses, and the entropy conditional on each individual question, to determine which questions are contributing the least information). Classical test theory measures such as the point-biserial coefficients of the questions, and the clustering work discussed in chapter 5 will also inform our work on shortening the MEGS.

After drafting a shortened MEGS, we can compare it to the full MEGS in field tests to decide on the value of the shortened form.

Chapter 5: Comparing factor analysis and network-based clustering methods in analyzing multiple-choice tests

5.1 Introduction

In chapter 4, I described the creation, validation, and early analysis of the MEGS survey. While working on this project, I became interested in the value of clustering and looking at distractors for both formative assessment and test validation. The link between MEGS questions and the constructs they attempt to measure is subtle. The MEGS does not directly ask students to evaluate the dimensions of some expression, for example, but instead asks a question that is made significantly easier by taking a dimensional analysis approach. I didn't know whether students who were answering questions correctly were doing so because they were learning to apply the techniques we were interested in. Also, in chapter 2 I described repeatedly watching students in interviews try out creative and expert-valued problem solving techniques, only to fail to get to the right answer in the end. Could the MEGS be generating many false negatives? In general, how were students seeing the questions, and how was that different from how we saw them?

I would usually have tried to answer these questions purely through qualitative research. I'd bring students in for interviews and ask them to solve problems, or ask a class of students to write out their work as they solved problems. I could then interpret the rich data set this provided to learn how students saw the MEGS.

That's part of what I did in chapter 4, but the data set is necessarily limited in size this way for interviews, and for artifacts generated by a large class the analysis is daunting due the quantity of responses. It would make sense to see whether analyzing just the multiple-choice test results could also speak to the questions of where the students were in their use of problem-solving techniques and how valid the test was. After all, we explicitly built the MEGS around four e-games

- dimensional analysis
- examining extreme cases and special cases of formulae
- estimation
- mapping abstract symbols to physical meaning

Wouldn't it make sense to check whether these same four e-games pop out of the survey data?

At the time the first MEGS results were coming in, I was taking a course in nonlinear dynamics [Girvan] with UMD's Michelle Girvan, and I thought the network-based techniques presented there had the potential to grant some insight here. We could model a class worth of test results as a network, with students and answer choices as nodes, and look for connections between the nodes. This could tell us what questions were related to each other in students' answering habits, and might yield results that either confirmed the qualitative work, challenged it, or showed new areas to investigate.

As a class final project, I proposed two methods of analysis for multiple-choice single-response concept inventories - factor analysis and community detection on a network. In the intervening period, researchers published methods analyzing concept inventories using both these techniques in ways that were substantially similar to what I had proposed, so the techniques in this chapter are either direct replications of those results or conceptually very close to them. What I try to add here is some validation of the techniques by comparing them to each other on simulated data sets to see how the insights they produce differ or agree. I also apply the techniques to the MEGS and interpret the results in light of our qualitative understanding of the MEGS.

We also built a second survey, the Mathematical Attitudes and Expectations Survey (MAX). It probes student opinions on how to study, how interrelated different academic fields are, and issues related to student identity as math users. It contains many sets of questions which we intend to ask very nearly the same point, although we did not build the survey with an explicit clustering of questions in mind. Similar surveys, such as the Maryland Physics Expectations Survey, are commonly broken into distinct clusters to be analyzed separately in the physics education research literature. Although I expect there's a lot to be gained by studying the MAX data, especially in relation to the MEGS data, I'll analyze only MEGS data in this chapter. This allows closer comparisons to most published work, because concept inventories and attitude surveys have usually been analyzed separately in

PER. Future work could look to gain new insights from the MAX data.

In both tests, there are sets of questions that are conceptually similar. However, just because questions appear to test similar concepts or ask about similar ideas does not mean that students will give similar answers. In interviews in which students explain their reasoning while solving physics problems, we have seen that they possess “knowledge in pieces”. That is, if a student answers a question about a sled sliding over ice at constant speed by saying it must have a constant force on it, this does not mean the student has a single, wrong theory, like “force requires motion”. The same student might answer an equivalent question about a meteor in outer space correctly, or might say that a constant force causes constant acceleration on a sled over ice. Students answer questions by activating “resources”, small, compartmentalized bits of knowledge or viewpoints like “bigger effects require bigger causes” or “always trust equations over intuition”. Different superficial features of a problem may call up different resources in students, so that they give apparently contradictory answers. The coherent theories we see in physics experts are the result of well-trained relationships between many resources, which, taken as a whole, appear like a single, coherent theory.

To better understand the ways that students view problems, we conduct interviews in which we have students “think aloud”, explaining their thought process as they solve problems. However, it is only possible to interview a small sample of students in this way. By analyzing the results of surveys students take, we may be able to learn about how entire classes of students are approaching problem-solving by detecting similar problems and reverse-engineering the features which cause the similarity.

5.2 Analyzing concept inventories with a focus on distractors

When hundreds or thousands of students take the same multiple-choice test, they are generating a large data set. This data might yield quantitative insights into how students think, complementing the qualitative data generated by field observations and student interviews, but only if we find effective ways to analyze it.

One way to get new types of insights from tests is to analyze the distractors. Distractors are the incorrect answer choices given alongside the correct answer choice on a multiple choice question. They're usually chosen to be "attractive", so that people are likely to think they make sense and could plausibly be correct.

Most published analysis of test data reduces a student response to a test item from one choice among (say) five options, to a binary correct/incorrect response. But over the course of a semester, students could go from picking one type of incorrect answer to a different type of incorrect answer. Depending on how that process worked, it might represent real progress and have important implications for instruction, but current test analyses would never pick up on it. (I will discuss two exceptions to the rule that PER analyses ignore distractors, model analysis and module analysis, later in the chapter. Below I review several other analyses that get information from distractors, or which are otherwise relevant to the approach of this chapter.)

Historically, physics education researchers have sometimes used factor analysis, a technique related to principal components analysis, to look for clusters of similarly-answered questions in hopes of finding that all questions can be understood as a linear combination of a few underlying "concepts". This technique begins by classifying question responses are correct or incorrect, and has no ability to learn anything from which incorrect answers students chose. However, recent work by Scott and Schumayer [2017] expanded factor analysis to include distractors. They found significant factors between distractors on the FCI and interpreted the result as evidence that students have coherent, incorrect theories of physics that they use when taking tests, as opposed to finding their incorrect answers by random guessing. I'll examine Scott and Schumayer's version of factor analysis alongside other analysis techniques later in this chapter.

Also recently, researchers have begun to use network-based techniques for community detection to discover clusters of questions and clusters in the potential responses to questions in test results. The most notable example here is "module analysis" introduced by Brewe et al. [2016]. They also analyzed FCI data and found six clusters, or "response modules" among the incorrect answer choices on FCI. They were able to find interpretations for each cluster in terms of the physical ideas behind them, for example an "impetus module" in which two particular question responses were central. In the impetus module, question responses deal with the idea that object carry a force along with them after being influenced by something that no longer touches them. For example, a tennis ball might be thought of as carrying the force of an impact with the tennis racket along with it as it travels towards the net.

Scott and Schumayer [2018], in a follow-up to their factor analysis work, introduced another network-based technique related to that of Brewe et. al. Their results were quite similar; they also discovered impetus related factors and a few very central nodes to their network. However, their analysis requires inputting some *a priori* factor structure from their earlier work, so we won't attempt to replicate it or examine it in great detail here because the same structure is not already well-recognized in the MEGS.

Factor analysis and network module detection are the two methods I'll focus on in this chapter, but these are not the only methods researchers use to get information from distractors selected in multiple choice tests.

When two students take a test, they can be given a score based on how similar their answers are, for example a number that is just the number of questions on which they selected the same answer. This could be high either because both students had very high scores, or because they selected the same wrong answers as each other. With a "distance" defined between any two students in this way, methods such as k-means clustering can identify groups of students in a class. It might, for example, separate those students who have an "impetus" view of mechanics from students with a Newtonian view of mechanics in FCI data. For example, Battaglia and Fazio used k-means clustering to analyze FCI data and found that both which questions students answered correctly and which incorrect answers they chose were important to the resulting cluster of students. I find results like these promising; cluster analysis might be a fruitful avenue for future investigation into the MEGS, although I omit it due to time and space limitations and because its tack of clustering groups of students is fundamentally a bit different than my research thrust.

Another approach to identifying clusters or factors in concept inventories is multi-dimensional item response theory (MIRT). Stewart et al. [2018] used this method on the FCI, comparing their results very closely to previously-published results from factor analysis. They believe MIRT is better than factor analysis at determining the optimal number of factors in a test, for example. As yet, MIRT doesn't look at distractors and its mission is very close to traditional factor analysis, so I haven't attempted to extend it to include distractors or applied it to the MEGS.

One other thing analyzing distractors can potentially do is point to false positives in tests. For example, if a slew of incorrect question responses are clustered together with a single correct question response, that single correct question may be a false positive. It could be that test-takers are approaching the question for that lone correct answer choice with an incorrect general theory of model or mode of reasoning. Sometimes, incorrect theories happen to give the right answer, but that's undesirable for a test.

While I'm not aware of quantitative analysis that detects candidate false positives in this way, work such as Yasuda et al. [2018] has investigated the rates of false positives with other methodology. They write sub-questions extending FCI questions to determine false positive rates, and show that the false positive rate varies considerably by question. If false positives remain an major issue on such well-validated tests as the FCI in 2018, more than three decades after the test was created, then they are likely an issue with many other concept inventories as well, and methods that can find likely false-positive candidates will contribute to our understanding of what scores on those tests mean.

5.2.1 Why I'm interested in new ways of analyzing concept inventory results

In physics education research, researchers often use multiple-choice tests such as the Force Concept Inventory (FCI) [Hestenes et al., 1992a] to assess how much students have learned, and thereby assess the quality of instruction. For example, Hake [1998a]'s large analysis showed that students' conceptual understanding improves dramatically more under interactive engagement than under traditional teaching methods, as measured by class average score on the FCI. It's a one-dimensional result. What comes out of an entire class taking the FCI twice is just a single number from -1 to 1, with 0 representing no average learning over the semester and 1 representing all students getting perfect post-test scores.

More-informative analyses of test responses could do a few different things for PER and physics teaching practice. First, they might have affordances for viewing concept inventories as formative assessment, i.e. assessment that exists not to evaluate, but to help improve the learning environment. Tests can say more than how well the students did or how effective the instruction was. They can also provide insight into ways that students are thinking about problem solving or specific places that students are making errors. That feedback guides instructors in adapting their classes to best serve students based on where they are.

The most famous concept inventory in physics, the FCI, was originally intended by its authors as a tool for formative assessment, but is today mostly used as summative assessment tool not of students, but of instructional strategies. So papers may, for example, cite FCI gains under a reformed curriculum and a traditional curriculum in order to claim that the reformed curriculum lead to better learning because that class had higher FCI gains. This is an important role, and FCI and similar concept inventories have contributed to important reforms in physics education, but it's only one role. We can get more from the time and effort students spend in taking concept inventories and researchers and instructors spend on creating, validating, administering, interpreting, and comparing them.

For concept inventories to be useful for formative assessment, they need to give feedback to instructors about where students are in their learning process. For example, we could break the FCI down into several different sub-tests that test different skills, report student scores on each of the subtests, and use these scores to decide which concepts in mechanics will receive the most instructional time over the remaining parts of the course.

With data from distractors, more nuanced forms of formative assessment become possible. For example, teachers might learn what specific misconceptions students hold, or what specific contexts tend to trigger their misconceptions. Model analysis uses data from distractors to show teachers the extent to which their class holds coherent, incorrect views about physics (giving answers that are wrong, but consistent among each other), as opposed to giving contradictory answers to seemingly-equivalent questions (and getting the same overall score). We might also hope that learning which question responses are related to each other in student responses to a test, coupled with reading those items carefully, could give us new insight into how students are thinking.

5.3 Specific Aims

In this study, I want to compare the effectiveness of factor analysis and network module detection for identifying known clusters in test questions and granting insight into student thinking on the MEGS surveys. To this end, the specific questions I will try to answer are

How well do factor analysis and module detection detect clusters we know are there?

> One way to judge the closeness of a detected cluster to the source cluster is the variation of information. When detecting clusters using factor analysis and module detection in simulated test results where the true clusters are known, which has the lower variation of information?

How do factor analysis and module detection scale?

> When simulating test results, we can change the number of test takers, number of questions, and number of different concepts. Does one test always outperform the other, or does the superior test depend on number of test takers, etc.?

Does the MEGS survey cluster on the dimensions along which it was designed?

> One might not expect the MEGS survey to show such clustering at the be-

ginning of the semester because students do not yet know, or do not yet recognize, the various tools around which the questions were designed. However, by the end of the semester, it's possible that some students will have learned techniques such as dimensional analysis and consistently apply them, whereas other students will not. In that case, the questions with dimensional analysis will strongly correlate and form a cluster.

5.4 Research Plan

This is a computational chapter. Here I'll describe the different type of tests I ran on real and simulated data. I wrote the code for this chapter in R, and I can provide code on request.

5.4.1 Tests on Simulated Data

The first part of this project will be a theoretical investigation into how factor analysis and network methods perform on simulated data sets which have a known underlying structure.

In section 5.5, I outline four different methods of simulating a class of students responding to a multiple-choice, single-response test. Some of these methods have inherent structure available to be discovered, and others don't. They vary in the subtlety behind the structure as well. In the "model analysis model" it is fairly clear which answer choices belong in a cluster together. In the "Bayes nets" model the edges are murky, even though connections between questions and question responses certainly exist.

Once these models are in place, I simulate taking a test by taking each student and each question and having the student answer the question based on the probabilities generated by the model. This generates data which may be sent to algorithms for factor analysis or used to build a network, then analyzed for community structure.

Questions testing the same concept should cluster together. We can test how well they did by computing the variation of information between the clustering resulting from detection algorithms and the known structure of the questions. In this part of the project, we determine how well the algorithms detect clusters as a function of the number of clusters and number of students. For each model where it is relevant, we plot the average variation of information between the true and computed clusters as a function of the model parameters.

5.4.2 Analysis of Data from Students' Tests

After analyzing the theoretical ability of factor analysis and network clustering techniques, I will use them on the data generated by students in Maryland's physics 131 between Fall of 2015 and Fall 2017. The biggest task in this area is interpretation of the results.

5.4.2.1 MEGS

For the MEGS test, detecting clusters of similarly-answered questions may indicate which of the mathematical concepts we designed the test for exist as categories in student answer patterns. We do not expect that questions on similar conceptual material will necessarily cluster together because students do not answer them in similar ways. For example, MEGS questions 1 and 10 both ask students to convert a measure from one unit to another. However, students usually answer question 1 by constructing a sort of chart. While this strategy would be appropriate for question 10 as well, few students use it, instead using a method based on recall of facts and "moving the decimal point". The different methods come about because in one case, it's only necessary to multiply or divide by powers of ten, whereas in the other case, different numbers are used and students need a chart to keep track of them.

It's unlikely that these two methods have especially strong correlation beyond the correlation all questions have based on students' general background in math and physical science, so we would not expect these two questions to cluster at the beginning of the course. If students learn to better recognize unit conversion problems as a single category over the course of instruction, these questions may then cluster together at the end of the semester instead of the beginning, or the pre-post improvement in these questions may cluster.

Deborah Hemingway and I have conducted a series of interviews with students in which they solve selected problems from the MEGS. The goal was to analyze the students' solution strategies and how they identify and categorize the problems. We were also interested in the students' stance towards knowledge while solving the problem (e.g. the question should be answered via an algorithm, via memorized solutions, via constructing an answer, via eliminating wrong potential answer choices, etc.). Finally, we wanted to know whether they find the question intimidating, interesting, or boring.

After identifying clusters of questions, we can cross-reference the clusters to our interviews to attempt to identify common features of the questions that cluster together. These clusters may not be apparent simply from reading the questions from the expert point of view.

5.4.2.2 MAX

Attitude surveys have often been broken into separate dimensions for analysis. For example, the MPEX survey questions probe student attitudes towards independence, coherence, concepts, reality link, math link, and effort [Mccaskey et al., 2003]. These dimensions are identified by the test creators. I am not aware of any attempts to validate attitude surveys by showing that these clusters emerge from the data itself.

For the MAX, our future efforts could compare the results of algorithmic cluster detection to the clusters identified by people reading the questions. If these clusters closely agree, we will have validated analyses that look at student changes on these clusters independently of each other.

Further, we could look for connection between MAX and MEGS results. For example, how well do different clusters of MAX responses predict MEGS improvement over a semester?

5.4.2.3 Further Tests

The Maryland physics education research group has collected extensive data from a number of multiple choice tests, including the Force and Motion Concept Evaluation, MPEX, and several smaller tests. If analysis of the MAX and MEGS proves fruitful, it should be relatively easy to analyze these tests as well for similar reasons.

5.5 Four models of how students answer test questions

As the first part of my investigation, I'd like to know whether factor analysis and modularity maximization on a network can detect signals that are known to exist. I'll do this by creating the data artificially with a simulation built on the type of signal I want to detect.

For example, if I create data where each question is coded to come in one of four distinct types, can factor analysis and modularity maximization correctly create four question clusters, each with the correct questions in them? How much data do they need to do this with a given reliability? Which technique is more effective for which models of student test-taking behavior?

Answering these questions provides a measure of validation of the techniques themselves, so that the results of applying the techniques to the real data are more meaningful. For example, if I explicitly construct data with three underlying clusters, but a certain data analysis technique, when given a sample of 1000 students, finds anywhere from two to seven clusters in the data, I know not to trust the number of clusters detected when I apply that technique to real data.

In the subsections below, I'll describe four models for how students respond to multiple-choice test questions, each based on an established method of analyzing multiple choice test data: the Rasch model (closely related to item response theory (IRT)), the factors model (related to factor analysis), model analysis, and Bayes nets (as used in Evidence-Centered Design).

5.5.1 The Rasch Model

The Rasch model is a simple model that imagines all questions on a test are fundamentally the same, except that some are harder than others. If the entire test is designed to measure a single underlying construct, the Rasch model might be appropriate. The basic assumption of the Rasch model is that every question has a difficulty and every test-taker has an "ability". When a test-taker answers a question on the test, they will get it right with a certain probability, so

$$P_{\text{correct}} = f(\text{question difficulty}, \text{test-taker ability})$$

We'll denote the questions difficulty by δ and the question number is indexed with a subscript i, so δ_i is the difficulty of question i on the test. The ability of the test-taker is denoted by β and the test-takers are indexed by a subscript n, so the ability of test-taker n is β_n. So the Rasch model posits there is some function

$f(\delta_i, \beta_n)$ that gives the probability to answer a question correctly.

The function should be increasing with increasing ability, and decreasing with increasing difficulty. Beyond that, the choice of the function is in theory arbitrary because different functions simply correspond to different rescalings of the ability and difficulty variables, and constructs such as "item difficulty" are a priori ordinal data only.

To make progress on what function to use (or how to scale ability and difficulty), the Rasch model makes a few more assumptions.

- abilities and difficulties are real numbers and vary on a scale from $-\infty$ to ∞
- $f(\beta_n, \delta_i) = f(\beta_n - \delta_i)$, i.e. only the difference between the ability and difficulty affects probability
- $f(0) = 0.5$
- Increasing the test-taker's ability by 1 is equivalent to giving the test-taker some piece evidence about the question with fixed odds ratio of e.

5.5.1.1 Evidence and odds ratios

The first three assumptions described in the previous section are straightforward, but the last requires some clarification. We can imagine it as if having greater ability is equivalent to receiving some hints on the problem. Suppose, for example, that for some test-taker and some problem, the test-taker's chance to get the question right is 66.67% based on their ability in the test's underlying construct and how hard the question is. For simplicity we will also imagine that the question is true/false (this assumption isn't necessary).

We might give the test-taker a hint to increase their chances. This might be telling them a piece of information useful to solving the problem, pointing out a mistake in their work so far, or giving a metacognitive prompt such as, "can you think of an example problem similar to this one, but that you've worked out previously, to use as a comparison?"

For the sake of building a mathematical model, we can abstract the details of these hints away and replace them with a box with two light bulbs: one for true and one for false. The box lights up one of the light bulbs. It usually lights up the light bulb corresponding to the correct answer to the question, but sometimes lights up the incorrect answer at random. So if the test-taker sees the "True" light bulb light up, they take this as a hint that makes them more confident in the answer "True", although they know it's still possible that the answer is "False". As the test-taker receives more and more hints, they become more and more confident (assuming the hints don't happen to split nearly 50-50 between True and False), so that after receiving many hints, their chance of getting the answer right is very high - as high as another test-taker with far higher ability, but who didn't receive the hints.

For a given gap in ability, there is a certain number of hints the lower-ability test-taker needs to receive to have the same chance of getting the answer right. Assumption 4 on the list says that this number of hints to achieve parity is proportional to the difference in abilities between the two test-takers.

To work with this story mathematically, it's useful to introduce odds in place of probabilities. Before the first hint, the test-taker's odds are 2:1, meaning that a large ensemble of test-takers with the same ability level would see two get the answer right for every one who gets it wrong. For example, with 300 test takers, 66.67%, or 200, would get the answer right, while 100 would get it wrong. (On expectation; the actual number would vary due to the random nature of the model.)

Let's suppose that the hint box gives the right answer 90% of the time, so it represents odds of 9:1. The test-taker was previously leaning towards the answer being "True". If the hint light bulb "True" lights up and the test-taker takes this into account the hint using Bayes' theorem, their new odds become 18:1; i.e. the original odds are multiplied by the odds associated with the hint to get the final odds.

If they are given a second hint, which is also "True", their odds become 162:1, and if they receive a third hint for "False", their odds go back to 18:1. The simplicity of this calculational scheme shows the benefits of using odds; using probabilities leads to more complicated calculations.

The iterated multiplication that comes from receiving multiple hints suggests looking at logarithms to turn the multiplication into addition. For a probability p, the log odds are defined to be $\ln\left(\frac{p}{1-p}\right)$. In our example, the test-taker starts out with log-odds of $\ln 2 \approx 0.69$. Each "correct" hint (light bulb that light matches correct answer to the question) increases their odds by $\ln 9 \approx 2.20$, while each incorrect hint decreases their odds by $\ln 9$. After receiving one correct hint, the test-taker's new odds are $\ln 2 + \ln 9 \approx 2.89$.

In the Rasch model, having higher ability is considered equivalent to receiving more hints. Specifically, an increase of ability by 1 improves the test-taker's log odds by 1, so it is equivalent to receiving a single hint with odds of $e : 1$, or a hint with probability $e/(1+e) \approx 0.731$ to be correct.

This is sufficient to complete the function f and define the Rasch model. We know that $f(0) = .5$, or 1:1 odds. Then $f(1)$ is the probability to get the question right after receiving one hint, or $f(1) = e/(1+e)$. To determine $f(x)$, we note that the log-odds associated with the probability should be equivalent to receiving x hints, each of which increases the log odds by one, so the log odds should be x. Then solving $x = \ln(p/1-p)$, we arrive at

$$f(\beta_n, \delta_i) = \frac{1}{1 + e^{-(\beta_n - \delta_i)}}$$

This defines the Rasch model for how a student answers a test question. To simulate data with the Rasch model, we first define a few parameters:

- N: the number of test-takers
- I: the number of items (questions) on the test. In this chapter $I = 30$.
- p_β: the distribution from which the abilities of test-takers is randomly sampled. In this chapter we will use normal distributions with a mean of zero and standard deviation σ_β.
- σ_β: the standard deviation of the distribution of test-taker abilities
- p_δ: the distribution from which the difficulties of questions is randomly sampled. In this chapter we will use normal distributions with a mean of 0 and a

standard deviation of 1.

I've specified several of these parameters here because I won't search through different values of them when evaluating the analysis methods against the Rasch model. The number of parameters is large, potentially arbitrarily large (depending on the allowed distributions p_β and p_δ), but it shouldn't be necessarily to explore every corner of parameter space, especially with a model chosen for the purpose of not having any special structure.

After specifying the above parameters, we can simulate a class of test-takers taking a test via the following procedure:

1. For each test-taker n, sample once from p_β and assign the result to β_n
2. For each test item i, sample once from p_δ and assign the result to δ_i
3. For each test taker n and test item i, generate a number from 0 to 1 independently, each uniformly at random. If the number is less than $f(\beta_n - \delta_i)$, call the test-taker's response to that item correct. Otherwise call it incorrect. Choose from among the incorrect answers randomly to be the test-taker's response to that item.

This allows us to generate a simulated class worth of responses under the Rasch model, for a certain model of how student abilities and question difficulties are distributed.

It would be possible to analyze more general distributions of student ability and question difficulty, but normal distributions are a common default distribution to study, and we choose them here to reduce the parameter space we are searching in this study.

The Rasch model presents the idea of a very unitary test - one that tests a single underlying construct. Thus, factor analysis or clustering algorithms are expected to return no significant results when analyzing Rasch data. If we run factor analysis or a clustering algorithm on real data and see very similar results to what we get when running them on Rasch-model simulations, it's evidence against those tests testing a number of distinct constructs.

5.5.1.2 Distractors in the Rasch model

The Rasch model, and IRT more generally, are only concerned with whether a student answered a question correctly or incorrectly. Simulations using the Rasch model won't let us learn anything about whether students who answer questions incorrectly do so in any sort of systematic way, but the Rasch model's role in this chapter is to provide a sort of null model where the signals we're looking for are not present.

5.5.2 The factors model

The Rasch model assumes that a test tests only a single underlying construct. We can adapt this model to a test for several independent underlying constructs, which I will call the factors model.

The factors model is inspired by factor analysis, in which a matrix of students responses to a test undergoes a change of basis. (Technically, it's not a raw matrix of responses, but a matrix of correlations between student responses to questions on a test.)

Ordinarily, we would view a student's responses (coded as 1 for a correct response and 0 for an incorrect response) to a test as a response to item 1, a response to item 2, a response to item 3, etc. But if we decide that items 1 and 2 are testing the same construct, we might switch from vectors $(1, 0, 0, \ldots)$ to represent question 1 and $(0, 1, 0, \ldots)$ to represent item 2 to a single vector $(1, 1, 0, \ldots)$ to represent both item 1 and item 2. A student response to the "composite item" represented by this vector could be the sum of their responses to items 1 and 2.

As student responses to items 1 and 2 are highly correlated, we would capture most of the information available in their responses with just this one vector. We could still represent the complete results without loss of information by adding a vector $(1, -1, 0, \ldots)$ to represent the extent to which students got different answers on items 1 and 2, but if items 1 and 2 test the same construct, we expect that the coefficients involved with this vector would be quite small, while the coefficients involved with the vector $(1, 1, 0, \ldots)$ would be much larger.

If items, 3, 5, and 8 were also testing the same construct, we could create another vector, $(0, 0, 1, 0, 1, 0, 0, 1, 0, 0, \ldots)$ to capture student response on that construct, and continue in this way until we had covered all the constructs on the test with their own basis vector. By representing student responses in the basis made of these "construct vectors" (not a standard term), we could measure their ability on each individual construct.

This is the essential idea behind factor analysis, although I'll give more detail on the exact procedure in section 5.6.1.

What's important here is the factor analysis is based on the idea that individual test items test a single construct, that the constructs are independent and do not interfere (so that a linear superposition of them is an appropriate description), and that we can group the items of a test onto a relatively-small number of important constructs (also called "factors" in a more general setting, since factor analysis is used many places besides analysis of tests).

This leads to a simple way to simulate a factor-based test, based on the Rasch model, but with C constructs.

- assign each item on the test one of C constructs
- take the items that pertain to a single construct. Think of these as a small, self-contained test, and use the Rasch model to simulate students responding to this test
- repeat the last step for each of the other constructs

The parameters that go into the factors model are the same as for the Rasch model, plus the number of constructs. As for the Rasch model, I'll assume item difficulties are drawn from the standard normal distribution and student abilities are drawn from a normal distribution with mean zero. I'll change the standard deviation of student abilities between runs.

In order to assign the questions on a test with I items among the C constructs, I assign the first I/C questions to the first construct, the next I/C to the next construct, etc. with appropriate integer rounding. This gives all constructs approximately the same number of questions, which tends to reflect the structure of concept inventories, which are often designed to be roughly balanced. The MEGS was constructed with a goal of achieving rough balance between the epistemic games played.

With these procedures in place, we can simulate a class's results when taking a test with multiple factors. If we perform factor analysis on such a test, we should ideally find the same number of large factors as there were constructs in the simulated data. Also, we can test whether factor analysis maps test items onto constructs reliably. This will take some minimum number of students to happen reliably, which sets a lower bound for how many students would need to be in a class for the results of performing factor analysis on real data to be meaningful.

We can also use simulation from the factor model to test other clustering techniques. Do they cluster questions according to their concepts correctly? Do they find the correct number of concepts? Do they require more or less data than factor analysis to accomplish this? This gives us some insight into what other techniques might be able to tell us when used instead of factor analysis, if we have an underlying assumption that the test we are analyzing is a test of a number of different, independent, underlying constructs.

As with the Rasch model, distractors play no role in the factor model except insofar as they change the item's difficulty, and therefore the student's probability to answer the item correctly. Again, we shouldn't be able to learn anything significant about which distractors students choose when looking at simulations of the factor model.

5.5.3 Model Analysis

"Model analysis" is a technique of analyzing data from multiple choice tests, described in a paper by Bao and Redish [2006].

Here, I adapt model analysis to the purpose of simulating data, rather than analyzing it. I'll first motivate model analysis as an analytic technique. Then I'll describe how I see it informing a model for simulating students responding to a multiple-choice test.

Model analysis is situated in the resource framework, wherein students who ultimately learn to answer physics questions correctly do so by reorganizing many small "knowledge elements", or small, irreducible bits of cognition, into coherent structures.

A typical progression for a student learning a physics concept might be that at first, they answer questions in a consistent way, based on a "folk physics" model they have, which is incorrect. This is the stage at which students could be described as "having misconceptions".

As students begin learning material, they examine examples, try out solutions, do mathematical analysis, etc. Their task is to generalize from the examples they're given and parse the explanations they hear until they have built a new, coherent model. However, in the process, their views are unlikely to be coherent.

I'll give an example of what this means one might observe. Consider two questions, based on questions used by Professor Redish in teaching PHYS 131:

> A baseball and a bowling ball are sitting on a frictionless table. Each has a string tied to it. You grab the strings and pull both balls towards you with the same acceleration. On which string do you have to pull harder?
>
> You hold a baseball and a bowling ball up above your head, then drop them. They fall to the ground, and air resistance is negligible. On which ball is the gravitational force greater?

These two questions are what Bao and Redish call "expert-equivalent question" because from an expert's point of view, they are exactly the same. In both cases, the two balls have the same acceleration, so reasoning from $F_{net} = ma$, the one with the greater mass (the bowling ball) must have a greater net force on it.

Before any physics instruction, it may be common that students in fact answer both questions correctly. They appear to be answering with a coherent model because their answers both fit in the same conceptual framework, which may be as simple as "bigger things need more force".

However, in PHYS 131, a common result from asking both these questions is that students will identify the bowling ball as requiring a greater force to be pulled horizontally, but will decide that the gravitational forces on the baseball and bowling ball are the same.

What's happened is that students have learned, against their intuition, that gravity accelerates all objects with the same acceleration. However, they haven't yet encoded fine details of this fact into their conceptual systems, and we may instead hear statements such as Bert's "gravity is independent on mass", which he said in the interview in which he worked on the half-Atwood problem, discussed in chapter 2. Without the distinction of whether it is gravitational force or gravitational acceleration which is independent of mass, many students choose an answer that seems to best match the idea "gravity acts on everything the same", and say that the bowling ball and baseball have the same gravitational force on them.

Later, as students mature in their understanding of Newtonian mechanics, they distinguish force and acceleration and correctly say that the balling ball has a larger gravitational force on it.

It's the middle state, in which students give conflicting answers to apparently-conceptually-identical questions (from the expert's point of view), which inspired the creation of model analysis.

Model analysis takes it as inevitable that we will not be able to capture all the many sensitivities that students have to the precise context in which a question is asked. Instead, all the things that can "throw students off", any irrelevant details which students are likely to cue on for a while, any matter of interpretation of variables that look different in different scenarios, any bits of the physical environment that trigger different types of reasoning, etc. are classified as "context dependence" of student answers and modeled probabilistically.

Specifically, model analysis posits that there are several competing models for how to answer a physics question, for example, an expert Newtonian model, a "there must be a force in the direction of motion" model, and an Aristotelian “things go to their natural place” model. While students are transitioning between novice and expert, their response to a question can be modeled as choosing one of these models at random according to a probability distribution, then answering the question based on what that model says about the situation. The random model selection step is a stand-in for all the small details of context (either aspects of the question itself, or things the student was thinking about just before taking the test, or conditions during the test, etc.) that don’t affect the question from an expert’s point of view, but do affect the students’ responses.

Model analysis does lack one aspect of analysis of a test that the Rasch model has: item difficulty. In model analysis, all expert-equivalent items are equally difficult, and in fact interchangeable. This belies the existing data, which shows that some expert-equivalent questions are significantly easier than others, but it’s beyond my scope here to attempt to modify model analysis to add this extra dimension in.

Bao and Redish say that a student is in a “model state” (a vector representing the student’s probability distribution, or probability amplitude, (because they take square roots of the probabilities) for each possible model used to answer a test item) within a “model space”. When students’ responses to test items are coherent, meaning they always select answers corresponding to the same model (regardless of what that model is), Bao and Redish refer to the student as being a in “pure state”. When students on a particular test have significant probability loaded onto two or more different models, Bao and Redish say the student is in a “mixed state”. I will use this language here, although it’s important to note that, although much of model analysis is inspired by analogies to quantum mechanics (for example, Bao and Redish construct “density matrices” from the outer product of a vector of the square roots of students probabilities to choose various models), the “pure state” and “mixed state” terminology is based on the intuition that students are purely using one model, or are mixing their responses between models. It isn’t a mathematical analogy to the pure states and mixed states of quantum mechanics. Instead, model analysis’s “pure states” and “mixed states” for an individual test-taker are both analogous to quantum pure states, in that they both result in density matrices whose trace is one. Model analysis “pure states” are eigenstates in the basis where individual models are basis vectors; model analysis “mixed states” (from here on I’ll drop the quotation marks distinguishing them from quantum states) are superpositions in this basis. The distinction is important because model analysis does involve something akin to a quantum mixed state - the average of the individual density matrices of each student in the class, which Bao and Redish call the “class model density matrix”. Their class model density matrices cannot be represented as coming from a single model state, unlike single student density matrices.

By looking at the answer choices students selected and assigning those answer choices to various models, model analysis can determine the model state for a particular student taking a particular test at a particular time and context. It then constructs and analyzes the class model density matrix to look at an average over an entire class. This lets one determine whether the class is, on the whole, mostly in pure states or mostly in mixed states. You can’t tell this from an analysis that only determines whether student responses are correct or incorrect; there would be no difference between getting answers wrong 40% of the time because the class consistently uses the wrong model on 40% of the questions and getting answers wrong 40% of the time because the students are in mixed states that load 40% probability onto incorrect models for all the questions.

For an example of the insights this generates, Bao and Redish compared a traditionally-taught and interactive-engagement course (differing in the structure of tutorial sessions) using model analysis. They showed that in both cases, the classes began largely in a pure states with incorrect models. Both classes improved over a semester of instruction, and traditional analysis might simply demonstrate that the normalized gain of the interactive-engagement class was higher. Model analysis showed this, but also showed that the traditionally-taught class moved into a highly-mixed-state class density matrix over the course of a semester, meaning that most students were still cuing off of irrelevant context by the end of the semester. Meanwhile, the interactive-engagement class moved to a predominantly pure state class density matrix, demonstrating that they had successfully taken significant steps not only towards accuracy, but expert-like coherence of model selection as well.

Model analysis coaxes a lot more from the data of a concept inventory administration than a traditional data analysis, or even a factor analysis, does. (Factor analysis, as explained by Redish and Bao in a toy example, won't show when students in mixed states in a class are all behaving similarly to each other, for example, while model analysis will; it will show up in large off-diagonal elements in the class density matrix.) However, model analysis has to pay for this extra insight with very detailed input.

To use model analysis, the researchers must decide which questions are expert-equivalent questions and which answer choices belong to the same model; that is all put in by hand. It's done by analysis of qualitative validation data from students, which allows the researchers to determine what sorts of models students were reasoning from when they selected various item responses. It's also done with expert validation.

Experts are needed to determine which items of a test form an "expert-equivalent" set. So while model analysis is powerful, it's very resource intensive to set up and difficult to validate. It doesn't allow for clusterings of questions to arise from the data of question responses themselves, and so it isn't a tool for discovering unexpected connections. (It does, however, have an inbuilt feature to suggest that such a discovery is needed by other means. It collects uncategorized item responses in a "null model", and when enough probability mass loads onto the null model, model analysis suggests creation of new models via further qualitative analysis, or via other quantitative techniques, such as those I explore in this chapter.)

Because our expert validation of the MEGS is not complete and our present analysis of student validation data doesn't attempt to categorize distractors by a small set of coherent competing models, we aren't able to use model analysis here to analyze MEGS data. However, it does provide a model we can use to simulate data in a way different than the Rasch model. The method is similar to the factors model, but more detailed in that it will give input into which incorrect answer students choose.

To simulate students taking a test according to the model analysis model, we

1. partition the test into sets of “expert equivalent” test items

2. for each set of expert equivalent test items, choose a number of competing models, including a “null model” to catch extraneous responses

3. for each item on the test, partition its responses between the models for that item’s set of expert-equivalent items

4. for each student in the class and for each set of expert equivalent items, construct a model state (probability distribution to select the various competing models)

5. for each item on the test and for each student in the class, simulate the student responding to the item by sampling from the student’s model state-specified distribution for that item’s expert-equivalent set of items

My procedure for partitioning the test into expert-equivalent test items will be the same as for the factors model.

For convenience and to limit the parameter space to search through when testing analysis methods with model analysis, I’ll always use four competing models for each expert-equivalent question. Choice A will correspond to the expert-like model (and correct answer), B to the second model, C to the third model, D to the fourth, and E to the null model.

This considerably simplifies model analysis, because in some situations, there could be false positives in which a naive model gives the correct answer; I gave an example of this with the bowling balls and baseballs earlier in this section. Also, sometimes in model analysis multiple answer choices are assigned to the same model, and the number of models changes from situation to situation.

I will assume the class average 50% on the test, which is roughly consistent with almost all runs of the MEGS survey. For each student and for each set of expert-equivalent questions, I'll assume that the student's log-odds of answering the question correctly are normally distributed with a standard deviation of 1 and a mean of 0. I'll sample from this distribution to give the student a probability of choosing the correct model on that group of questions.

To further allow me to simulate data with model analysis, I'll introduce one more parameter: the class mixedness.

The class mixedness, m, is a number from 0 to 1 that determines the extent to which students are in mixed states. If the class mixedness is low, then students who don't get the question right have a single incorrect model with most of their remaining probability. If the class mixedness is high, their probability is spread out nearly evenly.

Specifically, after assigning the student their probability p to get the answer right, I'll choose a second model at random. I'll assign it a probability $(1-p)(3m+1)/4$ and the remaining models equal probability to sum to 1.

This method of assigning probabilities to students loses a lot. There's no such thing as a general student ability. Student ability in different expert-equivalent questions is uncorrelated. Further, there's no spread in the extent to which students are in mixed or pure states. These are sacrifices I'm making for the sake of a smaller parameter space to explore.

A full modeling of model analysis would require me to drop the above assumptions and add more parameters. This considerable extra complexity poses a true challenge for analysis methods. For example, two answer choices to the same question that belong to the same model are still mutually exclusive on the test, so it might be difficult for factor analysis or network analysis to pick them as belonging to the same model, even though they do. Using model analysis to generate more-challenging clustering problems in future work could create some caveats to the conclusions I find here. However, my goal with model analysis is to simulate data with fairly clean real signals for the analysis methods to find and prove their utility on at least simple test cases. Bayes nets provide the murkier test case, so for this analysis, I'm okay with making all these simplifying assumptions to model analysis.

The model analysis simulated data provides an important test for analysis methods that inspect distractors. While the Rasch model and factors model had students choose between distractors randomly if they were simulated to answer the item incorrectly, the model analysis model is the first to include meaning to the incorrect answers students choose. With model analysis simulations, there is real structure to the distractors - each belongs to a certain set of competing models. For an analysis method to be fully successful in analyzing data simulated from the model analysis model, it should

1. successfully partition questions into the expert-equivalent sets of questions that were built into the simulation

2. within a set of expert equivalent questions, successfully partition responses according to the models they represent

Most analysis methods will not be completely successful in this way, because they aren't set up to make two levels of distinction in clustering the test. For example, an analysis method that clusters all individual item responses at will might be expected to assign a cluster to each of the different models in model analysis, but then it wouldn't explicitly state which questions are related. A human could read that information off from the results by noting that several clusters would contain responses from the same few questions. It would then be clear that we'd found a set of expert-equivalent questions.

Alternatively, the analysis method might cluster all the responses from all the items in a set of expert-equivalent questions without finding the competing models. Or it might be expected to do a mix of both of these. We would need to write an analysis method with the explicit goal of clustering questions at one level and item responses within clusters of questions if we were to perfectly reproduce the structure assumed in model analysis, and existing methods, such as module analysis, don't do this. I won't set this as a goal here. Instead, if a method clusters all answer choices from the same model together, I will consider it perfect.

5.5.4 Bayes nets

Bayes nets are a probabilistic graphical model used in many fields. I will follow the presentation by Almond et al. [2015] in their book on using Bayes nets for educational assessment.

A Bayes net is a network, so it consists of nodes and edges. The nodes represent variables, which in educational assessment could be items on a test or constructs we wish to measure. The variables in a Bayes net are discrete by definition, and each has an unknown value. The variables have probability distributions, which can be affected by taking data (i.e. administering a test).

The edges in a Bayes net are directed. They represent conditional dependencies between the variables. For example, suppose some Bayes net has variables A, B, C, and D. There is some overall joint probability distribution $P(A, B, C, D)$, as well as marginalized probability distributions (e.g. $P(A)$) and conditional distributions (e.g. $P(A|B)$). Suppose a Bayes net has edges directed from A to B and from B to C and no other edges. Then there are no dependencies between D and the other variables; D is independent of each of A, B, and C. A and C are connected, though, and so are probably not independent. However, they are not directly connected; they are only connected through B. A and C are then independent conditional on B. That is, if we know B, then learning the value of A does not affect the probability distribution for C and vice versa. In general, the probability distribution for a node is a function of all its "parents", those nodes that have edges leading from them to the node in question, and not a function of the values of any other nodes.

Almond et al. [2015] give an example Bayes net related to a test of language skills. It has several nodes for constructs, which are not directly observable. These are the student's true facility for reading, writing, listening, and speaking. Then it has several observable test performance variables. One is performance on a pure listening task, so there is an edge from listening to this node. Another observable variable is for a task that requires the test-taker both to read and to write, so both the reading and writing nodes connect to this variable.

The various observable nodes are not connected to each other because in this model, performance on one task does not directly affect performance on other tasks (for example, things you remember from a reading task earlier in the test don't prime you to answer differently on a later writing task), but the nodes could still have conditional dependencies because they are indirectly connected via the unobservable constructs. (Although all edges connected to the observable nodes are directed from unobservable constructs to the observables, learning the value of an observable still affects the probability distribution we have for the unobservables; the directedness of the edges doesn't imply that learning about variables only affects the distribution of other variables in one direction.) However, the unobservable constructs do have edges between them, indicating, for example, that if students use their facility with speaking to help them read by "sounding words out", then speaking may have a directed connection to reading (though in Almond et al. [2015]'s example, the connection actually runs the other way; the precise reasoning behind the connections isn't explained, and the connections are only drawn as an example).

5.5.4.1 Bayes nets and the resource framework

Each of the models we've discussed for simulating student responses to a test is based on some underlying theoretical understanding of how students think (and how their thinking interacts with the test and environment around them). For example, the Rasch model is based on the idea that students have a single true “ability” level for the material being tested. The factor model is similar, but posits that students have several different abilities for different concepts within the same test. The model analysis model is more complex, and is based on the idea that students switch between various models because they activate groups of resources together, but the activation may be based on irrelevant surface features for a time while students are learning. It acknowledges that the probability distributions it measures are not simply properties of students themselves (as the abilities in the Rasch model are), but emergent between the student, the test, and the context under which they take the test. Bao and Redish give a lengthy theoretical underpinning to their modeling, based on the resource framework for cognition and some findings in neuroscience.

Bayes nets are very flexible and general. This is a weakness in that it makes them hard to falsify, but its a strength in that they can be adapted to the researchers' theoretical framework for problem-solving. They seem particularly well-suited for work in the resource framework, overviewed in chapter 2.

One of the progenitors of the resource framework was Minsky [1991]'s concept of a “Society of Mind”, in which the human mind is suggested to comprise a huge number of individually-uncomplicated “agents”, which each perform simple functions. These might be low-level functions like line detection in a visual field, or high-level functions like playing an epistemic game. Each agent is, individually, lacking in intelligence and capable only of very simple tasks, but high-level agents exert control over and organize lower-level agents in a hierarchy which Minsky illustrates with a graph-like structure.

In the PER literature, this network-based picture lies at the heart of some resource framework literature. For example, in his Oersted lecture, Redish [2013] illustrated a network of resources. The resources were connected by edges, representing how the resources tend to activate or suppress each other. This was inspired by similar connections known to exist between neurons.

In this resource network, nodes can have many different types of characteristics. They might be memorized facts, such as the density of water. They could be reasoning primitives, such as p-prims [DiSessa, 1993].

They could have executive functions as well. For example, a node, when activated (i.e. if the node has two values, 0, and 1, the node would be in the state 1), might have edges that raise the probability of activating many different nodes all related to some sort of real-world knowledge, while reducing the probabilities to activate nodes related to formal knowledge, such as equations. When such a node is activated, it could map onto what we observe as an epistemic frame in which physical storytelling dominates, and problem-solvers wind up playing the "physical mechanism" e-game.

Nodes could also represent resources used in conceptual metaphor, for example there could be nodes related to "change occurring by specific agent" related to the ontological metaphors to be discussed in chapter 3.

In chapter 4, I outlined a number of other possible influences on test scores. For example, I discussed several lines of evidence that student effort was a significant factor, and Bayes nets could have resources for effort, attention, etc that would allow us to model the effects seen in that analysis. Questions near the beginning of a test might be test influenced by attention more than questions near the end of the test, for example. Bayes nets could reflect this by having no link from the attention variable to questions at the beginning of the test (or links that only make small changes to the probability distribution), while attention might be linked to every item near the end of the test, with the value of attention having a strong effect of the distribution of responses to each item.

The diversity of types of resources relevant to solving problems is in good accord with the diversity of nodes in Almond et al. [2015]'s view of Bayes nets, as they write, "Variables are introduced to stand for aspects of an examinee's proficiency - elements of knowledge, strategies and procedures, tendencies to solve problems that have certain properties, and so on."

Still, none of these analogies to Bayes nets are perfect. Minsky and Redish both constructed their nets with a very dynamic picture, with different agents or resources being called upon at different times during a problem-solving process, whereas Bayes nets simply represent a probability distribution. But the resource framework calls for a complicated interdependency between diverse influences such as student knowledge, affect, epistemology, self-efficacy, ontology, and more. Networked structures have much greater flexibility and thus ability to model this sort of complicated interplay than drastically-simplified models like the Rasch model. They have the potential to add more meaningful degrees of freedom than model analysis. Model analysis suppresses all the above-listed influences into a broad category of “context dependence” and replaces an attempt to model them with a simple probability distribution for choosing between alternative models.

5.5.4.2 Discrete values

By definition, the variables in Bayes nets are discrete. This condition could certainly be relaxed. In the Rasch model, ability is a real number from $-\infty$ to ∞. Constructs in probabilistic graphical models could likewise be real numbers. However, if we are modeling test-takers’ responses to test items on a multiple-choice test, those variables should be discrete. Choosing discrete variables for unobservable constructs is a restriction, but is useful because it makes the models computationally much easier to work with. Here, we’ll continue to use discrete variables simply because we wish to stay closely connected to the considerable existing work on Bayes nets in educational assessment.

5.5.4.3 Evidence-centered design and learning Bayes nets’ structure

Model analysis can only be performed on a test built in a certain way, with previously-known expert-equivalent items and previously-identified models associated with the item responses. Similarly, statistical inference on test results can allow us to learn the values of unobservable nodes in a Bayes net, and even to learn the structure of the Bayes net, but only on tests constructed in a certain way. Almond et al. [2015] outlines a holistic method for conceptualizing, writing, validating, and analyzing tests with Bayes nets called Evidence-Centered Design (ECD).

Our design of the MEGS outlined in chapter 4 was systematic, and shared many elements with ECD, but was not fundamentally conceived or executed as an ECD project. Specifically, ECD places a different kind of emphasis on test validity in early stages of the design process than we did. We constructed questions we believed would be best-solved by employing one of four e-games, and considered validating these questions to be a task for qualitative problem-solving interviews with students, artifacts from students solving the problems while showing written work, and expert opinion. In other words, we viewed validation as a step to take on the test after it was written. In ECD, validity takes a more fundamental role in the creation of the items. These and other discrepancies between ECD and our process lead me to shy away from applying Almond et al. [2015]'s techniques for analysis of test results via Bayes nets. Instead, I will used Bayes nets simply as a way to construct simulations of test data to be analyzed with other techniques.

5.5.4.4 Constructing a simulation with Bayes nets

To use Bayes nets to model a class of students taking a test, I do this:

1. for each item on the test, define an observable variable in the Bayes net, having as many potential values as there are choices for that item

2. define a number of unobservable "resources" as variables in the Bayes net. These may correspond to constructs, or to finer-grained resources of various types, as discussed in 5.5.4.1. Give each resource two potential states, "activated" or "dormant".

3. draw directed edges between the variables. There can be at most one edge between two variables. Edges can go between resources or from resources to test items.

4. for each edge between a resource and an item, decide whether the edge is "productive" or "counterproductive"

5. for each edge between two resource, decide whether the edge is "inhibitory" or "promoting"

6. for each variable (both resources and test items), define a joint probability distribution giving that variable's probability distribution as a function of the values of all its parents. for items in particular, load more probability onto the correct response when productive resources are activated and onto the incorrect responses when counterproductive resources are activated

7. identify all variables with no parents

8. define a prior distribution for each variable with no parents

9. for each student in the class, sample from the prior for each variable with no parents to assign the student values for those variables

10. for each student, using the joint-probability distributions defined earlier, find the probabilities for all the remaining variables in the network

11. for each student and each test item, sample from the item's probability distribution once and record the result as the student's response to the test item

The above procedure is fairly general; it doesn't specify exactly how all the probability distributions should be defined, for example. This is to give some space to explore different possibilities. However, I've narrowed resources down to having only two potential states. This greatly simplifies the model, and makes a superficial connection to neurons, which can fire or not fire, and computer bits, which are either 1 or 0. I don't claim any close connection between resources and either neurons or computer bits, but because neurons and computer bits can both represent a great diversity of things, they show that restricting certain variables to only two values, rather than many, does not necessarily mean the network can't have diverse behaviors.

I will introduce three parameters to explore Bayes nets, in addition to the number of students (making four total parameters). First is the number of resources. Second is p, the probability for an edge that can exist to exist. I will set this the same for edges between resources and from resources to answer choices. For edges between resources, if they are drawn, I choose the direction at random. Additionally, if a proposed edge would cause the graph to become cyclic, I don't add it.

Finally, there is connection strength, s. For links between resources, the first resource being activated increases / decreases the log-odds for the next resource to be activated by s (depending on if the link is promotional or inhibitory, respectively). For links from resources to answer choices, productive links increase log odds of correct answer choices by s and decrease the log-odds of incorrect answers by s. All probabilities for resources are 50% without considering edges, and all probability for answer choices are 20% without considering edges. After propagating probabilities through the network, probabilities for answer choices to a single question will no longer add to 1, so I'll normalize them.

Four total parameters is more than I want to search through, so for this chapter I will set $s = 1$ and leave exploring the effect of changing s to future investigations.

I also need to describe how I'll set the values of resources with no parents. For these, I'll choose a probability, uniform from 0 to 1, for each resource. Then for each student, the resource will be activated with probability equal to the probability associated with that resource. This completes the specification for how I'm coding Bayes nets.

5.5.4.5 The role of Bayes nets in evaluating analysis techniques

In this chapter, I am examining various analysis techniques abilities to learn about student cognition from examining the choices they make on multiple choice tests, especially learning what distractors are related to each other via clustering. Data simulated via Bayes nets provides a true challenge here, because although there is underlying structure, that structure is much less explicit than in other models.

For example, in model analysis, each item response belongs to a particular model. There is a single correct answer to the question, "under the model by which this data was simulated, which item responses belong together?". For Bayes nets, this is less true. A single resource may influence the probabilities for, say, six different test items. But in each case, it will amplify the probability for some test items and suppress the probability for others. A resource might be productive for one question and unproductive for another. Should the analysis method then group the two questions together or not?

Additionally, Bayes nets lead to fuzzy borders. One resource might influence five questions, but of those five questions, perhaps two are both influenced by some other resource, two more have individual resources that influence them and well as other resources not connected to the original five, and one is influenced only by the resource in question. Then how should a grouping of items be made? It's hard to say because the sets of item influenced by the resources may overlap considerably.

Bayes nets add the most potential for sophisticated, complicated modeling to my stable of models for simulating test responses, but they also represent the toughest case to evaluate because the underlying structure to the simulated data is messy. This is valuable in moving from analysis of simulations to analysis of real data because my theoretical framework of the resource framework predicts complicated real-world connections that could likely lead to murky-looking results. I would *a priori* expect the real results to resemble results from Bayes net simulations more than the results from the other types of simulations discussed in this chapter.

5.6 Methods for analysis of test data

Here, I'll describe three methods of analyzing test data, each of which takes advantage of which specific distractors students select in order to try to gain some new insights from the test.

I applied each of these three methods to simulated data to understand better how the methods behave, then I applied each method to real data from the MEGS to analyze that survey.

5.6.1 Factor analysis

In this section I'll first review factor analysis as it is usually used in educational assessment to set up a contrast with the new technique of Scott and Schumayer [2017] I'll introduce next for doing factor analysis with the test distractors.

5.6.1.1 Traditional factor analysis (no distractors)

Factor analysis is a common technique in research on education assessments. Its goal is to understand a many-item test as essentially a linear superposition of a few independent factors. For example, suppose you gave a test on identifying species of birds, then gave another test on naming the capitals of various countries, then mixed up the questions to create one test with the questions interspersed. If you gave the test to a bunch of people, graded the questions as right or wrong, and handed those results over to someone (without giving them information on the questions, just which people got which right), they would have a good chance of being able to tell which questions were about birds and which were about capitals (or at least, which questions were of the same type. They might not know whether that type was birds or capitals). The reason is that some people know a lot about capitals. If they get one capital right, they're more likely to get the others right. But someone who knows a lot of country capitals doesn't necessarily know how to identify birds and vice versa. There might be some correlation between bird and capital questions, but the correlation within a category would presumably be much stronger than the correlation between categories. We could probably represent someone's test result pretty well using a score for that person on how well they know capitals, another score for how well they know birds, and a bunch of other little scores for random mixes of questions that encode all the remaining noise in the data. Factor analysis is a way to do this using principal components analysis.

Factor analysis is relevant to PER. The FCI, for example was designed to test several independent concepts related to force, such as Newton's third law, velocity and acceleration, and gravitational force. These concepts might play a role similar to the roles of birds and capitals in the previous paragraph, but whether they actually do is an empirical question. In 1995, Huffman and Heller [1995] challenged this idea by performing a factor analysis on FCI scores.

The idea behind a factor analysis is that if the FCI tests, say, six distinct concepts, each of which students answer independently, then student scores on this 29-question test should be effectively six-dimensional. We can treat student responses as vectors in a vector space, and a projection into six dimensions should lose very little information. Further, each of the six basis vectors, called "factors" (or "latent traits"; we will use "factors" here), should consist of several related questions that exist only in that factor and not in others.

To perform a factor analysis on the results of a test administration, we first compute a matrix whose i, jth entry is the correlation coefficient between questions i and j, with correct responses receiving a score of 1 and incorrect responses a score of 0. This gives a symmetric real matrix, and we compute its eigensystem. These eigenvectors will contain some non-zero amount of all of the questions on the test if for no other reason then because there is noise in the responses, but the eigenvectors should ideally be mostly focused on a few questions. These eigenvectors necessarily have zero correlation with each other because they are an orthogonal basis for the correlation matrix.

It is usual to perform a sort of "rotation" to the eigenvectors to transform them so they have zero coefficient for most questions and a coefficient of one for the questions that truly belong to that eigenvector, or as close to this ideal as possible.

We interpret the eigenvectors as "factors" and the coefficients of each individual question in a given factor are called "loadings". Factors with high eigenvalues are considered significant, while factors with low eigenvalues are not. For a given factor, the questions with high loadings should all have fairly similar correlations, and thus we interpret them as testing the same concept. Ideally, there should be six large eigenvalues and 23 small ones (for a 29-question test with 6 independent concepts). To see whether this is the case, we apply some objective criterion to decide which eigenvalues are large enough to be considered significant. (There are various possible objective criteria, but an example is to choose all factors with loadings greater than one.)

Applying this procedure to FCI data, Huffman and Heller concluded

> All in all, the large number of insignificant a factors produced by this factor analysis and the limited number of items that grouped together

> on the few significant factors indicates that the question on the FCI are only loosely related to each other and do not necessarily measure a single force concept or the six conceptual dimensions of the force concept as originally proposed by its authors. [Huffman and Heller, 1995]

In their response, Heller and Huffman [1995] claimed that the FCI did not intend to measure six independent coherent theories held by students; only that the six categories they created were logically-independent standards against which to hold students as they learned about force.

Both the FCI's authors and its critics agreed that students do not have a coherent set of concepts, so that two questions both testing Newton's first law will not necessarily be correlated because students may respond mostly to surface features of the questions, such as whether they ask about familiar or unfamiliar scenarios (e.g. hockey pucks vs rockets in space), rather than by the underlying physical concepts. Thus, they did not expect a factor analysis to be able to discover strong factors (factors with high eigenvalue).

However, Scott et al. [2012] used a different data set and slight adjustments to the data analysis and found a single strong factor in FCI data, and additionally five meaningful conceptual factors after performing a non-orthogonal transformation on the correlation matrix.

Ultimately, it appears that physics education researchers do not yet agree on the extent to which students answer test questions based on coherent physical theories (correct or not) or on the interpretation of factor analysis performed on test results. However, factor analysis continues to be a tool used to group questions together.

Here, we won't attempt factor analysis on the MEGS. We already know that factor analysis of FCI produces somewhat-murky results at best. The questions of the MEGS are related more subtly than those of the FCI. Whereas related FCI questions test the same physical concept, related MEGS questions are related only in that a particular problem-solving strategy, and one which students may not even know at the beginning of the course, is useful for solving them, from the view of the test creators. While we wrote the MEGS hoping that learning these problem solving strategies would lead to MEGS improvement and that MEGS improvement demonstrates a genuine improvement to student facility with recognizing and applying the strategies, we didn't expect the questions to be strongly-enough linked that the factors we put into the test would pop out of a factor analysis.

That's not to say we believe a factor analysis of the MEGS would not be useful; it might still give valuable information about which questions are connected from the students' point of view, but in this chapter I'm more interested in extracting information from distractors, so I'll introduce a modified factor analysis which distinguishes different incorrect responses from each other, unlike traditional factor analysis.

5.6.1.2 Factor analysis with distractors

In order to learn from distractors, we can modify the factor analysis procedure slightly. Here I will follow the work of Scott and Schumayer [2017], which describes several important technical steps to overcome problems that occur in trying to extend factor analysis to include distractors.

One generally begins factor analysis with a matrix of 1's and 0's. Each row represents a test taker and each column a test item. 1 means the test taker got that test item right. 0 means they got it wrong.

We can instead begin with a matrix where each column represents a response to a test item. 1 means the student selected that response, 0 means they did not. We can then apply factor analysis exactly as before on this matrix.

One problem with this procedure is that responses to the same question will necessarily be strongly anti-correlated because they are mutually exclusive. The correlation coefficient will in fact be -1 exactly for any two choices to the same question, so long as at least one student picked each choice (otherwise the correlation coefficient is undefined, and we can set it to -1 for convenience here). Factor analysis may simply pick out the question choices for the same question as factors. To sidestep this problem, I'll set the corresponding elements of the correlation matrix to zero rather than their true correlation, following Scott and Schumayer.

5.6.1.3 Deciding how many factors to keep

For a 30-question test where each question has 5 answer choices, traditional factor analysis results in 30 factors, whereas this procedure results in 150 factors. I need a small number of interpretable factors to come out of factor analysis, so I'll need to adopt some criterion for decided which factors to keep.

It's universally agreed that one should keep the factors with large eigenvalues and dismiss those with small eigenvalues. Beyond that, there are many competing criteria for where to make the cutoff between significant and non-significant factors. Some of these criteria are very simple. For example, the Kaiser criterion suggests keeping those factors with eigenvalues greater than one. However, this method is known to result in keeping too many factors for large datasets, and trying it out in some preliminary tests on simulated data, I saw it could keep up to 50 factors when the simulated data should only have had 3.

Other criteria are rather ad-hoc, involving a human looking at a plot of the eigenvalues and subjectively deciding where they see a significant drop-off.

Then there are a large number of analytical techniques, and reviews of the literature on this topic don't lead to a single, universal recommendation [Courtney and Gordon, 2013]. In examining MEGS data, I'll be most interested in the few factors with the largest eigenvalues, so keeping one or two extra or throwing one or two extra out over an ideal solution isn't too big a deal.

For this study, I tested the methods provided in the R package 'nFactors' [Raiche, 2010]. This package outputs several different estimates for the number of

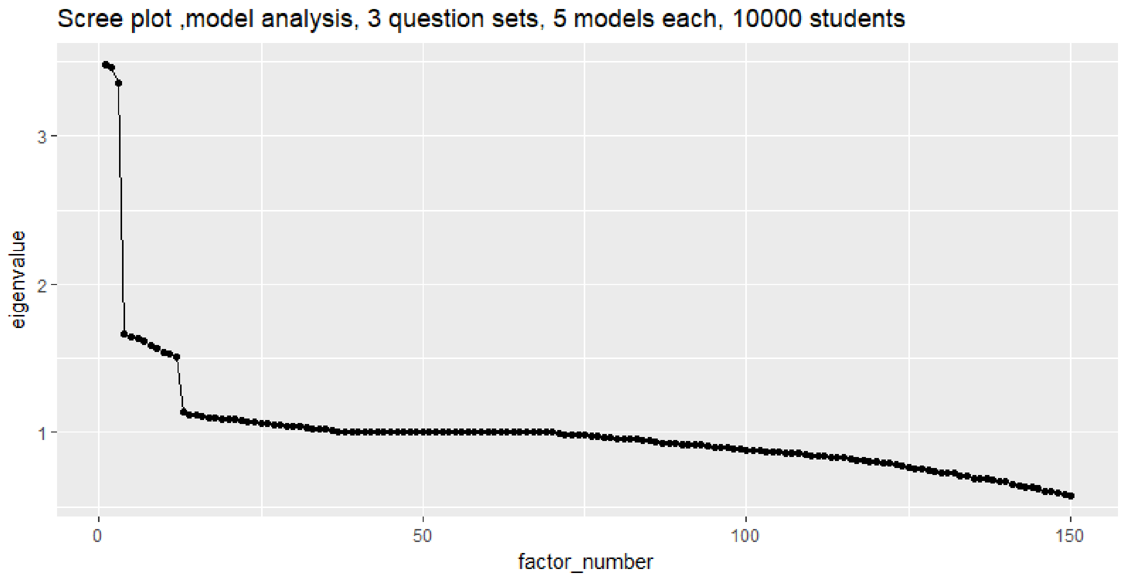


Figure 5.1: Scree plot for the factor analysis eigenvectors generated from simulated model analysis, with mixedness of 0.3. The three large eigenvalues correspond to three correct models. The 12 medium eigenvalues correspond to 12 incorrect models.

factors to keep. I tested it on a simulated data set from model analysis with three expert-equivalent question sets and five models for each question set (including the null model). The scree plot (plot of size of eigenvectors versus eigenvector number, ranked from highest to lower), is shown in figure 5.1.

As expected, the 15 models in the simulated data led to 15 significant eigenvalues, when evaluated by human inspection. These came in one group of three for the three correct models, and one group of 12 for the 12 incorrect models. A good method of determining the number of eigenvalues to keep should generate 15, but could potentially generate 3 as well.

When I ran this test multiple times, the results were visually consistent -

there were always three large eigenvalues of $\approx 3 - 3.5$, twelve medium eigenvalues of $\approx 1.5 - 1.75$, and 135 small eigenvalues from $\approx 0.2 - 1.2$.

The nFactors package gives four estimates for the number of significant factors. The Kaiser method (eigenvalue ¿ 1) always vastly overestimated, as did the "parallel analysis" method, which compares the eigenvectors to those that would be expected from completely random data.

The "acceleration factor" method always found 3 significant eigenvalues, corresponding to the three correct models. It never recognized the medium eigenvalues as significant, though. This method attempts to search for kinks in the scree plot.

The "optimal coordinates" method uses a slightly different methodology to look for kinks, and had varied results. I've plotted the number of eigenvalues it generated in 100 runs in figure 5.2.

The optimal coordinates method sometimes finds the correct number of eigenvalues in the model analysis model, but in general it's quite poor. It was successful at finding the 15 factors in only 6 of 100 runs, even though a human would find the factors by inspection easily. Optimal coordinates sometimes finds zero significant eigenvalues. That's with the overly-clean data that comes from simulations of 10,000 students. I repeated the test with 100 students, and plotted the number of significant factors found by the acceleration factor and optimal coordinates methods in figures 5.3 and 5.4.

With a smaller class size to work with, the acceleration factor is no longer consistent, and sometimes finds the three correct models as factors, but sometimes finds more or fewer factors. The optimal coordinates method never found exactly

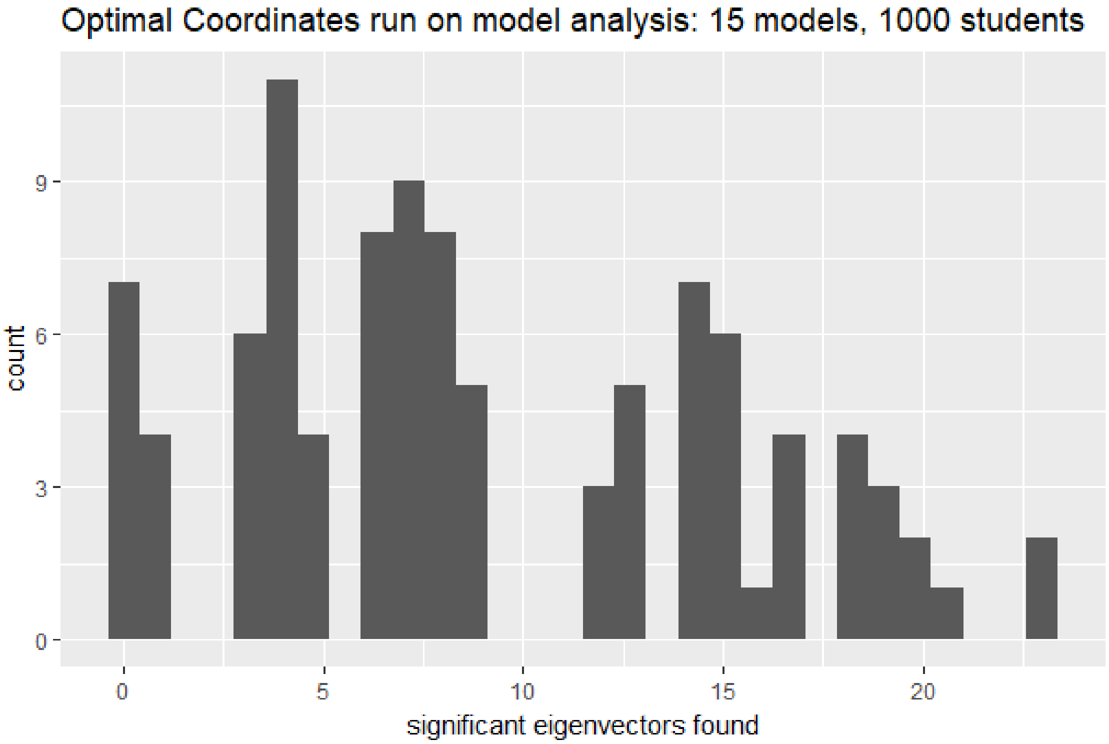


Figure 5.2: Number of significant factors detected by the optimal coordinates method in 100 runs of factor analysis on a simulated test result. Test was simulated under model analysis with three expert-equivalent question sets, each with five models (including the null model).

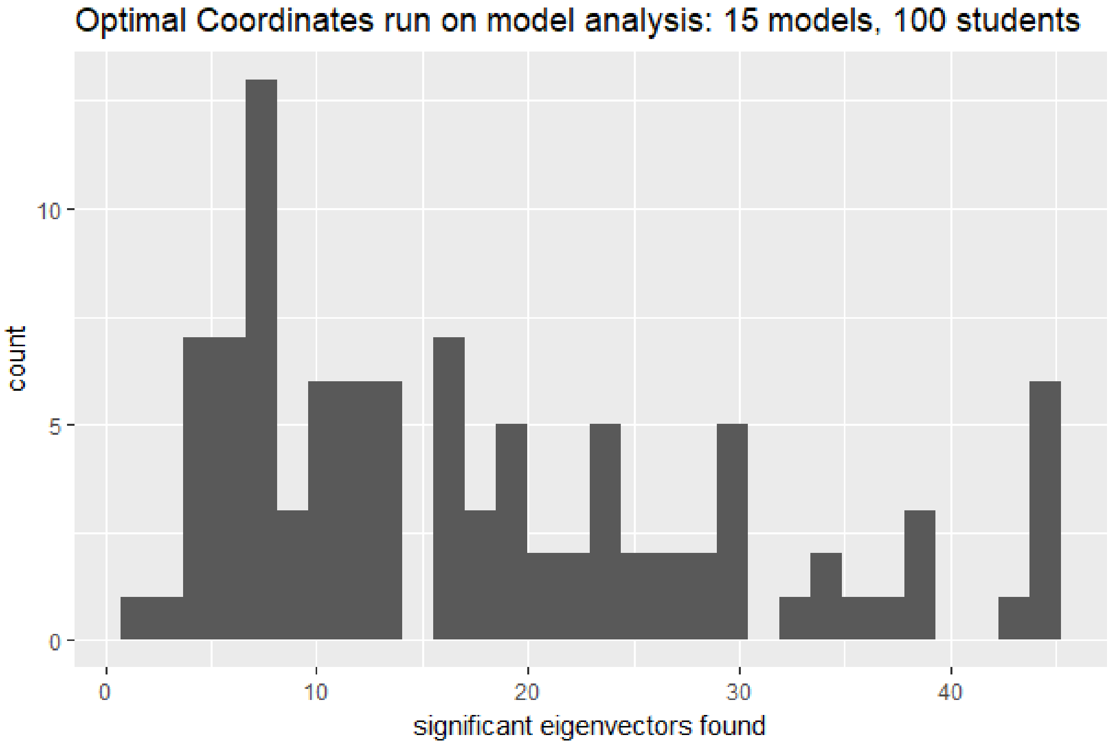


Figure 5.3: Number of significant factors detected by the optimal coordinates method in 100 runs of factor analysis on a simulated test result. Test was simulated under model analysis with three expert-equivalent question sets, each with five models (including the null model).

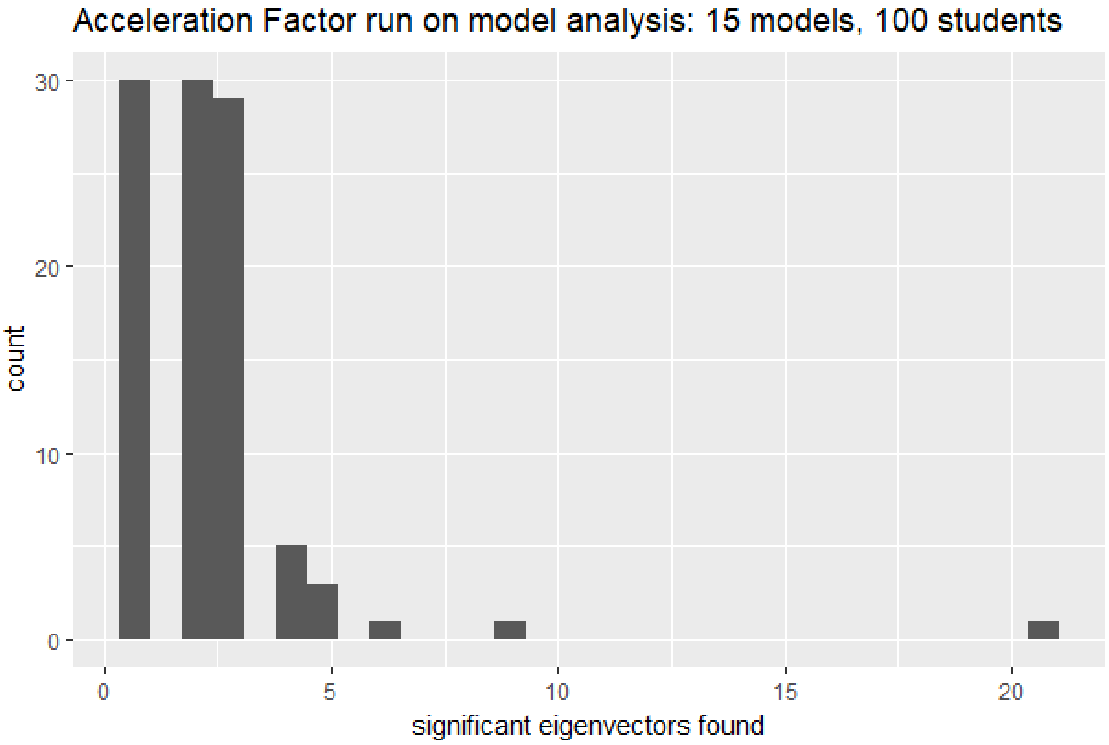


Figure 5.4: Number of significant factors detected by the acceleration factor method in 100 runs of factor analysis on a simulated test result. Test was simulated under model analysis with three expert-equivalent question sets, each with five models (including the null model).

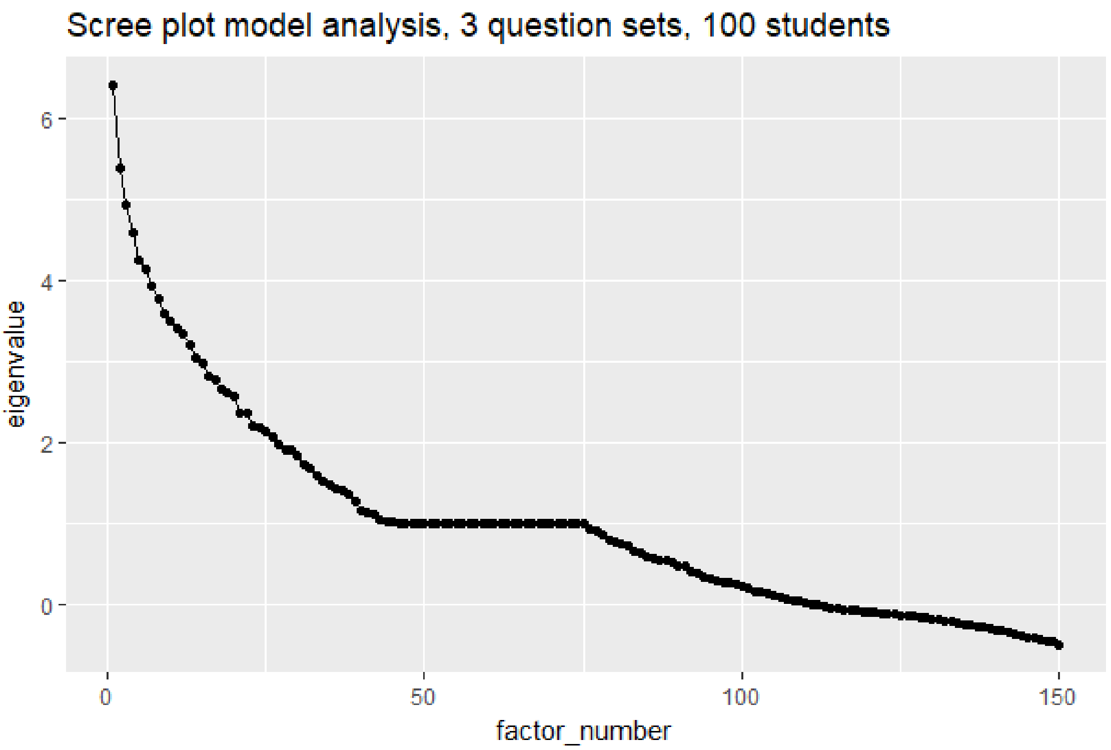


Figure 5.5: Eigenvalues for factor analysis on simulated model analysis data with 3 sets of expert equivalent questions, 5 models per set, and 100 students

15 factors.

A typical scree plot for this tests with 100 students is in figure 5.5

The structure is now much less obvious, and the analytical methods are hard to fault for only having moderate success in finding the correct number of factors.

Model analysis is somewhat subtle due to the mixed states students are in (the above tests were with mixedness 0.3) The factors model is more straightforward. With 100 students and a 4-factor simulation, a typical scree plot looks like figure 5.6

By inspection, there are four eigenvectors separated from the remaining 146. This is expected because there are four factors in the simulation.

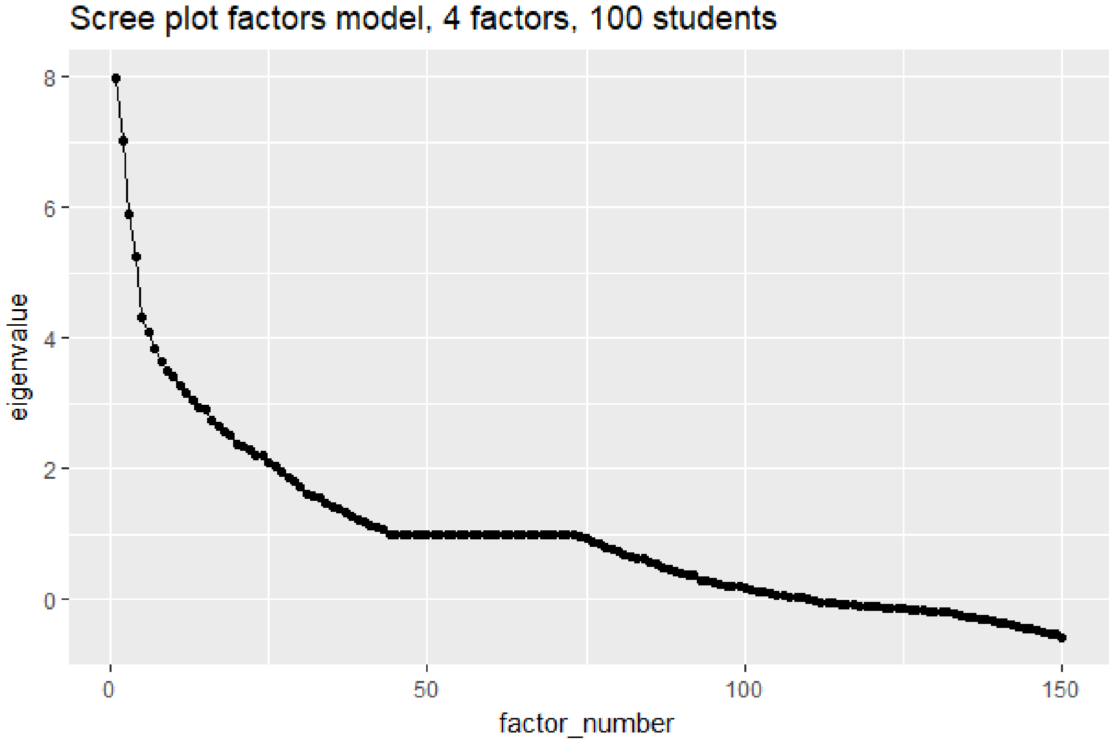


Figure 5.6: Eigenvectors for factor analysis run on the factors model with 100 students and 4 factors. Student mean ability is zero, normally distributed with a standard deviation of 2.

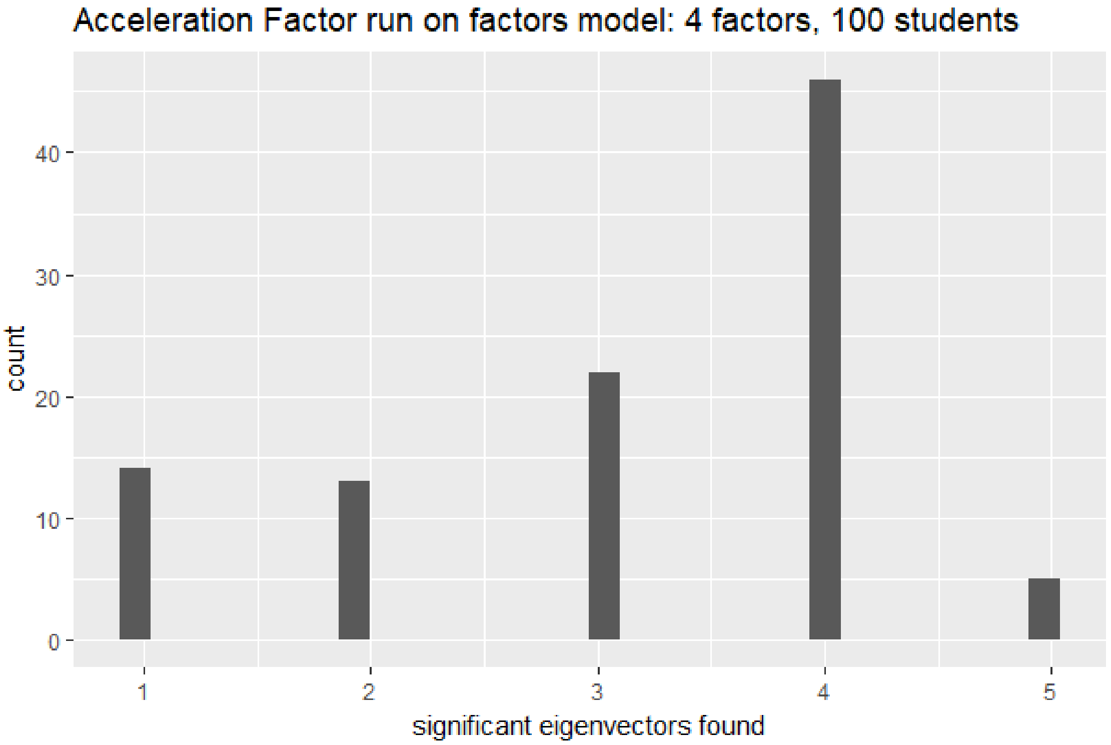


Figure 5.7: Number of significant factor found by the acceleration factor algorithm for factor analysis run on simulated data from the factors model with 4 factors, 100 students, mean ability zero, ability standard deviation 2

The acceleration factor is quite good at detecting the four factors. In 100 trials with the same conditions just described, it found four significant factors in 46 trials. A histogram of its results is in figure 5.7.

The optimal coordinates method had more varied results, as shown in figure 5.8.

When I increased the number of students to 1000, the four factors from the factors model were very clear to human inspection, see figure 5.9. In this case, the acceleration factor correctly discovered four significant factors every time.

The results for the optimal coordinates for 100 trials of this condition are

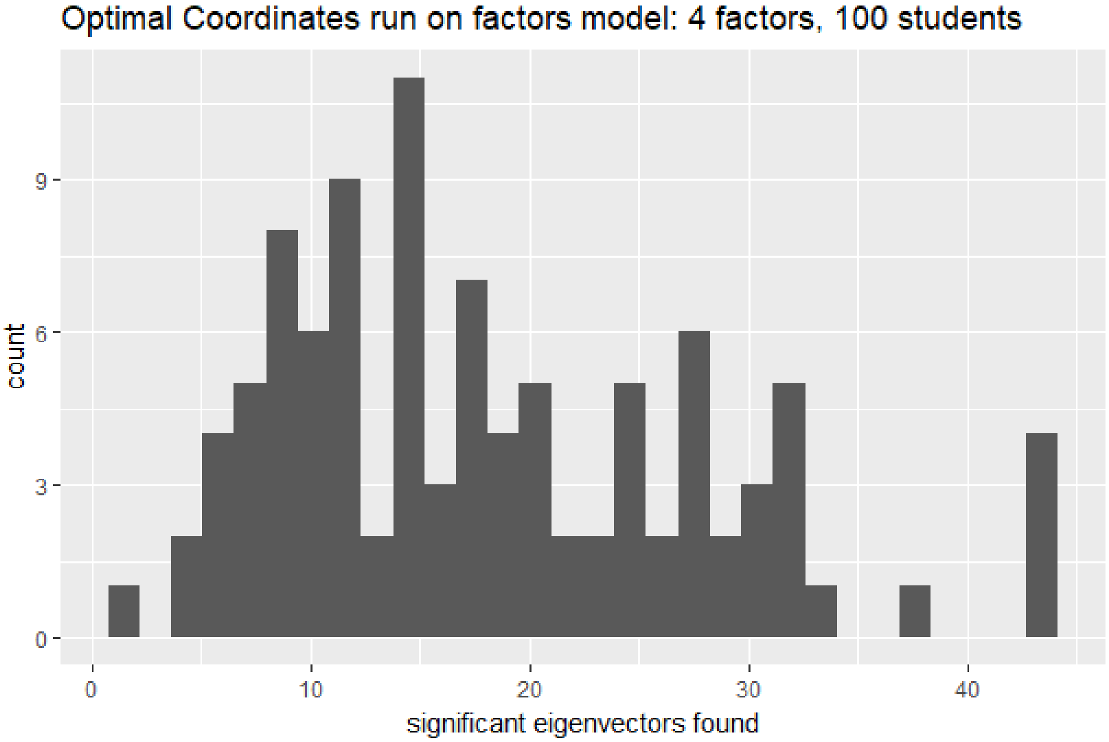


Figure 5.8: Number of significant factor found by the optimal coordinates algorithm for factor analysis run on simulated data from the factors model with 4 factors, 100 students, mean ability zero, ability standard deviation 2

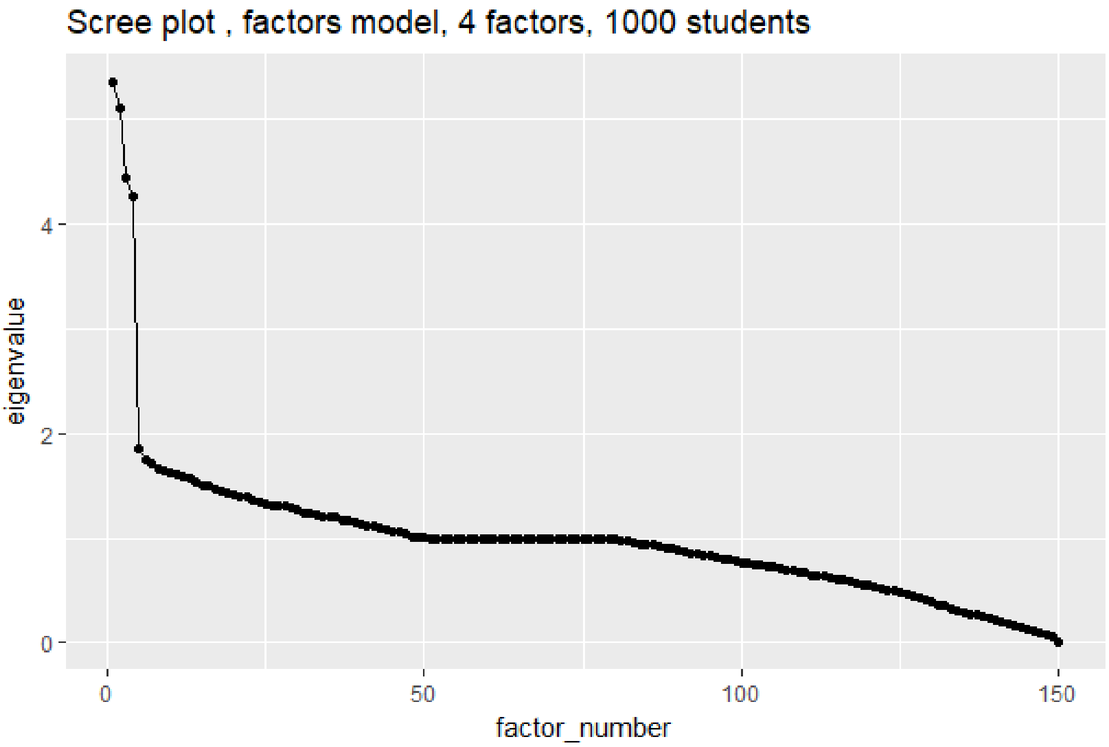


Figure 5.9: Eigenvalues from performing factor analysis on simulated data from the factors model with 1000 students, mean ability zero, standard deviation 2.

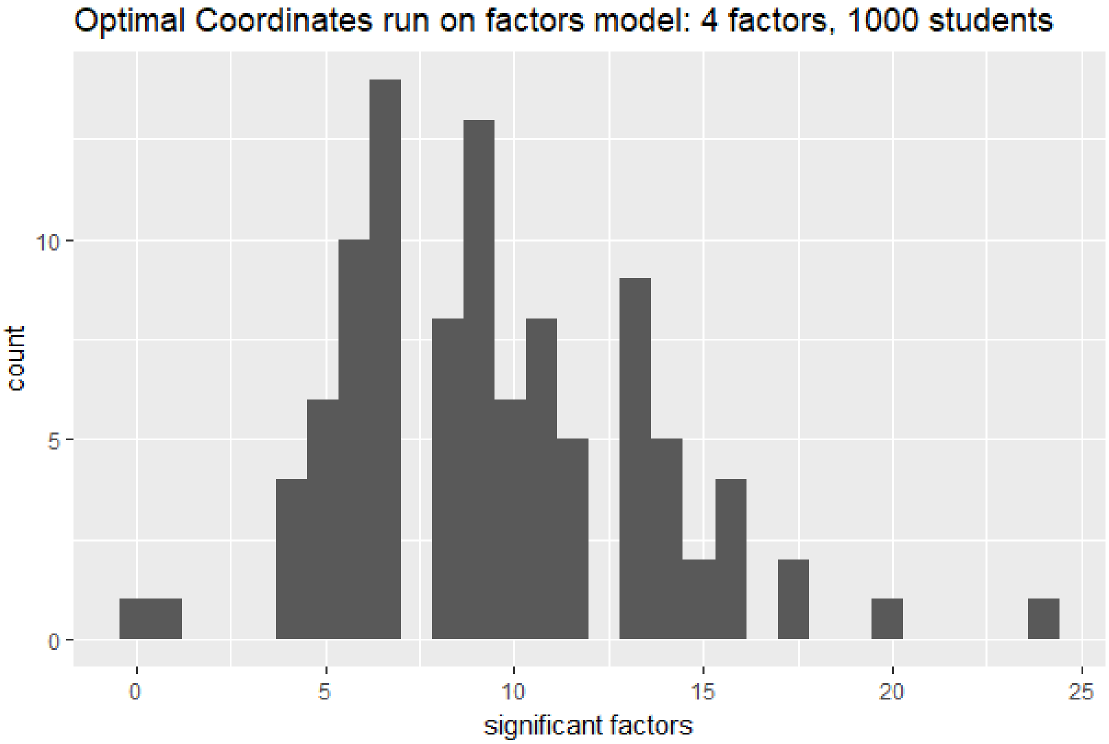


Figure 5.10: Number of significant factor found by the optimal coordinates algorithm for factor analysis run on simulated data from the factors model with 4 factors, 1000 students, mean ability zero, ability standard deviation 2

shown in figure 5.10.

My goal is to detect signals in the distractors, so the optimal coordinates method has something to recommend it. This is the only method I tested that can find the existing structure in distractors in model analysis, but it was very inconsistent in doing so. The optimal coordinates method occasionally fails completely and returns nothing, and also often wildly overestimates the number of significant factors.

The acceleration factor method is likely to find the most significant eigenvectors. Although it never found the medium-sized eigenvectors in my test of model analysis, its performance was superior over all. For this reason, I'll use the acceleration factor method to determine how many factors to keep in factor analysis.

5.6.1.4 Rotation

Even keeping only the most significant eigenvalues, the eigenvalues of the correlation matrix in factor analysis come out messy. They don't simply pick out, say, questions 1, 3, 5, 6, and 10 as a factor. Instead, they pick out some amount, possibly negative, of every question on the test (or question responses in our case) and combine them all together. We then have to judge how to turn this into an interpretable factor that wholly includes or wholly excludes each question choice.

The problem is especially bad when the true structure of the data has several factors that all contribute about equally to the makeup of the test. If the test has 40 questions and four factors, each of which comprise ten questions, there is a sort of degeneracy in the test structure, as several factors should ideally have the same eigenvalue. The eigenvectors will be especially mixed up, from a human points of view. Factor analysis is just a linear algebra technique. Its goal is to choose factors that maximize the variance explained, which has little to do with human readability or sensibility. Thus, factor analysis is happy to take linear superpositions of vectors that we'd like to keep separate for human interpretation.

For example, if knowing a student's coefficients for factors 1 and 2 contribute about equally to understanding how the student will respond to the test, factor analysis has no real reason to pick out factors 1 and 2 as opposed to, say, factor

$1+2$ and factor $1-2$, or some other linear combination. In the degenerate or nearly-degenerate situation, small amounts of noise from the randomness of the simulation or from the context-dependence of real life will make factor analysis choose between such superpositions essentially at random.

The problem is minimized, but still present, when the true factors contribute significantly different amounts to our understanding of a student's responses.

To deal with these messy eigenvectors, factor analysis includes a last step of "rotation", in which we change the eigenvectors into cleaner, interpretable vectors. There are many procedures extant for exactly how to do this. The biggest philosophical divide between these procedures is whether to enforce that the resulting factors are orthogonal. (If we don't force the results to be orthogonal, we do sacrifice some of the aptness of the word "rotation", but I'm willing to make that sacrifice.) In the MEGS, it's entirely plausible that some answer choice could be a part of multiple factors, and so I'll use the method called direct oblimin rotation, as implemented in the R package "GPArotation" [Bernaards and Jennrich, 2012].

Different rotation methods give similar results, and *direct oblimin* is simply a popular one already implemented in statistical software packages. It minimizes the distance between the eigenvectors we've found and the results of the rotation according to some technical measure of distance.

Direct oblimin rotation will still give a variety of loadings onto factors; it doesn't turn factors into clean lists of 1's and 0's. That's a necessary step in order to have factor analysis identify which questions belong to which factor, though, which is my plan for analyzing the various simulations. To that end, I transformed the results of oblimin rotation so that the question with the highest loading and any other questions within 80% of that loading are changed to loading 1, and other questions are changed to loading 0. This will produce subsets of the test for each factor. They may or may not form a partition, but I can calculate the variation of information regardless. It no longer has all the nice properties, such as being a mathematical metric, but still gives a idea of how close two ways of subsetting the questions are.

5.6.2 Module Analysis for Multiple Choice Responses

Recently, several researchers have begun analyzing physics education research data using network analysis. Many of these efforts focus on the interactions between students, for example while working in collaborative tutorial sessions. However, a recent paper by Brewe et al. [2016] models FCI responses as a network and uses community detection algorithms to "identify clusters of common responses which map to models held by students". They call this technique Module Analysis for Multiple Choice Responses (MAMCR). In MAMCR, to model responses as a network, each question response and each student are treated as a node. A question response is then linked to a student if the student selected that question response. The resulting network has unweighted, undirected edges. The network is bipartite, meaning that there are two different sets of nodes (students and question responses). Edges run only between the two sets, not within them.

This model includes distractors. For a test with 30 questions and 5 answer choices per question, there will be 150 distinct question response nodes. The inclusion of distractors was the primary motivation behind Brewe et al. [2016] analysis. They hoped to use distractors to shift concept inventories from generating a number that evaluates instruction to giving insight into how questions and their distractors are linked from the viewpoint of students. This is in contrast to traditional factor analysis, where answers are considered either correct or incorrect, and the analysis does not consider which incorrect answer a student chose. It runs parallel to Scott and Schumayer's version of factor analysis that uses responses, not questions. This means we can hope to compare the results of the MAMCR and factor analysis with distractors.

Network-based clustering techniques are not yet well-tested in physics education research on educational assessment outside of Brewe et al. [2016]'s first effort and a few preliminary follow-ups. These methods have the potential to grant new insights by distinguishing between various incorrect answers, but we do not yet know how well they will agree with factor analysis or how well they pick out clusters when they truly exist. That is why this project proposes to compare the effectiveness and interpretation of factor analysis and network-based methods.

5.6.2.1 Projection

There are a many different techniques for detecting clusters in networks. Most well-developed techniques are not based on bipartite networks like the one generated as the first step in MAMCR. To generate a network they can use with standard module detection techniques, Brewe et. al. begin by projecting their network onto question responses. Projection is a common technique in bipartite network analysis. It turns the bipartite network of students and question responses into a network of only question responses. There are different definitions of projection, but in this case, it creates a weighted network. The weight of an edge between two question response nodes is equal to the number of students who chose both question responses while taking the test. This sort of projection can be achieved with a simple matrix multiplication (where the matrix is just a matrix of 1's and 0's representing which nodes in the original network have edges between them).

I have followed Brewe et. al.'s lead, implementing their projection in R.

5.6.2.2 Simplifying the network

Next, Brewe et. al. removed nodes from the question response network. First, they removed any nodes with no edges (these were nodes that no student chose). Then, they removed all the nodes corresponding to correct answers. Their reason was that these nodes were by far the most-commonly chosen (the FCI class average score was 85% in their data set) and they all clustered together very strongly and obscured the rest of the data. Brewe et. al. stress that this step might not be necessary for other data sets. For analyzing simulated data and finding the variation of information, the correct answers are important, so I'll leave them in. I will run the MEGS analysis both with and without the correct questions responses in the network.

After removing those nodes, the resulting network still has a lot of noise - links where just one student happened to choose both answers, possibly at random. It might make sense to delete all such nodes, but that might be too hasty. If a node has very few total links, a link with weight 1 might still be meaningful to it, whereas that link would be less meaningful to a node with many links with weights in the tens or hundreds. Acknowledging the varied important of low-weight links, Brewe et. al. pared the network down using locally adaptive network sparsification. I will do the same, but using the disparity filter of Serrano et al. [2009], which accomplishes a similar function. I used the implementation in the semnet R package by Welbers.

5.6.2.3 Detecting modules

At last, the network was ready for module detection. Brewe et. al.'s chosen algorithm was InfoMap, citing it having "proven both stable and useful in physics education research". InfoMap has a charming interpretation in terms of random walks on the network and the coding them with minimal information, but there doesn't seem to be a very direct analogy to student cognition, so I will leave the details of this algorithm to the paper cited by Brewe et. al. [Rosvall et al., 2009]. The takeaway is that it finds clusters that have many internal connections, while also making sure there are few connections between the clusters.

The results of running InfoMap are modules of question responses. Humans can then look at which question responses they are, and speculate on why they're connected. In Brewe et. al.'s FCI analysis, the two most strongly-connected nodes, for example, were 30E and 13C. 13C says that if you throw a ball up the in air, while it's rising, it has a constant gravitational force downward and a force from your hand that gradually diminishes as the ball reaches the top of its arc. 30E says that after hitting a ball with a tennis racket, while it's flying over the net, it feels a downward force from gravity, a force from the air it's passing through, and the force from the hit from the racket.

It's not surprising these questions choices were so strongly linked. They deal with the same apparent misconception of objects "carrying force with them", and they are both about balls moving through the air under the influence of gravity. Taken alone, we can't conclude that these nodes are linked because of a the physical theories we see in them. They might be linked due to surface features. However, the module Brewe et. al. found for 30E and 13C contained other question responses, such as 6A, that seem more aligned with a "physics theory" interpretation than any surface features. (6A says that when a ball leaves a frictionless curved tube, it continues curving for a while). Together, this module became the "impetus module" because it contained responses consistent with a physics theory in which things remember what they were doing a little while ago and carry some influence along with them.

By showing that these conceptually-coherent distractors were linked in student responses, Brewe et. al. found good evidence that students in their cohort were, in the language of model analysis, in a pure state using an incorrect model. Brewe et. al. found five other modules, each arguably human-interpretable as belonging to a distinct physical theory. This represented a significant advance for MAMCR over factor analysis to learn about student thinking from test analysis.

5.6.2.4 Implementing MAMCR

I had several difficulties implementing MAMCR which may affect the results. None of the network backbone extraction implementations I tried proved reliable on the networks I was using. The implementation by Welbers, which I used in this study, would often remove all edges until I decreased the algorithms "significance" parameter (i.e. made the algorithm less strict). This means the algorithm regularly stated, under default settings, that none of the edges were significant. I continually adjusted the significance parameter each time I ran the algorithm to ensure I got results.

I used the implementation of InfoMap in the igraph package in R. The InfoMap algorithm often returned the entire network in one giant cluster unless enough backbone was cut out. So I performed a balancing act, moving the significance parameter for backbone extraction up and down to get the algorithm to run and produce non-trivial results.

5.6.3 Modularity maximization

My method using networks to model multiple choice tests is very similar to Brewe et al. [2016]'s, so here I'll focus mostly on the differences.

5.6.3.1 Lack of projection

I defined the same bipartite network as Brewe et al. [2016], with answer choices connected to students who selected that answer choice. However, I didn't project it that network onto answer choices, as Brewe et al. [2016] did. Instead, I performed module detection directly on the full network. I believe this method is worth testing out and comparing to MAMCR because projection loses data. For example, suppose that in the projected network, question responses 1A, 2A, and 3A are all strongly connected to each other. It would appear that they should go together in a single cluster. However, it is possible that 1A and 2A are connected by a third of the students in a class, 2A and 3A by a completely separate third of the students, and 1A and 3A by the remaining students. In this case, the three questions don't intuitively necessarily belong together because no student ever picked all three choices. By taking a projection, we lose data on which students were connecting the answer choices, and potentially might find connections we would reject when looking at the full picture, although whether this is happening in practice is not yet known. To get some sense of what the extra data of the full network is telling us, I'll do clustering on the full, un-projected network and compare the results to MAMCR.

One side benefit of using the full network is that it includes student nodes in the clusters. This gives users the ability to find which students hold which models, potentially allowing differentiated instruction and more responsive teaching. I won't explore this aspect module detection on the full network here, but it might be of interest in future work.

5.6.3.2 Modularity

One method of detecting communities in such networks is to find a function of partitions which is higher for greater clustering, and then search for the partition that maximizes this function. This is the idea behind measuring the modularity of a network.

To measure the modularity, one first breaks a network down into modules, i.e. partitions the nodes in the network. Then each link within the network is either between two nodes in the same module (edge is within a module), or two nodes in different modules (edge is between modules). A certain fraction of edges will be within modules. This fraction will be high for networks that are very-cleanly partitioned (and for which we've chosen a good partitioning).

The fraction of edges that are within a module would not, by itself, be a good measure of a clustering, since it could be maximized for every graph by putting all nodes in the same module. There should instead be some sort of advantage to having a larger number of modules (or disadvantage to having a small number of modules) to encourage placing only those nodes that truly belong together in the same module.

Taking this into account, the *modularity* is defined as the fraction of edges that are within modules minus the expectation value for the fraction of edges that would be within modules if the edges were chosen between all nodes uniformly at random [Newman, 2006]. This definition encourages an algorithm looking for high modularity to create a lot of modules, if it can do so without making many links go between the modules, because this reduces the expected fraction of links between modules for a random graph. When the number of modules is large, the modularity can be close to 1 for a very-clustered network.

Brewe et al. [2016] use modularity in their investigation, but with a different epistemological role in evaluating clusters. Here, we will use an algorithm that attempts to search through all possible clusters maximizing modularity. Modularity is therefore used to generate clusters. Brewe et. al. instead use modularity as a check on the quality of the clustering developed by a different algorithm. They find that their clustering results in a graph with modularity 0.39, which they describe as "fairly high". So modularity is a common tool used in different roles in different analyses.

5.6.3.3 Adjustments to modularity

The standard concept of modularity requires some modification for the graphs I'm generating. The graphs are bipartite, so not all possible edges are allowed. This changes the standard for what's meant by the "random" graphs we use.

Barber [2007] defined a modularity for bipartite networks. This can form the starting point for detecting communities on networks of test questions. However, even this definition is not completely appropriate for the networks generated by single-response multiple choice tests.

The networks generated from single-response multiple-choice tests have additional structure beyond being bipartite. Each individual answer choice is a node, and students are constrained to choose exactly one answer choice for each question. We saw this caveat come up in the procedure for factor analysis on question responses. There, it caused a large number of -1 correlations which are set to zero to make the resulting matrix tractable. Here, it means that most random graphs constrained to have a bipartite structure still would not be valid graphs. Instead, the relevant random graphs are those resulting from students guessing randomly (with uniform probability) on each question.

Modularity on this sort of graph is not difficult to define; the paragraph above is essentially a definition. However, finding the partition that maximizes modularity is quite difficult. The number of partitions of a set of cardinality n is called B_n, the nth Bell number. Suffice it to say that they are large. The first few are $1, 2, 5, 15, 52, 203, 877, 4140, 21147, 115975, \ldots$. They grow super-exponentially, and the number of partitions for a 250-student class taking a 30-question, 5 choice test is B_{400}.

Any algorithm for modularity maximization needs to search a huge space. They use various techniques, such as simulated annealing (or analogies to it) to do this. Developing an algorithm to maximize a new definition of modularity would be its own research project, outside the scope of this one. For this project, I'll use the DIRTLPAwb+ algorithm described by Beckett [2016] for modularity detection on a bipartite network. Beckett implemented this algorithm in the R package "bipartite"[Dormann et al., 2017].

DIRTLPAwb+ generalizes Barber's work, finding modules on weighted networks. It was designed and tested on binary networks as well, so its design for weighted networks doesn't hurt its applicability to my networks.

Using DIRTLPwb+ is a compromise between practicability and the specific nature of the computational problem I'm solving. Unlike methods based on projection, DIRTLPAwb+ doesn't discard any information about the network, so it has the potential to find new things that previous networks analysis of PER concept inventories has not. However, it's optimized for a network without the extra structure of mutually-exclusive answer choices (one of which is required), so it is only a partial step towards a solution customized to the problem generated in this project.

DIRTLPwb+, like InfoMap, is a probabilistic algorithm, and can arrive at different solutions in different runs. I'll run it ten times and choose the highest modularity it finds. (The implementation in the bipartite package does this by default.)

Feeding networks from either real or simulated data into DIRTLPwb+ and searching for maximum modularity gives one new way to look for structure in both correct answer and distractors.

5.7 Testing Factor Analysis and Network Methods on Simulated Data

For each model I can use to simulate a class worth of responses (Rasch, factors, model analysis, Bayes nets) and for each analysis method (factor analysis, MAMCR, modularity maximization), I'll run two tests:

1. Generate a very large pool students, simulate a test result, run the analysis method, and compare the results to my expectations qualitatively

2. Choose two or three parameters in the model to vary. Define the ideal behavior of the analysis method, and measure quantitatively how well the analysis method performs as we vary these parameters using the variation of information between the ideal clustering and the analysis' discovered clustering.

The first step will let us see what types of results and insights the method can produce in an ideal case where the analysis method has plenty of data to work with. When there are other parameters besides the number of students, I'll make choices that intuitively should make the analysis method's task easier. The "take a look as see what you can see" approach to examining these results gives me latitude in judging what the different analysis methods do.

The second step gives an objective measure of how well the analysis method discovers structure when we know it is there and of a certain type.

I'll use a 30-item, 5-choice, single-response test throughout the analysis. This is the format of FCI and close to the format of MEGS. To make it easier to interpret the results, I'll set the correct answer for each question to A, and when the distractors have a clear clustering, response B's will be in the same cluster as each other for questions that contribute to that cluster, likewise for response C, etc. When questions have a clustering input into the simulation, questions in the same cluster will have adjacent numbers. E.g. questions 1-7 might be the first cluster, questions 8-15 the second cluster, etc. These conventions are equivalent to simply shuffling around the order of questions and answer choices in a more natural-looking test, and don't affect the analysis.

5.7.1 Measuring distance between clusters

When simulating a class of students taking a concept inventory, I sometimes know what the intended clustering of the questions or responses should be. For example, in the factors model, each question belongs to one of several factors, and we can expect a good analysis method (for distinguishing questions) to pick up on that. In model analysis, each question response belongs to a specific model, and so we can expect to discover those models in the data.

When the class size is small enough, the analysis methods will make errors in detecting the known cluster structure. (They may make errors even with large class sizes, but with small class sizes, eventually random noise makes this inevitable.)

I'll then want to compare how close the detected clusters are to the clusters put in by the simulation. I'll do this using the variation of information between the clusters, which is one such measure.

To give some intuition for variation of information values, I've calculated the variation of information for a few possible discovered and true clusterings for a 24-item set in figure 5.11. The tests I'll work on have 150 items, which is too busy for an illustration, but because the variation of information is an information theory-based metric, if we were to split each dot in the visualization into n smaller dots, the variation of information would remain the same, so we can imagine splitting each dot in the visualization into six smaller dots to approximate the actual conditions in the tests.

If $X = \{X_1, X_2, .., , X_k\}$ and $Y = \{Y_1, Y_2, .., , Y_l\}$ are both partitions of some set A, the definition of the variation of information is [Wikipedia]

$$\mathrm{VI}(X;Y) = -\sum_{i,j} r_{ij} \left[\log(r_{ij}/p_i) + \log(r_{ij}/q_j)\right]$$

where r_{ij} is the number of elements X_i and Y_j have in common, p_i is the fraction of

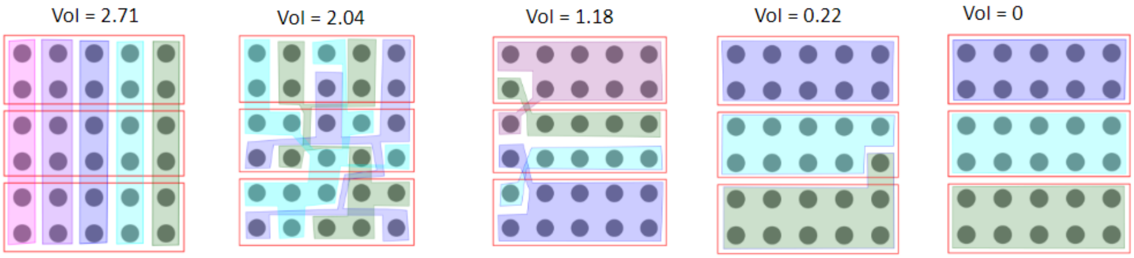


Figure 5.11: Some possible discovered clusterings (colors) compared to the true clusterings (rectangles) and the variation of information (labeled "VoI") between them

elements in A that are also in X_i and q_j is the same quantity, but for Y_j.

Other metrics would work as well for my purposes here, but the variation of information is easy to measure.

5.7.2 Testing against the Rasch model

The Rasch model is a sort of null model in my stable of models. I'll begin by running each analysis on this model to see what the results look like when the things we're looking for aren't present.

5.7.2.1 Expectations

The Rasch model treats all distractors the same, but correct answers are special (because students with high ability choose most of them and students with low ability choose few of them). So I expect all analyses to put the correct answers into a single cluster, or into a small number of clusters (separated by question difficulty).

The Rasch model doesn't have inherent cluster structure to the questions. They're all the same except for difficulty, so there's some chance that analyses run on the Rasch model will group questions into a "high difficulty" and "low difficulty" or similar. However, this isn't the behavior we're looking for in clustering. I want to know when analysis methods detect different types of questions, which I see as a different dimension than question difficulty, so if an analysis clusters by difficulty on the Rasch model, I count it as a strike against it in terms of finding useful insights from the results when run on real data.

The clustering I'll use to measure the variation of information is all correct answers in one cluster (i.e. 1A, 2A, 3A, etc.), all incorrect answers in another cluster.

5.7.2.2 Factor analysis results

I ran factor analysis on a simulation of the Rasch model with "friendly" parameters: student mean ability of 1 and standard deviation of 1, and 10 000 students. The high student mean results in an expected median score of about $1/(1+e^{-1}) \approx 0.73$. This should make the correct answers stand out significantly. The 10 000 students is a large number which should create cleanly-separated eigenvalues.

The goal of this test run was to determine that factor analysis will identify the correct clustering in the most extreme cases.

The result for the scree plot is shown in figure 5.12

Factor analysis returned a single factor. It consisted of answer choices 1A, 2A, 3A, etc. In other words, it found all the correct answers as a single factor.

This test shows that factor analysis is minimally viable in the Rasch model. It doesn't find all sorts of supposed structure in the distractors where non exists.

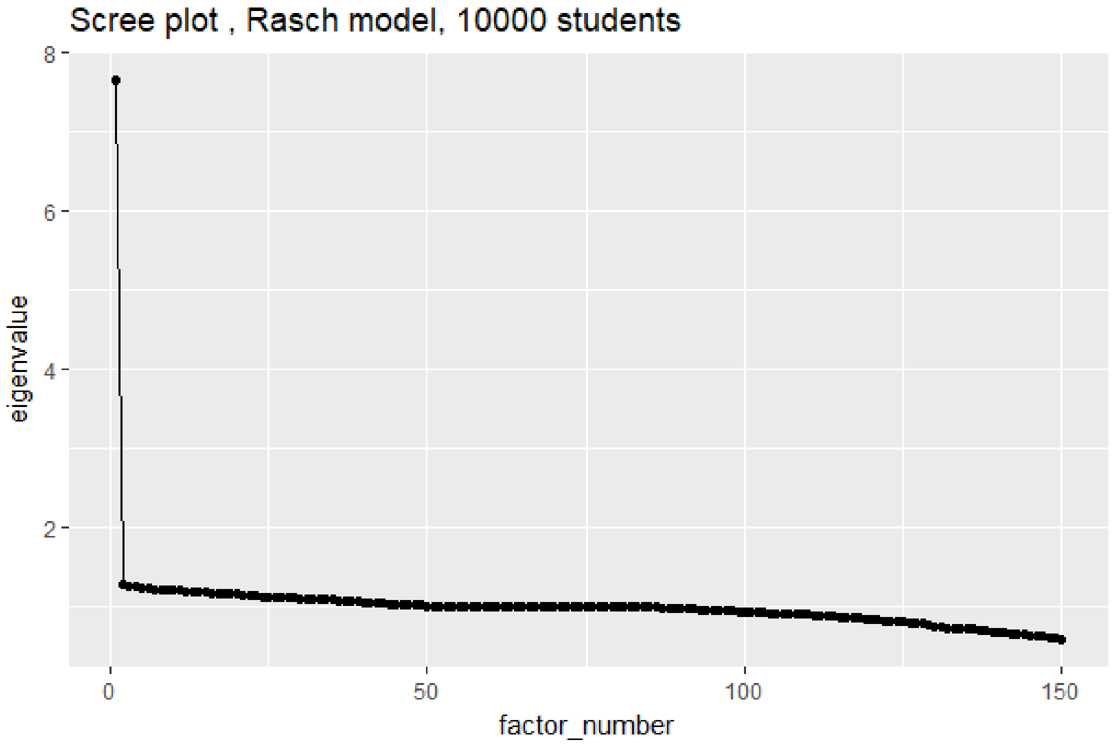


Figure 5.12: Eigenvectors from factor analysis on a simulation of the Rasch model with 1000 students. Student mean ability was 1 and standard deviation was 1.

To test factor analysis more extensively, I measured the variation of information between the clustering factor analysis found and the true clustering (all correct answer choices in one cluster, all remaining answer choices in the other cluster) over a range of parameters.

I tested with student average abilities set at 0, 30 questions on the test each with 5 choices. I varied the standard deviation of student abilities between 0, .5, 1, 2, and 3 and varied the number of students between 10, 30, 100, 300, and 1000.

While it would make sense to test varying the mean student ability as well, that would introduce one new parameter. To continue testing different mean student abilities in the factors model would then require four parameters (because I will vary the number of factors there), which would be unwieldy, so for consistency I kept the mean student ability fixed. A mean student ability of 0 matches well to our actual dataset, in which class average scores of about 50% are common.

For each set of parameters, I ran the test 10 times and took the median variation of information. The results are visualized in figure 5.13.

As expected, the variation of information goes down as we increase the number of students, meaning that with more students, factor analysis is less likely to "discover" structure that isn't there.

I was surprised to see the strong effect of increasing the standard deviation of student abilities on variation of information. The more spread in student abilities, the better factor analysis is at picking up on the correct answers. I suspect this is because there are more students with very high scores. These students make all the correct answers correlate to each other more strongly than if they were absent.

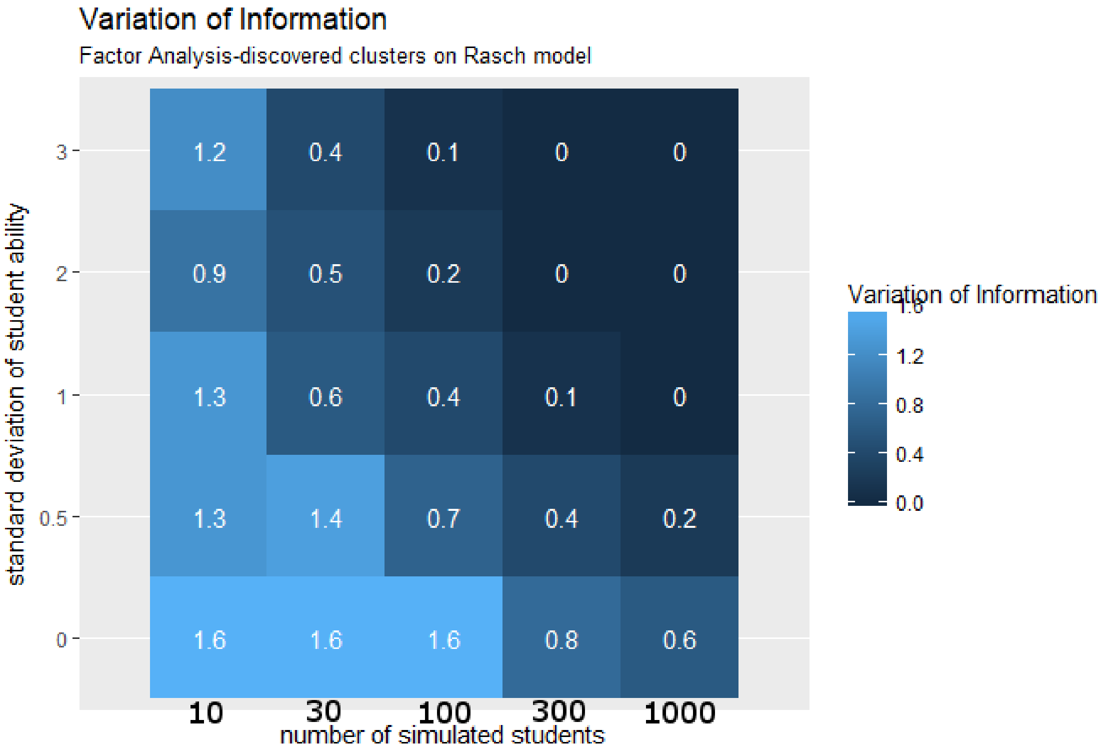


Figure 5.13: Variation of information between ideal modules and modules detected by factor analysis on Rasch model data with student mean ability 0.

This is worth keeping in mind when interpreting or deciding whether to use factor analysis. The less spread there is in the students, the more difficulty factor analysis is likely to have picking up signals, and the more likely its findings are spurious.

5.7.2.3 MAMCR results

I originally expected MAMCR to select the correct answers as a single cluster when given data from the Rasch model.

As shown in figure 5.14, MAMCR performed poorly on the Rasch model in my tests. For the most part, it put all answer choices in one giant cluster, which is why so many values in the table are identical.

Even by playing around with the parameters at will (which is possible because the implementation is very fast), I could only occasionally get more than one cluster as a result, and when I did, the clusters had no apparent structure. For example, setting student ability standard deviation to 3, mean to zero, number of students to 100, and class size to 30, I could get non-trivial clusterings using a significance parameter of 0.1 for the backbone extraction algorithm.

With these settings, InfoMap would return between ten and twenty modules. There would be one large module and the rest would have between 2 and 10 members. There wasn't any obvious association between them. The correct answers were spread between the single large module and the small ones.

It appears that my implementation of InfoMap failed on the Rasch model.

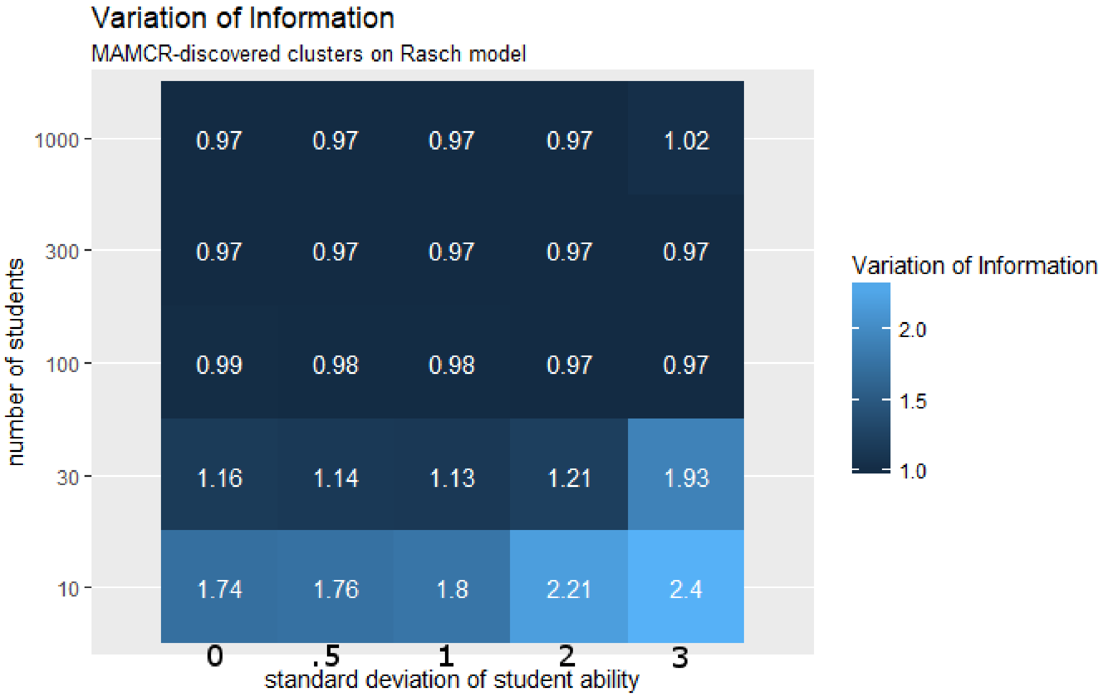


Figure 5.14: Variation of information between ideal modules and modules detected by factor analysis on Rasch model data with student mean ability 0.

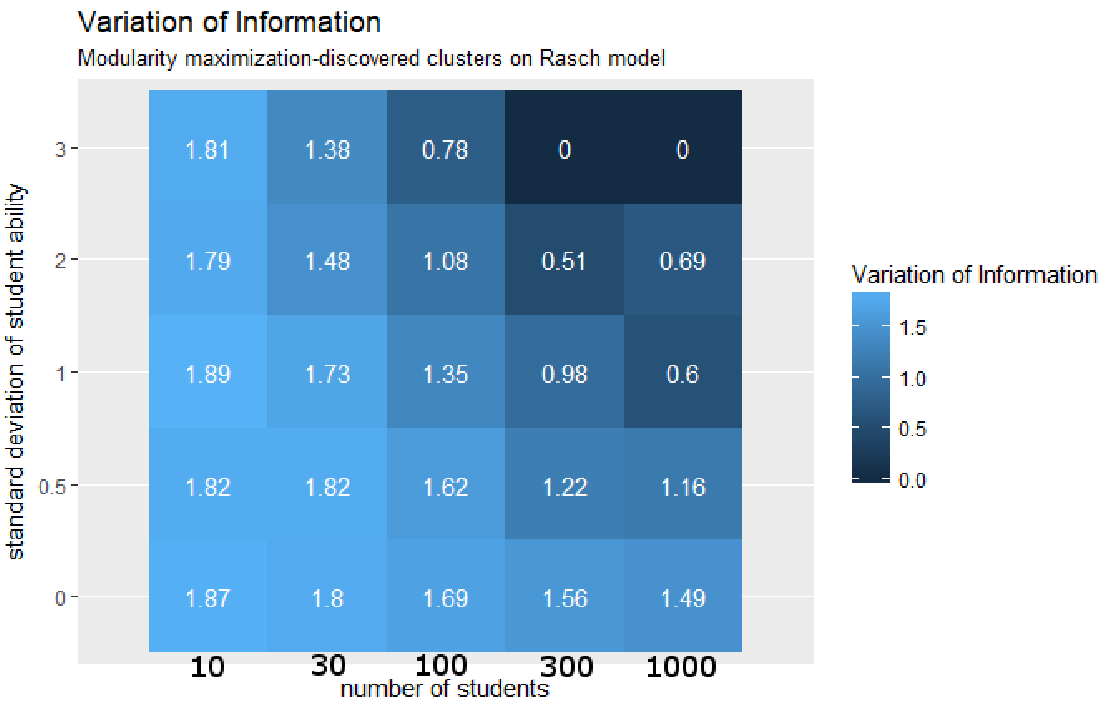


Figure 5.15: Variation of information between ideal modules and modules detected by modularity maximization on Rasch model data with student mean ability 0.

This means I can't be confident in its results on real data without other evidence that the results are significant, because my implementation of MAMCR can return results where there should be none.

5.7.2.4 Modularity Maximization results

In figure 5.15 I've plotted the variation of information between the modules discovered by modularity maximization and the ideal module structure for the Rasch module. I've tested the same parameter space as for factor analysis.

Modularity maximization follows the same general performance trend as factor analysis. Increasing both number of students and standard deviation of students improves the match between detected and ideal modules.

For any given set of parameters, modularity maximization performed worse than factor analysis.

This means that modularity maximization passes the Rasch model's basic test. It doesn't detect structure that isn't there, at least when given enough data of high enough quality. However, the Rasch model's results in themselves don't give any reason to prefer modularity maximization over factor analysis.

5.7.3 Testing against the factors model

The factors model is, for the most part, a second null model for the purposes of this chapter. I'll run each analysis method on the factors model to see to what extent they find the moderate amount of structure that's in there, while not finding extraneous stuff.

5.7.3.1 Expectations

As in the Rasch model, distractors are not distinguished from each other in the factors model. However, questions do fall into several distinct clusters, or factors. The correct answers from these clusters should definitely cluster together. It also makes sense for the incorrect answers from a factor to cluster together, since they all get more or less likely to be chosen by a single student together (i.e. if the student has very high ability on that factor, all incorrect answers for that factor get less likely to be chosen).

The clustering I'll use to measure the variation of information is all correct answers from a single factor in one cluster, all incorrect answers from a single factor in a cluster.

The variation of information, even for well-spread out students with standard deviation 1 in their ability, remains large below 100 students, suggesting that about 100 student is a bare minimum to perform factor analysis. Otherwise, factor analysis will discover many factors it shouldn't or miss the factors it should find.

5.7.3.2 Factor analysis results

I've already visualized the results of running factor analysis on the factors model in section 5.6.1.3, where I ran several such tests to determine which algorithms for deciding how many factors to keep were most effective. Those tests demonstrated that, given a large enough class size, factor analysis generates one large eigenvalue for each factor I put in the factors model.

To test factor analysis against the factors model systematically, I'll vary two parameters while finding the variation of information between the ideal modules (all correct answers for a single factor in a single module) and the modules found by factor analysis.

I know that having more students will lead to better results, so I'll visualize three different student counts (100, 300, 1000). I'll drop the 10 and 30 student levels because running factor analysis on the Rasch model showed that that's very unlikely to produce satisfactory results.

I'll vary the number of factors between 2, 3, and 6. 6 factors is about the most that factor analyses of concept inventories attempt to find. 1 factor reproduces the Rasch model so we can leave it out, and I'll leave out 4 and 5 because they wouldn't add much new information between 3 and 6.

I'll keep student ability at a mean of zero, as I did for the Rasch model. I'll also fix standard ability standard deviation to 1, because at this level the Rasch model worked very well with many students and very poorly with few students. It also represents a spread such at 68 percent of students fall within the range of 27% and 73% (abilities of -1 and 1 respectively). This is a realistic range for MEGS data, though it has a few more low-scoring students than is realistic. This might be partially because random guessing has an expected score of 20%, so that a student with an ability of -100, who is overwhelmingly likely to score zero on the test, actually knows more about the material than a student with ability -1.39, whose expected score if 20%, in the sense that in order to secure a score of 0, the test-taker needs to know enough about the material to avoid getting an answer right by accident.

Running ten trials for each set of parameters and keeping the median, I found the results shown in figure 5.16.

Compared to the Rasch data, detecting factors was more challenging. Even with 1000 students, factor analysis made small mistakes on average.

This suggests that factor analysis requires large data sets, probably gathered over several years or brought together from several institutions in order to have enough data to see real signals.

The MEGS data does exceed 1000 data points because it has been collected

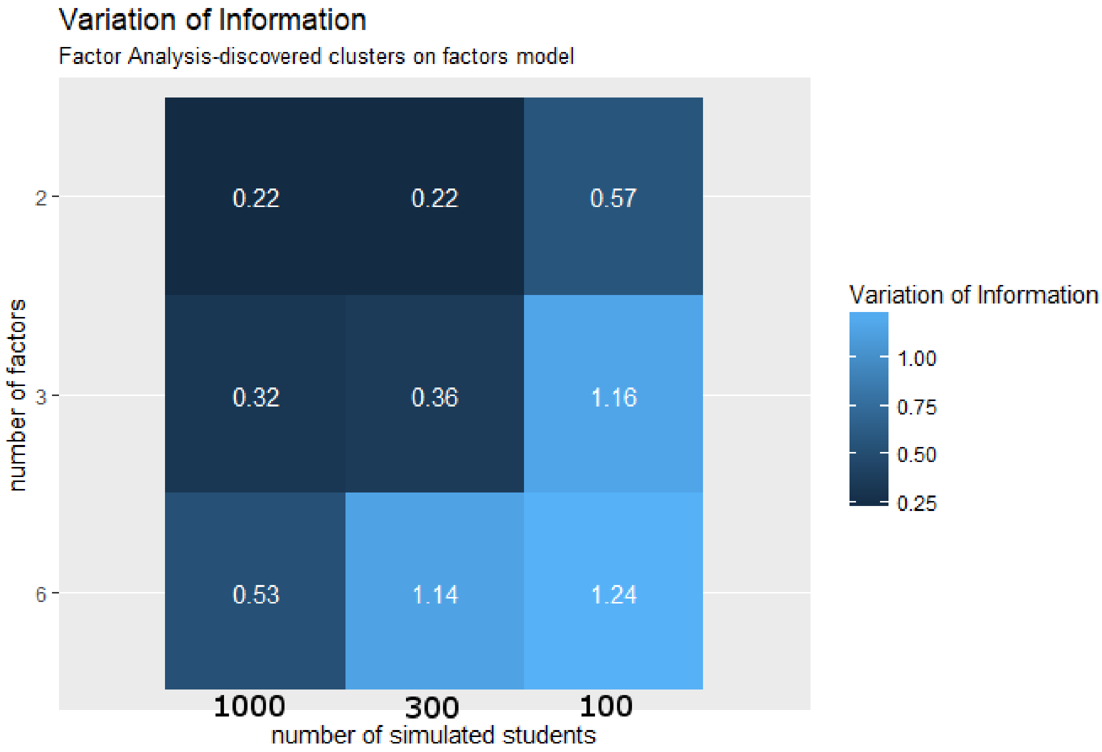


Figure 5.16: Variation of information between ideal modules and modules detected by factor analysis on factor model data with student mean ability 0 and student standard deviation 1.

over several cohorts of students, so factor analysis is still a viable alternative there, according to the results from the factors model.

5.7.3.3 MAMCR results

Because my MAMCR implementation seems to depend on the significance parameter quite sensitively, I began by exploring some example simulations.

I was able to fine-tune the parameters to get a non-trivial module structure, but non consistently or systematically. For example, I tried 4 factors, 1000 students, ability mean 0, ability standard deviation 2, and significance 0.02. This usually gave four modules, but sometimes gave two. There was often one large module with most of the answer choices. The modules weren't focused on correct answers. Instead, they often included answer choices from the same question. For example, one run has all the choices for question 1, 2A, 2B, 2C, 3A, 3B, 3D, 4A, 4C, 4E, 5B, 5D, 6A, 7D, 8A, 8B, and 9A as a single module, and no other answer choices. The next three modules had answer choices from the first 9 questions that were omitted from the first module. No answer choices from questions above 9 were included.

Evidently, backbone extraction worked fairly poorly on data from the factors model, as a large number of answer choices, including correct ones, were discarded.

To test MAMCR on the factors model systematically, I chose the same parameters as I used testing factor analysis on the factors model. I set the significance to 0.016, about the smallest value for which the algorithm would run. (At smaller values, the backbone extraction rejected all nodes and returned an empty network for some parameter values. At larger values, the algorithm tended to return just one giant cluster with all answer choices in it.) I obtained the results in figure 5.17.

As the trials I described earlier in this section suggest, my implementation of MAMCR did not detect the structure of the factors model well in general. Having more students did generally lead to slight better results, hinting at some level of validity, but this is mostly because with large student sizes, the algorithm was more likely to choose one big cluster, which is an improvement in variation of information over a number of smaller clusters almost completely misaligned with the ideal module structure.

This test of factor analysis suggests that MAMCR is not able, under my implementation, to discover true connections.

5.7.3.4 Modularity Maximization results

As shown in figure 5.18, modularity maximization performed poorly on simulations from the factors model.

Increasing the number of students led to only a small improvement, and even with only two factors, modularity maximization's median performance was quite poor compared to factor analysis.

To try to get some insight into why modularity maximization was performing so poorly, I ran it on a set of factors data with 6 factors and 1000 students. I ran two trials and looked at the results.

In both cases, modularity maximization found 5 modules, not the expected

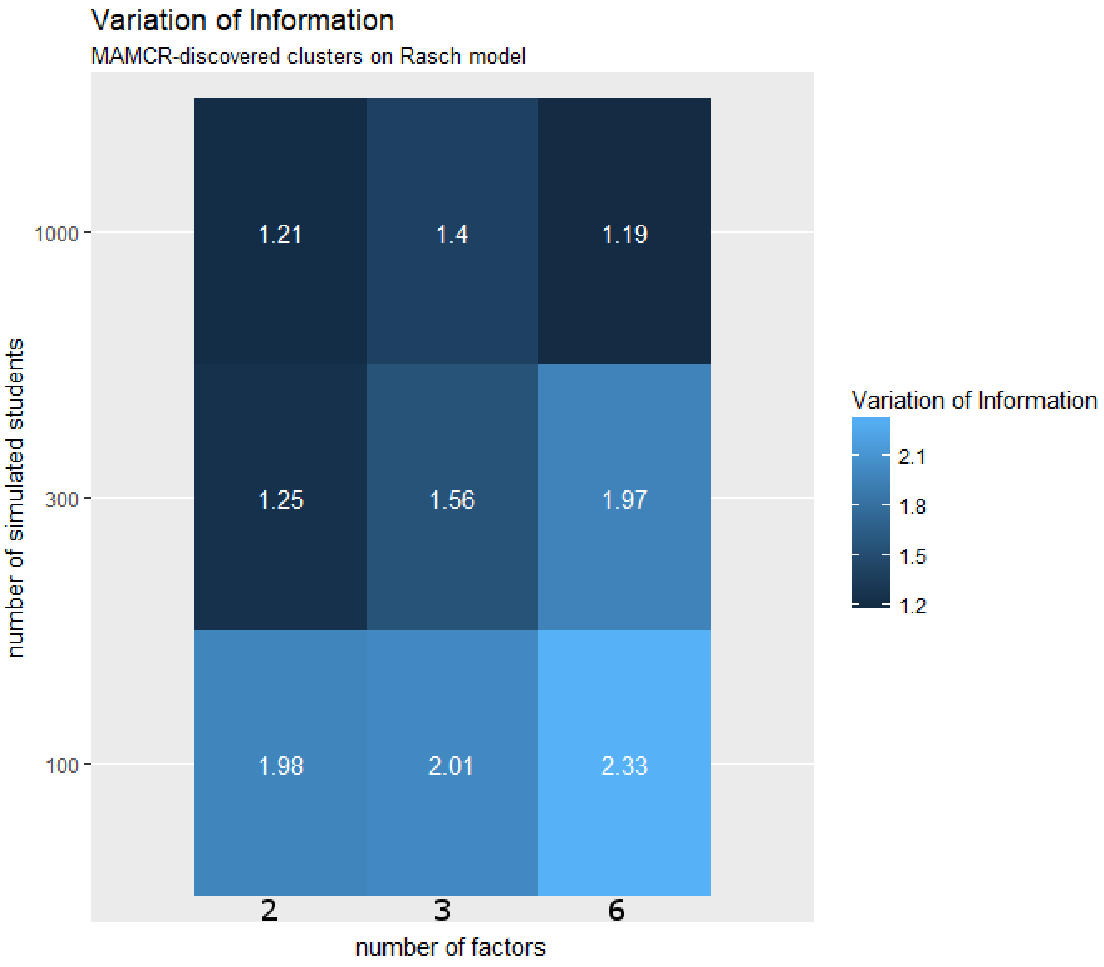


Figure 5.17: Variation of information between ideal modules and modules detected by MAMCR on factor model data with student mean ability 0 and student standard deviation 1.

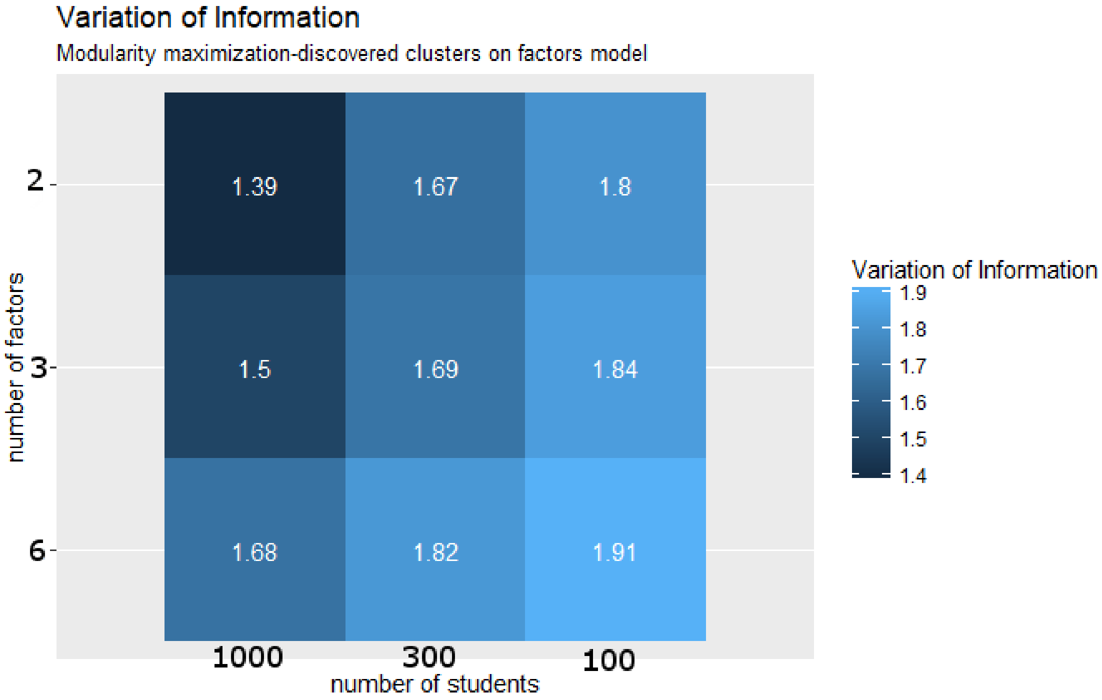


Figure 5.18: Variation of information between ideal modules and modules detected by modularity maximization analysis on factor model data with student mean ability 0 and student standard deviation 1.

six. There was no module with all the correct answers, either. The modules were of roughly equal size, mostly around 30 answer choices, but from as few as 22 to as many as 40. But by inspection, there were only occasionally groups of correct answers in the same cluster.

For example, one factor contained the correct answers to questions 10, 11, 12, 13, 14, 20, 21, 22, 23, 24, 25, 26, 28. This is far too many consecutive correct answers to happen by chance, but none of the correct answers from the first factor were in this module.

Next I ran a test with 2 factors and 1000 students. Modularity maximization found 7 clusters. This was again too many. The clusters varied in size from two to 45 answer choice nodes. Modularity maximization seemed largely at a loss outside of correct answers. It did find one cluster with the correct answers to questions 1 - 15 (the entire first factor) and another cluster with the correct answers to questions 16-30 (the second entire factor). These clusters didn't include any incorrect answers to the questions for which they had the correct answers. So in this sense, modularity maximization found the existing structure. But it failed to isolate that structure from noise, since both clusters contained extraneous incorrect answers.

Modularity maximization mostly discovered the single-choice multiple-response nature of the test as well, in that in all the runs I've described, it was very rare for multiple responses to the same question to appear in the same cluster.

One thing to remember about modularity maximization, as I've used it in this chapter, is that it's not just trying to find connections between questions. It finds connections between students as well, and this may be biasing it towards balancing the number of answer responses in each cluster. If a cluster had only five answer responses in it, and students in that cluster would necessarily have many links out to other cluster. So although those five answer responses may be tightly linked, they still might not be favored by modularity maximization.

This hypothesis suggests that modularity maximization should perform much better on model analysis simulations, where the ideal clusters all have the same size, as opposed to the factors model, where some are much larger than others.

For analyzing the MEGS, this means that modularity maximization would not be likely to detect any signal in which only a few answer choices are connected. Additionally, it appears that modularity maximization may be likely to find a real signal, but then add on extraneous answer responses into the modules, so we shouldn't expect every answer choice in a module discovered to be meaningful.

5.7.4 Testing against the model analysis model

The model analysis model is where analysis methods have the best chance to prove themselves. There is very clear structure to all the answer choices, and the analysis methods should find it.

5.7.4.1 Expectations

Each question belongs to a set of expert-equivalent questions in model analysis. Within those questions, answer choices belong to a set of distinct models. Clustering should happen at the level of answer choices to be useful in my analysis, because the goal is to test what information is available in the distractors students choose.

The clustering I'll use to measure the variation of information is all answers corresponding to a single model in the same cluster together, and no answers from different models in a cluster.

5.7.4.2 Factor analysis results

As with the factors model, we've already seen scree plots from factor analysis in section 5.6.1.3. These results showed that factor analysis should ideally be able to pick out all the models in model analysis with a large sample size. I know this because we tested model analysis with 15 models, three correct and 12 incorrect. This led to a visually-apparent structure in the scree plot in which there were three large eigenvalues and 12 medium eigenvalues, followed by 135 small eigenvalues.

An appropriately-tuned algorithm should then have been able to identify all the models, because one model corresponded to each eigenvalue. However, none of the methods I tested for analytically determining eigenvalue cutoffs were able to identify the correct number of eigenvalues with any regularity. Instead, I chose a method which, on the original test data, only identified the three correct models.

It's worth seeing whether this analysis holds across some variety of parameters. I'll test 2,3,and 6 models with mixedness of 0, .1, .3, and .6.

There's no need to test a mixedness of 1 because this reproduces the factors model. I'll use 1000 students for each test, because previous tests have established that factor analysis struggles to correctly identify the appropriate models in the

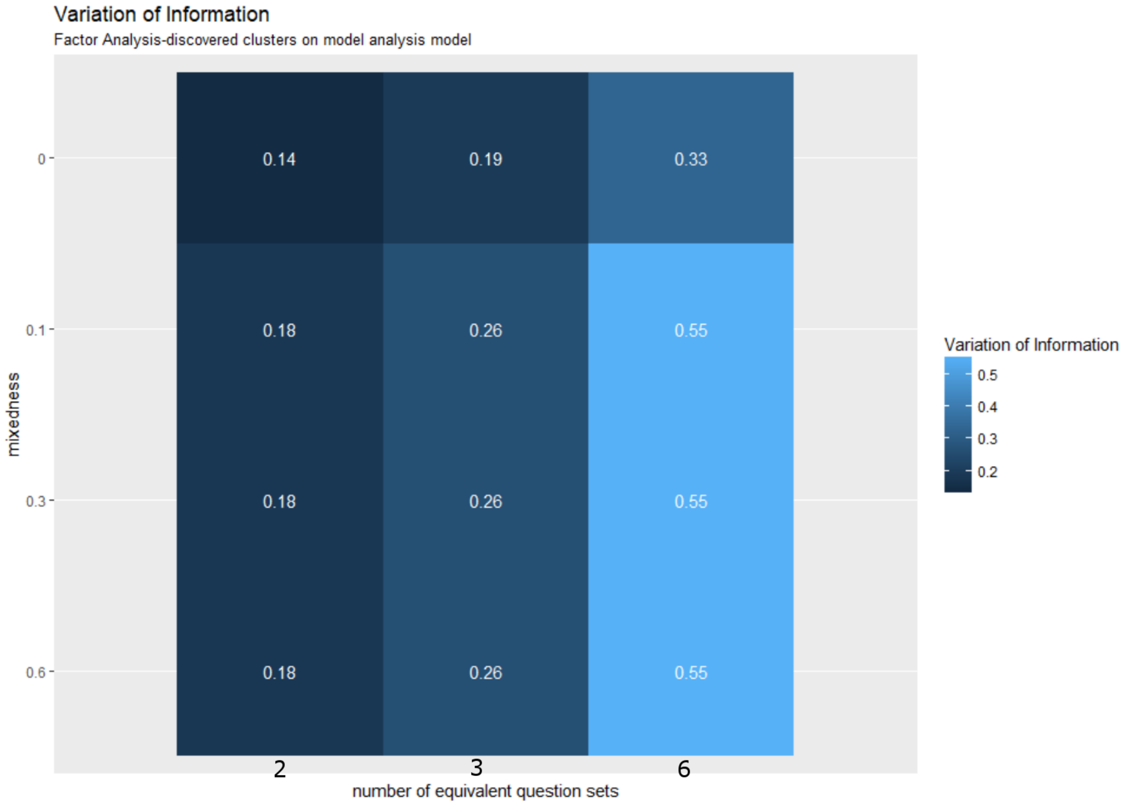


Figure 5.19: Variation of information between ideal modules and modules detected by factor analysis on model analysis model simulations.

model analysis data, so simulating small numbers of students is unlikely to be productive. I want the results of the simulation to be relevant to analysis of the MEGS, which has on the order of 1000 student responses, so this is an appropriate figure to use. Not varying the number of students allows me to keep my search to two parameters.

Running ten simulations and keeping the median variation of information for each one, the results I got from the model analysis simulations are shown in figure 5.19.

The first thing I note is that factor analysis did fairly well. Given that when I was deciding how many factor to keep, factor analysis never discovered the incorrect models, I expected the variation of information to be significantly higher for model analysis than for the factors model, but the opposite was true.

Next, as expected, increasing the number of expert-equivalent question sets made the structure more difficult to detect, just as increasing the number of factors does in factor analysis.

I expected decreasing the mixedness to improve the model's fit, and it did, but the effect was not dramatic. With very low mixedness, students are consistent in which incorrect model they choose. This helps the answer choices corresponding to the incorrect model to correlate highly, giving factor analysis a better chance to detect them, but even with fairly high mixedness of 0.6, factor analysis wasn't catastrophically harmed, especially when there were only two factors in the underlying data.

This test bodes well for factor analysis. It was able, on median, to come quite close to discovering the true stucture of model analysis on a data set of 1000 students. (See figure 5.11 for a visualization of how close two clustering must be to earn a certain variation of information score.)

5.7.4.3 MAMCR results

To experiment, I first tried running MAMCR on a class of simulated data from model analysis using 1000 students, 3 expert-equivalent question sets, and mixedness 0.3.

As in other trials of MAMCR, the results depended strongly on the significance parameter I put into the algorithm. At 0.04, the algorithm often didn't run, or returned just a few nodes, because the backbone was empty or almost empty. At 0.1, the algorithm consistently lumped all answer choices into a single large cluster.

With a significance parameter of 0.6, MAMCR didn't through out many nodes, and found three modules. The answer choices in the modules were from different questions, though, so they didn't correspond to the three expert equivalent questions, and were too large to correspond to individual models (of which there were 15).

I ran MAMCR on simulations of 1000 students, varying the number of expert-equivalent questions and the mixedness. The results are shown in figure 5.20. I kept the significance parameter at 0.06, the best value I found from experimentation.

Although the variation of information in figure 5.20 is generally low, this is deceptive. My MAMCR implementation was throwing out a lot of nodes, and variation of information is ideally to be used on partitions. The low variation of information simply reflects the large number of nodes discarded.

My conclusion was that MAMCR was not effective on model analysis. Model analysis presented the most and clearest structure to distractors of my models for simulating a class of data. Given that MAMCR didn't find any significant results here, I conclude that it wouldn't be informative to run MAMCR on the MEGS data or Bayes nets.

However, I think it's possible that my implementation of MAMCR is not as effective as that of Brewe et al. [2016], perhaps due to differences in the network backbone extraction, or perhaps we used different implementations of InfoMap.

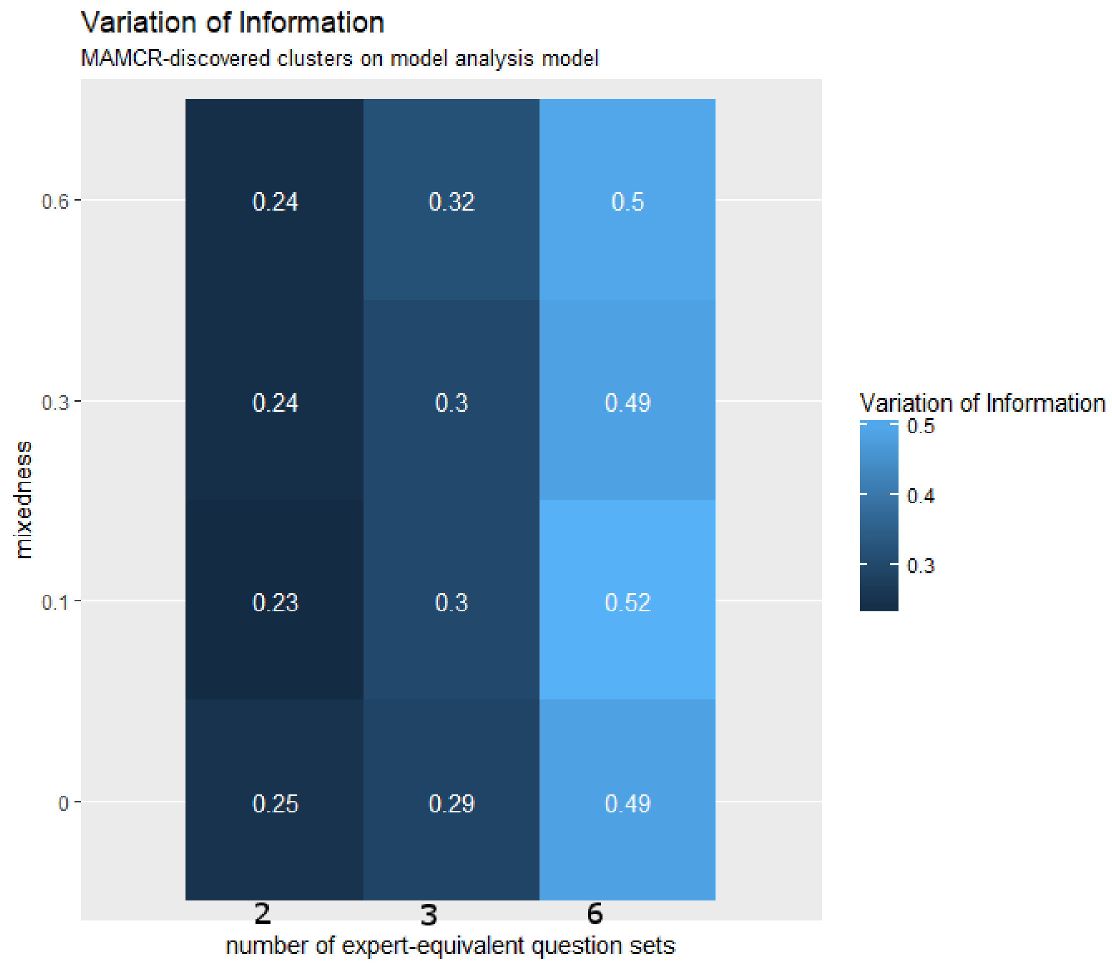


Figure 5.20: Variation of information between ideal modules and modules detected by MAMCR on model analysis model simulations with 1000 students.

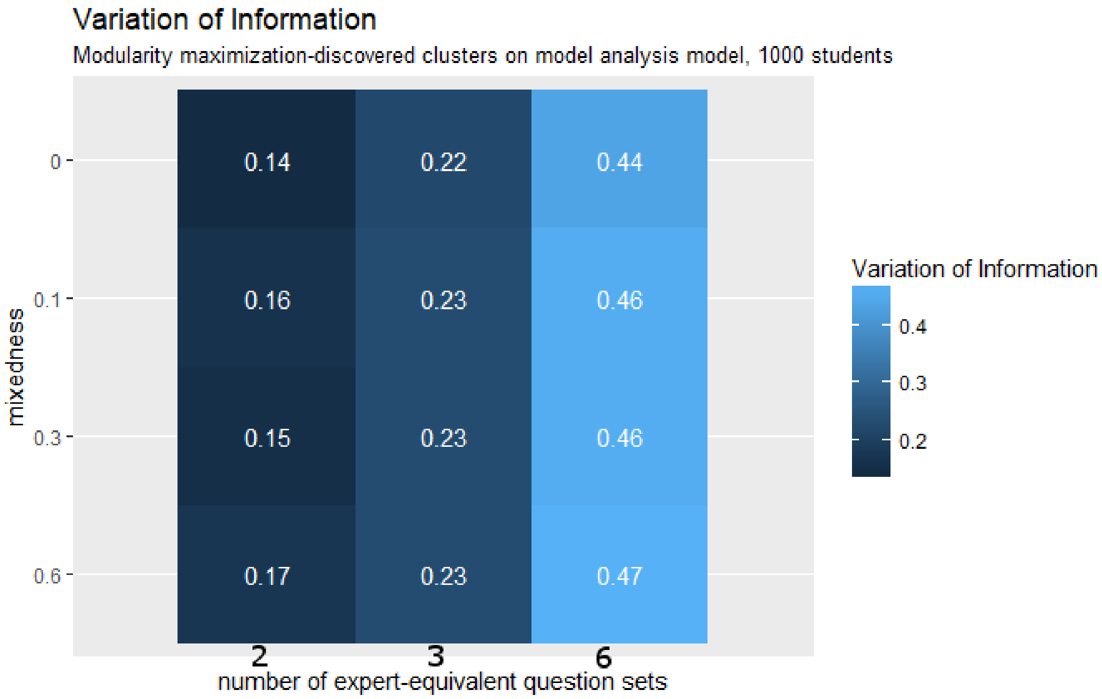


Figure 5.21: Variation of information between ideal modules and modules detected by modularity maximization on model analysis model simulations with 1000 students.

MAMCR worked well on FCI data for Brewe et. al. it is worth revisiting in the future to see if my implementation can be improved.

5.7.4.4 Modularity Maximization results

The results for modularity maximization on the model analysis are shown in figure 5.21.

Because the variation of information is generally low, modularity maximization works well on model analysis simulations, given a student sample size similar to what we'll see on MEGS data.

As expected, having more models meant the variation of information increased. For interpreting real data, this means that I should be wary of interpreting a very large number of clusters. It also means that the more real clusters there are in the data, the harder they will be to detect.

To test whether modularity maximization was fundamentally unable to find all the existing structure or just needed more data, I did a single run on a group of 10,000 students with 3 sets of expert-equivalent questions and mixedness 0.3.

Even with 10,000 students, the fit was not perfect. Modularity maximization found only four modules, one of which contained almost all the correct answers. Ideally, it should have found 15 modules of equal size. This suggests that the method itself has inherent limitations; it isn't just a lack of data that results in imperfect fits.

5.7.5 Testing against Bayes net model

I described my scheme for simulating Bayes nets in section 5.5.4. The Bayes net model, depending on its parameters, is the most complicated of my four models and can present the toughest test to my analysis methods.

5.7.5.1 Expectations

The structure behind Bayes nets is not nearly so straightforward as for model analysis. Any number of resources can contribute to any answer choice, either promoting or inhibiting it. If we take some particular resource, all the answer choices it interacts with might be expected to cluster together, or maybe just the ones it promotes (with the ones it inhibits a separate cluster). But this can't be universally true. It leads to no single clustering structure because a single answer choice might be influence by multiple resources.

For this reason, I'll mostly analyze Bayes nets results just by looking at the networks and the resulting clusters and analyzing them qualitatively, skipping the variation of information. However, my first Bayes net A (see section 5.7.5.2) is simple enough to admit an ambiguous ideal module structure, so I'll measure variation of information for that.

I will generate three different Bayes nets of increasing complexity, visualize them, and analyze them with each of factor analysis, MAMCR, and modularity maximization.

This adds to the analysis by presenting the most-realistic case, in the sense that it is messiest. The results can tell us what sorts of clusters the analysis methods find when the real situation is more complicated than the types of assumptions that underly the analysis methods themselves.

5.7.5.2 Bayes net A - 2 resources, p=.3

This is the simplest Bayes net. After removing nodes with no connection, I visualized the net with igraph's plot function. The result is in figure 5.22

The answer choices for this model fall into four modules:

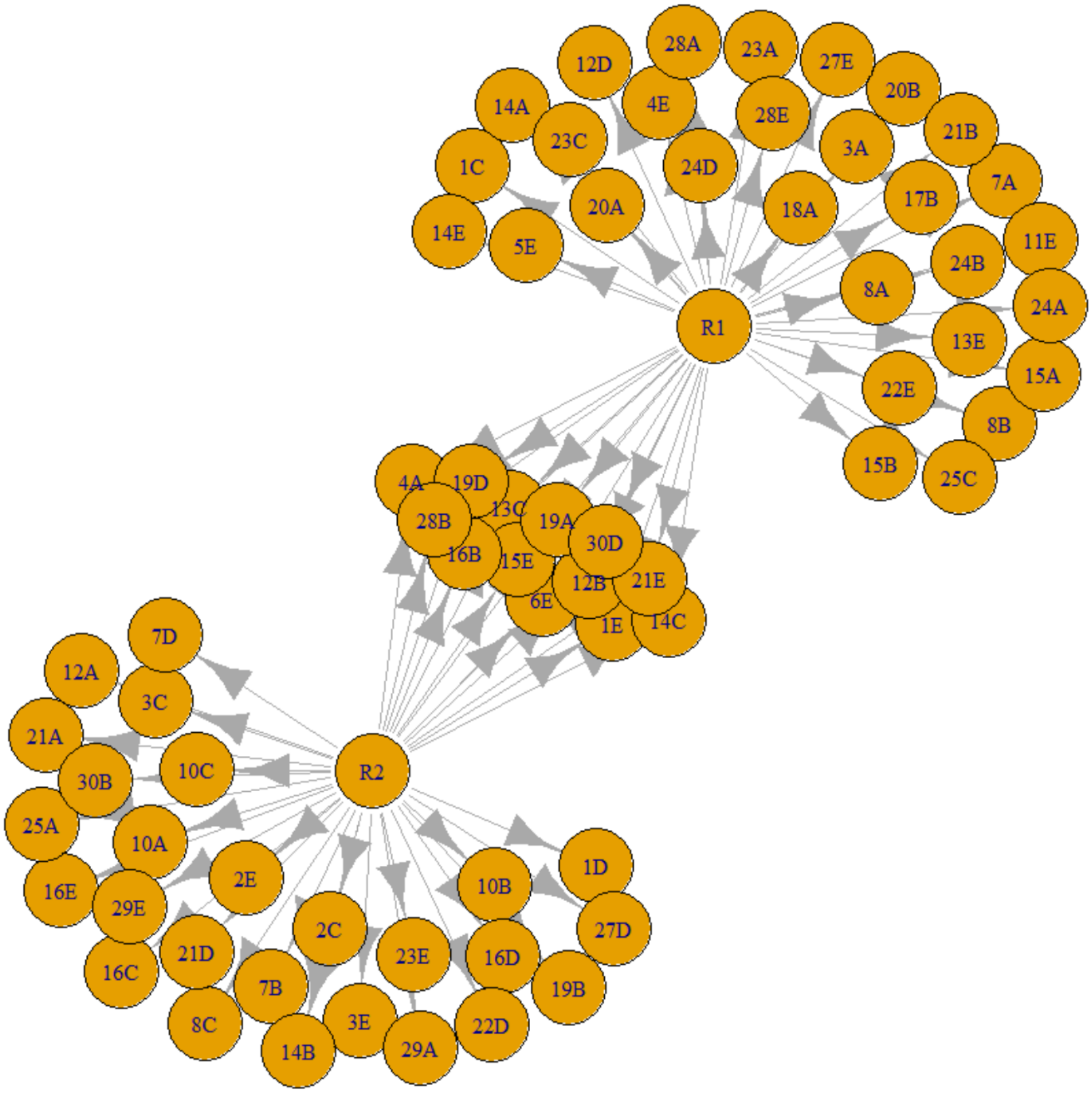


Figure 5.22: A simple Bayes net with two resources.

Table 5.1: Variation of information for analysis methods run on simulations from Bayes net A

Students	Factor Analysis	MAMCR	Modularity maximization
10	1.86	1.52	2.23
30	1.79	1.30	2.21
100	1.57	1.30	2.16
300	1.68	2.31	2.14
1000	1.77	0.31	2.12

- answer choices connected only to resource R1
- answer choices connected only to resource R2
- answer choices connected to both resources
- answer choices not connected to either resource

This provides a structure against which I can test the performance of factor analysis, MAMCR, and modularity maximization with varying numbers of students.

The variation of information that results from running these analysis methods on Bayes net A with varying numbers of students are shown in table 5.1.

The surprising result is that increasing the number of students didn't have much effect on the variation of information between the discovered and ideal clusters.

As in previous investigations, my MAMCR implementation simply put everything in one large cluster. The low variation of information for 1000 students occurred because it discarded a bunch of answer choice nodes, subverting variation of information as a useful measure. In a second trial with 1000 students, MAMCR still discarded most nodes, but found 8 clusters with variation of information 2.88. This reinforces my earlier conclusion not to use MAMCR on MEGS data until I improve my implementation of it.

Next I looked at the modules discovered by factor analysis run with 1000 students. Instead of finding 4 factors, it found 21, one of which contained over 100 elements. I repeated this several times, and factor analysis alternated between finding a large number of factors (e.g. 20) and a small number (e.g. 3 or 4). However, it always put most of the nodes in the same large factor. The smaller factors weren't, by visual inspection, closely related to any of the ideal clusters. In general, factor analysis failed to find the structure in this Bayes net.

Finally, I look at the modules from modularity maximization run with 1000 students. Five clusters were discovered instead of four, and as with the factor analysis results, I saw no obvious relationship between the contents of the clusters and the ideal clustering.

Overall, none of my analysis methods passed the test of a simple Bayes net. It's possible this is because I need to change the implementation. For example, I've set it currently so that each connection from a positive resource increases or decreases the log-odds of that node by 1 (before normalizing across the answer choices for a question). Alternatively, I could change connection from being both reinforcing and inhibitive to only reinforcing, but the detected clusters weren't, for example, all the reinforced answer choices connected to a single resource.

As it stands, I take this experiment with a Bayes net to show that my analysis methods, though they work on some types of simulations, produce mostly-meaningless results on other types. I'll need to be very cautious in interpreting MEGS data.

5.7.5.3 Bayes net B - 4 resources, $p = .3$

Next I generated a more-complicated Bayes net. With 4 resources, there were connections between resources, and answer responses could be connected to resources in multiple, complicated ways. The network is shown in figure 5.23

There is still an apparent cluster structure, so we can run this network through each analysis method and see if they return any clusters similar to those show in the figure. It's hard to say that there is an ideal clustering, but we might still be able to recognize elements of clusters as corresponding to the things we can visually detect.

Factor analysis run with 1000 students returned 11 factors, with a single factor containing 109 answer choices. The smaller clusters had between 3 and 6 elements, but never more than two of those elements were near each other in figure 5.23.

Next I ran MAMCR with simulated data from 1000 students on the same network. I used a significance parameter of 0.2. It found seven clusters, but with a maximum of six answer choices per node. As with factor analysis, I didn't see any connections between the clusters and figure 5.23.

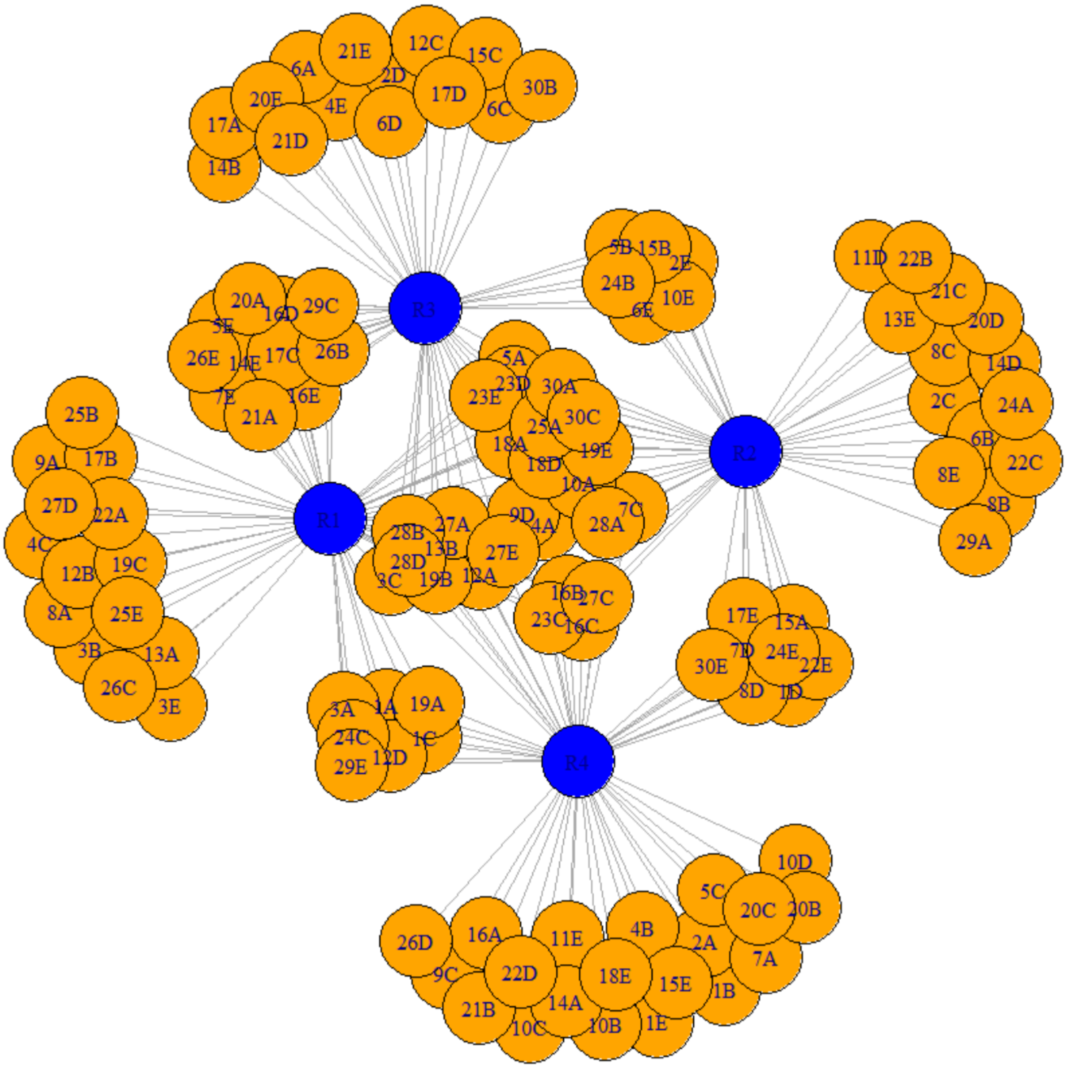


Figure 5.23: A Bayes net with four resources (blue) and 150 answer choices.

Finally I ran modularity maximization with 1000 simulated students on the same Bayes net. It discovered six modules, with between 17 and 30 answer choices each. But again, there was no clear connection between the clusters generated and figure 5.23

5.7.5.4 Bayes net C - 10 resources, $p = .2$

I made one final Bayes net, this time with 10 resources, shown in figure 5.24.

A cluster structure is no longer visible, and attempts to run and compare factor analysis or network-based clustering algorithms were too unsuccessful to be worth noting.

This example shows that if there is some "real" structure in the data, it might still be extremely subtle and difficult to detect via multiple choice tests. Just because analysis of multiple choice tests fails to validate a model of student cognition doesn't mean the model is incorrect. It simply isn't so simplistic as to have easily-observable results in multiple choice data.

5.7.6 Conclusions from tests on simulations

After running three different module-detection algorithms on four types of simulations of multiple choice test data and a variety of parameters, I believe I must use extreme caution when attempting to interpret MEGS data with more-complicated data analysis techniques than those used in chapter 4.

On some very simple data sets, factor analysis and modularity maximization

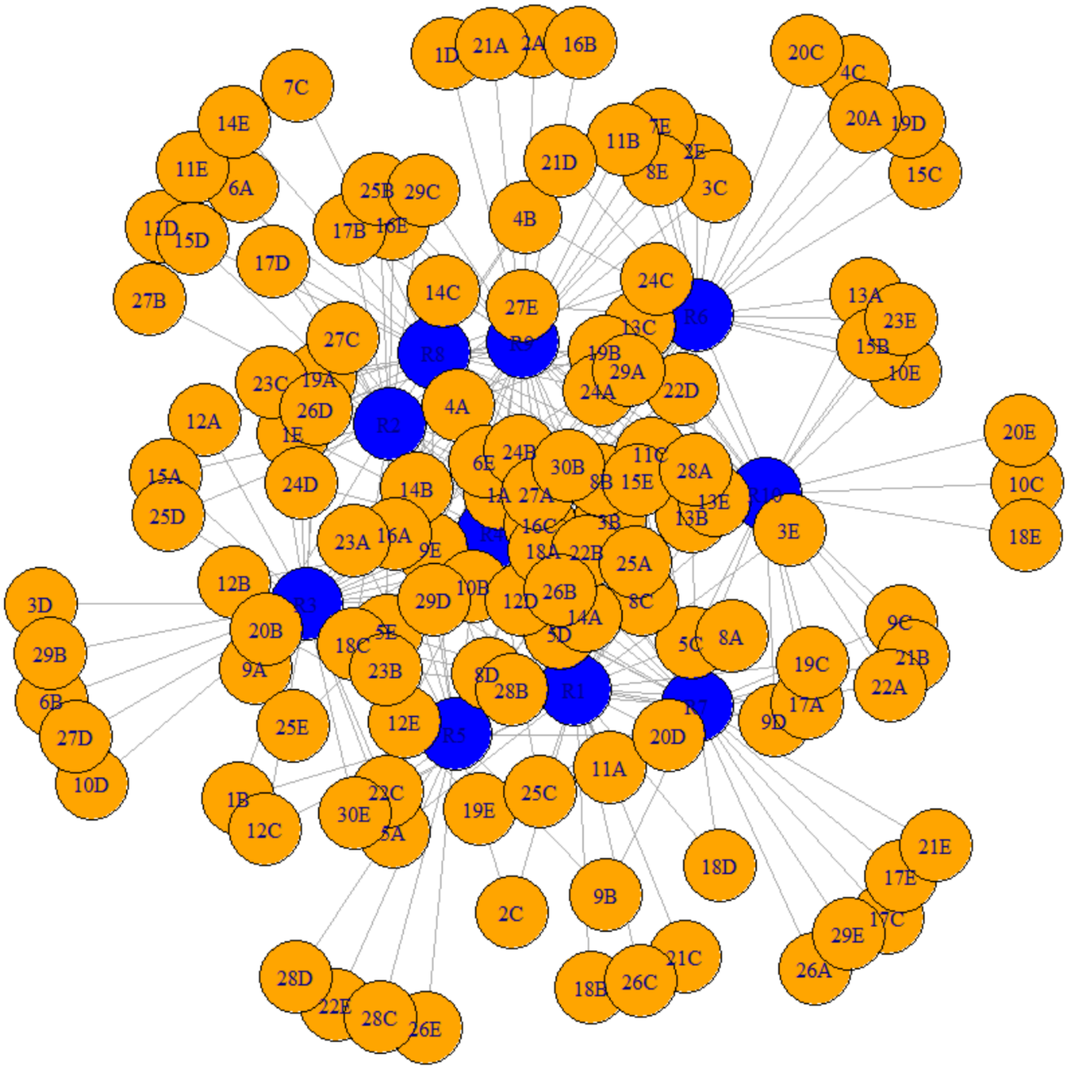


Figure 5.24: A complicated Bayes net with ten resources (blue) and 150 answer choices.

were able to detect the existing structure, sometimes even with perfect accuracy. Generally, factor analysis performed better in my tests, although see section 5.10.2 for notes on how I might improve modularity maximization to give better results.

Once the underlying models started to become more complicated, the analysis methods faltered, then failed completely. Factor analysis, for example, could pick out the different models in model analysis according to a read of the eigenvalues on a scree plot, but it was not able to consistently pick out all the right factors without human intervention, and the clean-looking separation of eigenvalues we saw in section 5.6.1.3 is unlikely in real data.

I take the simulations described above as tests only of my implementations of factor analysis, modularity maximization, and MAMCR. These methods are all based on widely-used and extensively-validated techniques that give results on real data that are sensible and informative. However, they may not work on every data set, and may only work given a lot of fine-tuning.

So my conclusion is that I won't place a strong epistemic weight on the techniques for analysis of MEGS data. This includes refraining from claiming that a module structure to MEGS data doesn't exist just because I haven't found it. I don't, however, think the tests described in this chapter have a strong bearing on interpreting other implementations of the same or similar techniques.

5.8 Analysis of MEGS data

For a complete description of the MEGS data, please see chapter 4.

For this analysis, I divided the MEGS into all pre-semester data and all post-semester data. We know that experts and novices, when asked explicitly to group physics problems into different groups, find very different structures [Chi et al., 1981]. We might then expect to find different modules coming out of pre and post-instruction data, especially if the instruction was effective.

Brewe et al. [2016] found a modular structure in FCI data using post-instruction results. Why post-instruction? Bao and Redish [2006]'s original work on model analysis grants some insight. They showed that at the beginning of the semester, students began mostly in pure states of incorrect models. A traditionally-taught course moved students mostly to mixed states over a semester, while a course with interactive engagement based problem solving tutorials led students mostly to pure states with the correct model. As people move over a learning progression, it is likely a fairly general pattern that they move into and through mixed states, where the clustering is less pronounced (we earlier in this chapter that high mixedness in model analysis simulations made it harder to detect their structure).

I combined data from different semesters to get a larger dataset because tests on simulations showed that a fairly large dataset was necessary to observe existing structure in most cases. This resulted in 724 pre-instruction test results and 714 post-instruction test results.

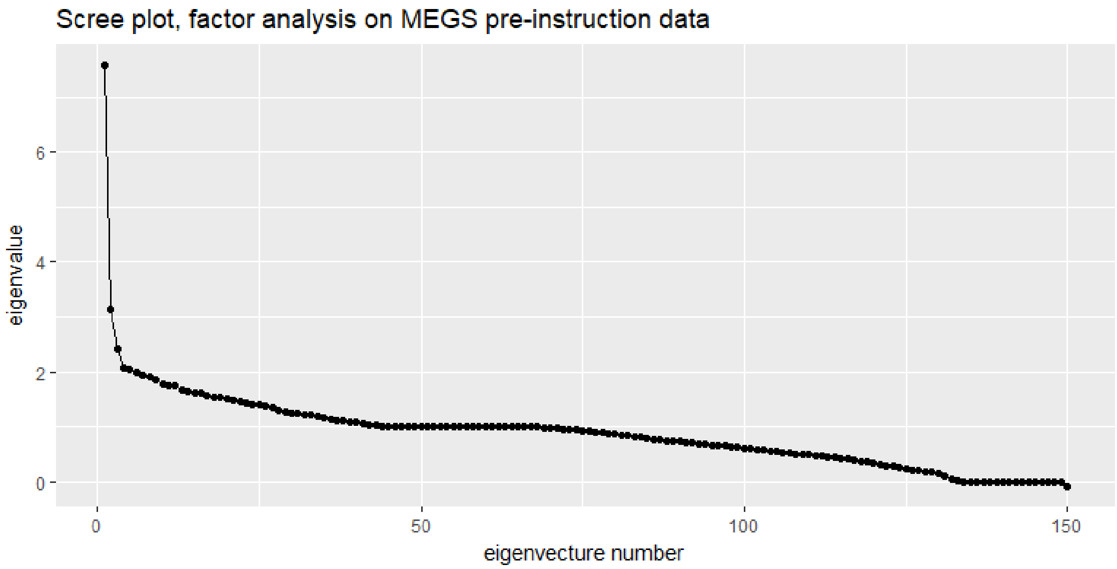


Figure 5.25: Eigenvalues from factor analysis on pre-instruction MEGS data.

5.8.1 Pre-instruction MEGS results

First I looked at the 724 tests taken by students who consented to participate in the research study over three years.

5.8.1.1 Factor Analysis

The scree plot for the eigenvalues of factor analysis run on the pre-instruction MEGS data are shown in figure 5.25.

Factor analysis identified one significant factor, which contained answer choices 5C, 10A, 10B, 13A, 13B, 15B, 17C, 24A, 26D, 27B, and 27C.

Of these, 5c, 10A, 13A, 15B, 17C, 24A, 26D, and 27B are correct. 10B and 13B are not. This is probably a “correct answers” factor.

While the finding is clearly not a statistical fluke, it is also not very informative.

I'll discuss some ways that the results of factor analysis might be improved, including excluding correct answers before performing the factor analysis, in section 5.10.1.

5.8.1.2 MAMCR

Because my implementation of MAMCR was almost wholly unsuccessful at finding known structure, I won't execute the entire MAMCR procedure on the MEGS. However, it would be interesting to visualize the MEGS network. I performed the project and network backbone extraction steps on the MEGS data, and the resulting network is shown in figure 5.26.

Unlike in examples such as figure 5.22, there isn't a strong, obvious structure here, except a cluster of nodes as the center, which inspection reveals to consist primary of correct answers.

I ran the process again, removing the correct answers from the network first. The result is figure 5.27.

I still didn't see any obvious structure, although it would be interesting to try to investigate this network more thoroughly in the future.

5.8.1.3 Modularity maximization

Next, I ran modularity maximization on the MEGS pre-instruction results. It returned 5 clusters. The first cluster had 81 answer choices, so I will omit it here. I'll describe the remaining clusters in the next few paragraphs.

The next-longest cluster consisted of every correct answer except to question

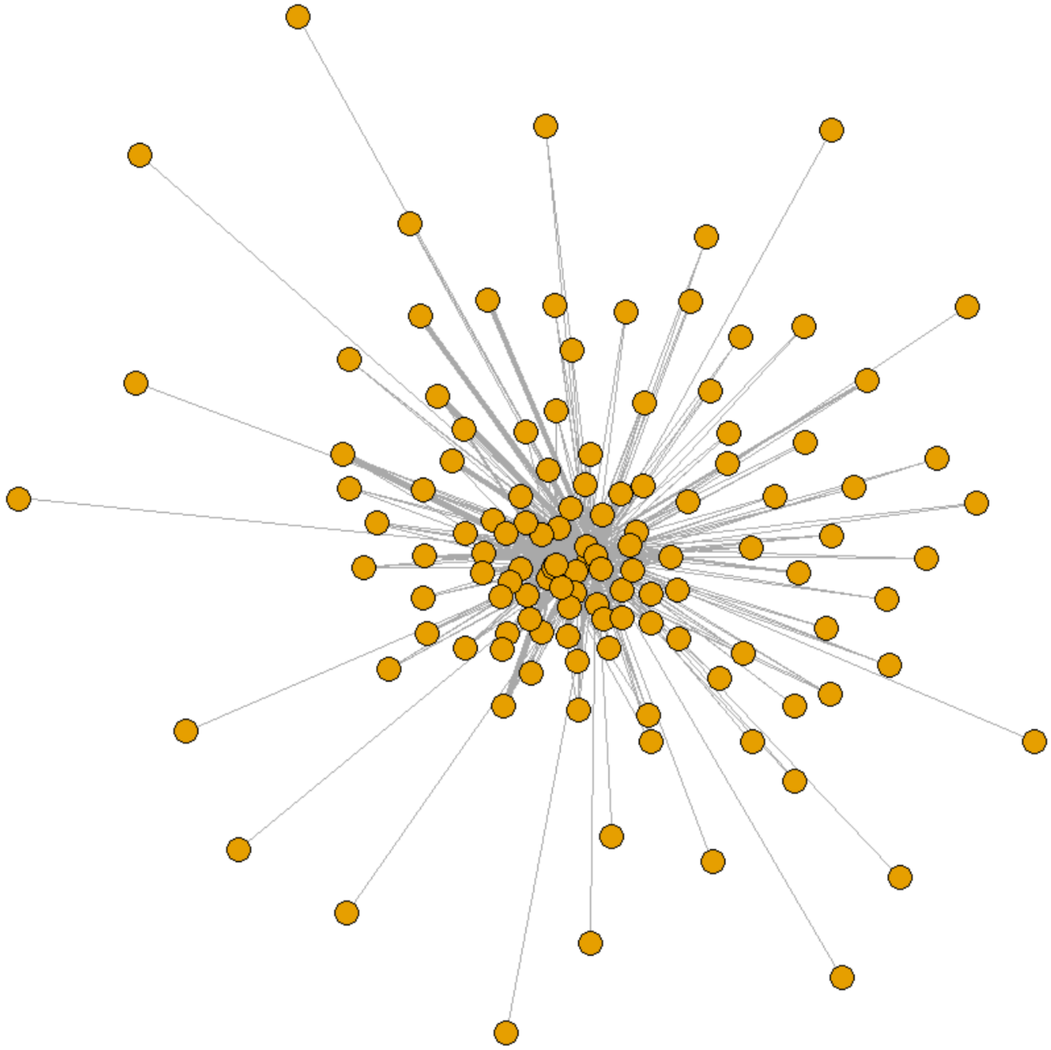


Figure 5.26: Backbone of the network of MEGS pre-instruction results projected onto answer choices. Edge thickness is adjusted to weight.

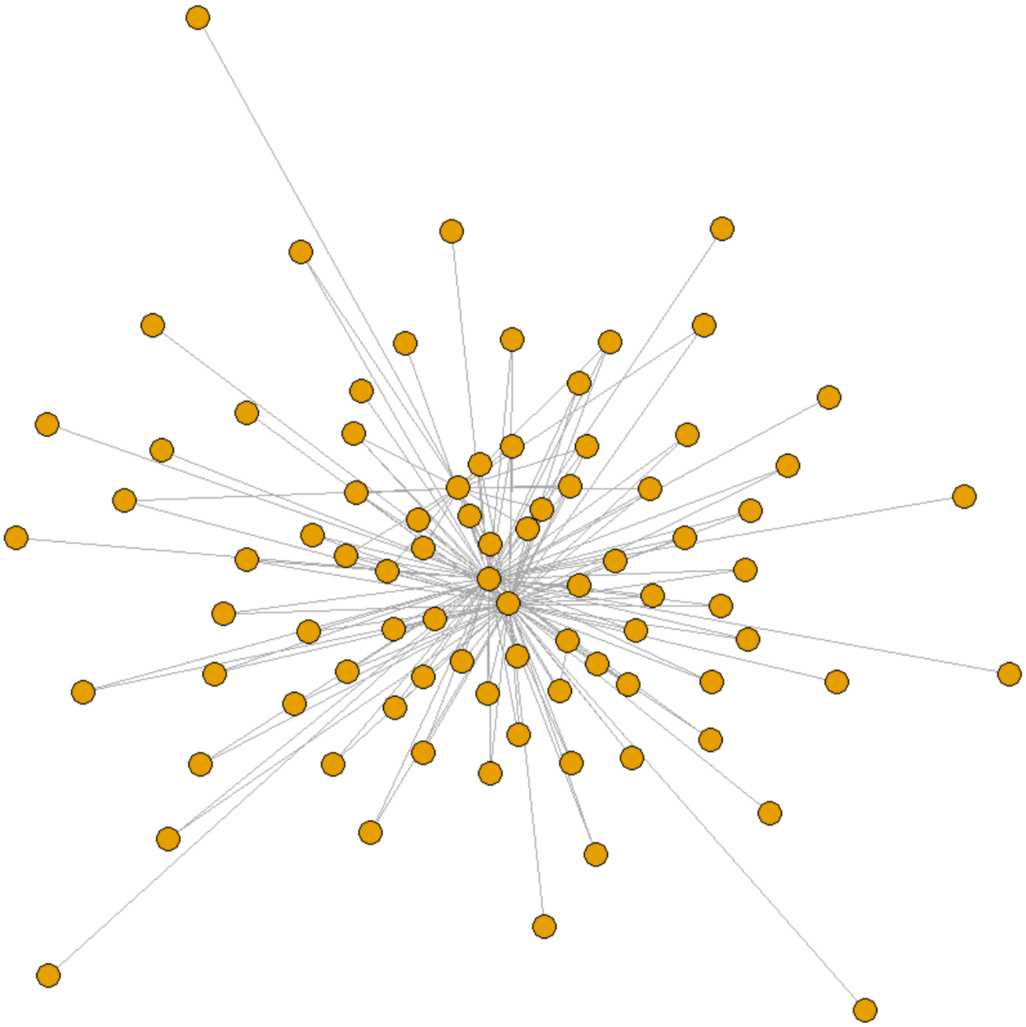


Figure 5.27: Backbone of the network of MEGS pre-instruction results projected onto answer choices. Edge thickness is adjusted to weight.

30. It additionally included 7A, 9E, 12E, 13C, 16B, 28E, and 30C.

The incorrect 30C was chosen 252 times in the data set, as opposed to 279 for the correct 30A. I interpret 30C's presence and 30A's absence as the cohort collectively choosing the wrong answer to question 30, something rare on the MEGS. 16B is similar; it is a massively-powerful distractor, chosen 528 times as compared to the correct 16D being chosen only 48 times. 7A was also a strong distractor, chosen 151 times.

9E, 12E, and 28E were in fact never chosen, so we can ignore their presence.

13B was chosen rarely, 28 times. Question 13 is about interpreting several equation qualitatively. It's unclear why this particular choice to this particular question is included. Nonetheless, this cluster represents the correct answers, plus several strong distractors.

The next longest cluster included mostly strong distractors. They were

Each distractor on the MEGS was chosen an average of 90 times, but these distractors were chosen an average of 177 times, despite several strong distractors appearing in the "correct answer" cluster. So this cluster might appear to be a "strong distractors" cluster. I wasn't able to find any other unifying factor behind this cluster.

It's important to remember that modularity maximization did not include projection, so the algorithm is trying to cluster not only answer choices, but students as well. With this in mind, the first cluster I discussed might be regarded as a "high-scoring" cluster and the second as a "low scoring" cluster.

The next cluster detected by modularity maximization was 1A (unit conver-

Table 5.2: Answer choices included in "strong distractors" MEGS cluster.

answer choice	number of times chosen
1E	18
4C	153
5D	21
6D	249
7E	59
9B	173
12B	186
14A	203
17E	85
20E	475
23E	215
28A	178
29C	187
30A	279

sion), 2E (scaling), 4E (dimensional analysis), 8C (unit conversion), 19B (estimation), 23B (turning words into an equation), 26E (estimation), 29E (scaling), and 30D (functional forms). The questions are spread out among various epistemic games, and I wasn't able to find any other connection between them. This cluster may be a statistical artifact.

The last cluster detected by modularity maximization included 22D, 27A, and 28D. Questions 22 and 27 are both about turning a verbal statement into an equation. However, question 23 fits in with this theme and is not included. Question 28 is about simplifying an expression in a limit, similar to question 12 and question 5. Although questions 22 and 27 seem related, I can't interpret this cluster as a conceptual cluster because modularity maximization often finds spurious connections, and conceptually-related questions aren't included here.

5.8.2 Post-instruction MEGS results

Next I looked at the post-instruction data, hoping to see whether there were any significant difference from the pre-instruction findings.

5.8.2.1 Factor Analysis

Factor analysis on the post-instruction MEGS data generated the scree plot shown in figure 5.28.

Factor analysis returned one significant factor. Its answer choices were 1A, 1B, 2B, 2C, 3A, 5C, 8E, 11A, 13A, 13C, 15B, 17C, 21A, 24A, 25A, 25D.

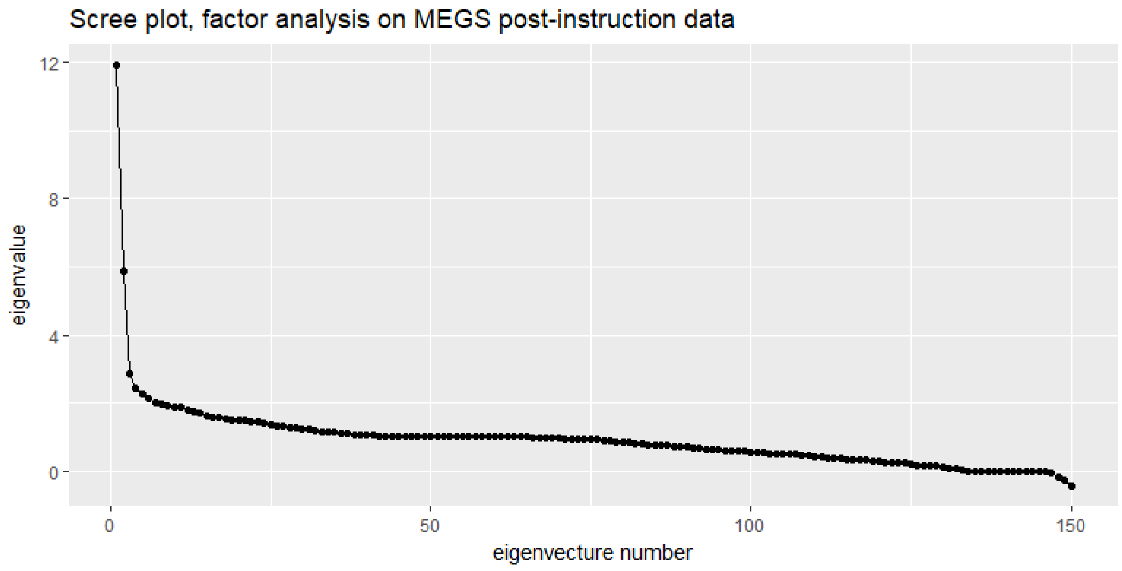


Figure 5.28: Eigenvalues from factor analysis on pre-instruction MEGS data.

Of these, 1B, 2C, 3A, 5C, 8E, 13A, 15B, 17C, 21A, 24A, and 25D are correct, so 11 of 16 answers in this factor are correct, and 4 more are answers to questions where the correct answer is also in the eigenvector. There is only one question represented (11) that has answer choices in the eigenvector, none of which are correct. I think it's safe to conclude this is the "correct answers" eigenvector. The question included are also, for the most part, the questions with a high percentage of correct answers. The average score on this subset of questions was 70%, while the average score on the entire MEGS was only 49%. Only two of the questions (8 and 11) were questions that were harder than average. The eight questions with the highest average scores are all included in this eigenvector, which is probably what allows it to have such a large eigenvalue (about 12).

Question 11 is "Approximately how many breaths does an average person take in their lifetime". The correct answer is one billion, and answer A, which appears in the correct answer eigenvector, is one thousand. However, this answer was almost never chosen; this probably creates high variance in how it correlates. It was chosen twice in the entire dataset. If those two students happened to get high scores generally but, for example, accidentally bubble the wrong answer on question 11, 11A's presence in the eigenvector is not meaningful.

It's also surprising that multiple answers to the same question are included in this eigenvector, since these answers cannot be associated with each other (although I set their correlations to 0, instead of their actual -1, to perform the factor analysis). Unlike answer 11A, these incorrect responses (1A, 2B, 13C, 25A) were chosen, respectively, 65, 34, 28, and 0 times. So 25A is presumably a statistical artifact. The other responses' presence is unclear. They could potentially represent distractors especially good at attracting high-achieving students. However, I generated many warnings about over-interpreting the results of factor analysis while looking at simulations; it is better to leave their presence unexplained.

The questions included in this eigenvector cover estimation, dimensional analysis, extreme cases, and functional relationships. All the epistemic games we wrote the test to evaluate are covered, and none of the questions are very similar (e.g. questions 6 and 11 are both estimates of everyday quantities). Thus, it appears that "correct answers", more than anything, is the connection.

The questions vary in their length from very short (e.g. 8) to long (e.g. 14). In chapter 4, I showed that classes improve over a semester on short questions and get worse over a semester on long questions. This made it plausible that a clustering algorithm might separate out short and long questions because students

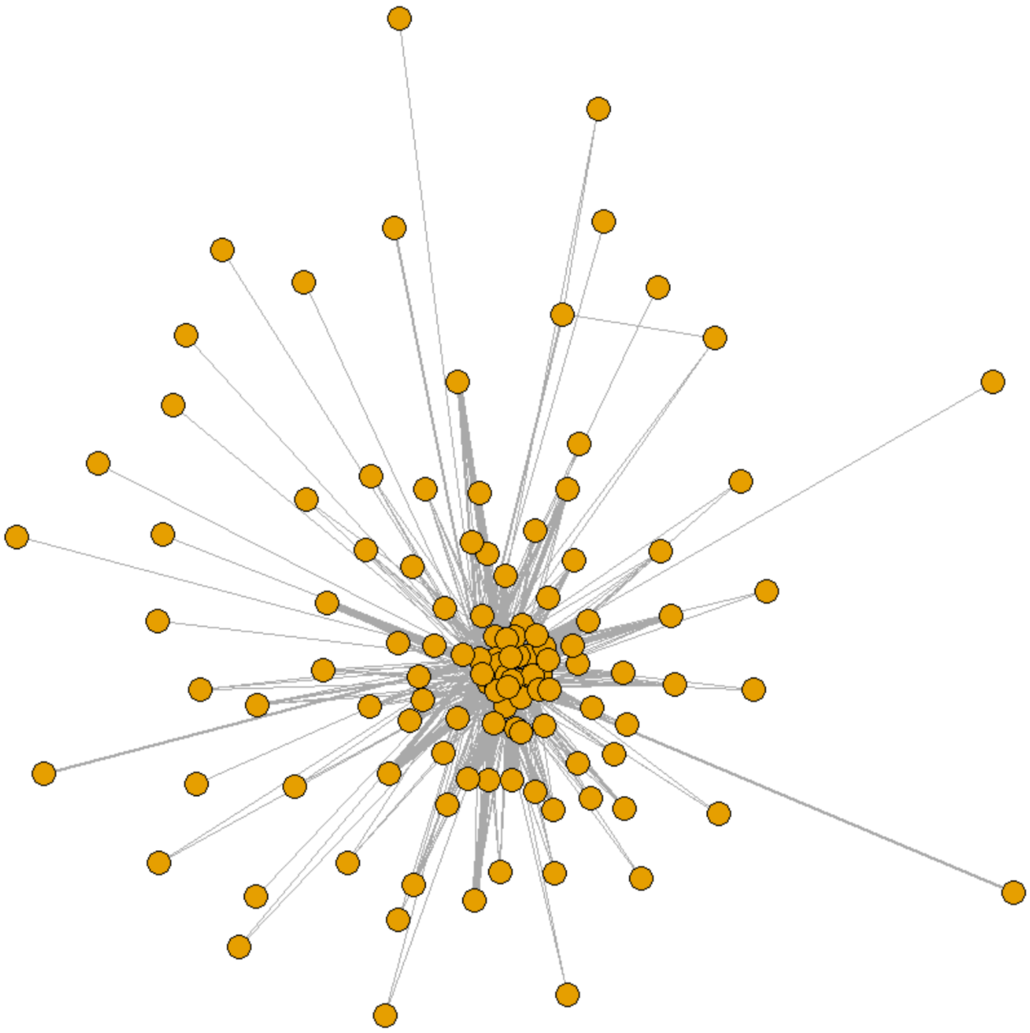


Figure 5.29: Backbone of network of MEGS post-instruction data projected onto questions.

not interested in reading long questions would be unlikely to answer them correctly, creating correlations, but this wasn't the case.

5.8.2.2 MAMCR

The network for the MEGS, analogous to figure 5.26 but for the post-instruction data, is in figure 5.29.

I didn't see any special structure aside from the correct answers clumped in the middle, and removing them was unenlightening.

5.8.2.3 Modularity maximization

Modularity maximization on the post-instruction MEGS discovered three modules. One was a giant module containing 97 answer choices.

As expected, there was a "correct answer choice" module, very similar to that discovered in the pre-instruction data. It contained exactly the same answer choices, except that it dropped the distractors 7A and 16B (the other distractors that were in the pre-instruction correct answer choice module remained).

This indicates to me that students improved at the MEGS over the semester. Although both 7A and 16B remained popular distractors (198 and 529 times chosen, respectively), they were no longer associated with the correct answers by modularity maximization. Excluding 16B is a remarkable result. The correct answer, 16D, was chosen only 79 times by 714 students. Still, modularity maximization picked it out as belonging to the "correct answer choice" cluster while ignoring the popular distractor. It nothing else, this shows that modularity maximization could be used to guess the answers to a test given only the responses a class of students made, and would probably outperform the method of choosing the most popular answer every time.

Question 16 is, "Bob and Fred have the same body proportions and body density, but Bob is 5'0" tall and Fred is 6'0" tall. Bob weighs 100 pounds. How much does Fred weigh?"

This is a question about scaling. The very strong distractor 16B is 120 pounds, which students arrive at by setting up a proportion that the ratio of Bob and Fred's heights is equal to the ratio of their weights. My interpretation of this distractor dropping from the correct answer choice module is that this is also a "high scoring students" module. Those students with the highest scores were the ones most likely to learn to recognize this problem as about three-dimensional scaling over the course of a semester.

Question 7 is on a very similar topic. It reads:

> Individual, single-celled *Dictyostelium discoideum* amoeba sometimes combine to form a small slug, typically about 500 μm by 60 μm by 60 μm . Scientists estimated the number of amoeba per a slug, assuming that the radius of a typical amoeba was 5 μm , but they later found that the radius of a typical amoeba is actually 2.5 μm . How far off was the original estimate for the number of amoeba in a slug?

The distractor 7A is "it was too big by a factor of 8". This question is closely related to question 16, also about scaling. Both question require understanding that the linear dimensions given in the problem should cubed because they are typical lengths of three-dimensional objects. That distractors from questions 7 and 16 were the two distractors dropped from the correct answer choice cluster over the course of instruction suggests that UMD PHYS131 is making strong progress on teaching the concept of scaling.

This is borne out by the pre-post score improvement on those questions. The cohort improved from 14.0% to 23.8% on Question 7 and from 6.6% to 11.1% on question 16. So the result that the class improved on these questions is not new, but it validates that modularity maximization was finding real results, even at the level of two individual answer choices dropped from a cluster over the course of a semester.

5.9 Conclusions

Both factor analysis and modularity maximization returned meaningful results when applied to the MEGS data, in that they detected modules of answer choices which clearly belong together. However, they simply detected correct answers, or strong distractors, which doesn't grant any new insight into the MEGS. Although they occasionally generated some other modules, these modules had no clear interpretation, and running the analysis on simulated data showed that such modules could easily pop up without carrying any real meaning.

I would have to continue to refine all three analysis methods' implementations in order to draw strong results about the MEGS from their outputs.

5.10 Future directions

All three methods I've implemented here could likely be improved to give more insight into the MEGS and how students think about it.

5.10.1 Factor analysis on distractors only

One step of MAMCR (which Brewe et al. [2016] stressed was optional depending on the cohort the data came from) was to throw out the correct answers. This might be a valuable step to take in factor analysis as well.

In factor analysis, the correct answers will correlate with each other strongly. As we saw in figure 5.1, this leads to a few very large eigenvectors in model analysis. The remaining incorrect models correspond to medium sized eigenvectors, and methods for determining the number of factors to keep were usually not successful in detecting these factors.

If I were to discard the correct answers and perform factor analysis on what remains, it might lead to a clearer view of the structure in the distractors.

Scott and Schumayer [2017] in pioneering the answer-choice form of factor analysis used here, found factors that include both distractors and correct answers. Such factors might be especially helpful in identifying false positive answers. So while it would be worthwhile to investigate this technique, it is also important to keep in mind that we lose some potential findings as well.

Another possible improvement to factor analysis as I'm using it would be to find a better algorithm for how many eigenvectors to keep, (see section 5.6.1.3) since none of the algorithms I tested performed well on model analysis, even though it was visually apparent that such an algorithm (e.g. the one my brain was using) should exist.

5.10.2 Improving modularity maximization

It may be that the modularity maximization algorithm isn't finding maxima to the extent that it could, so simply running it more times could potentially lead to improvements at the cost of computer time. The same isn't true for factor analysis, which, given a set of data, will obtain the same result every time it is run.

A more important way to try to improve modularity maximization is to change the definition of modularity to better suit a multiple-choice, single-response test.

Currently, my modularity maximization tests compared the the graph of a test to the sort of graph that would be generated by guessing randomly on a test were you could choose any number of responses to any question, including zero. Because on a real test, students are constrained to choose exactly one of every five answers, modularity maximization was comparing to the wrong standard.

Fixing this would require writing a new algorithm to deal with the new definition of modularity, but could potentially lead to improved results.

I know that modularity maximization, in its current implementation, has at least some validity. It correctly found the modules in the Rasch model, for example, given a realistic dataset of 1000 students. Improving its performance might allow researchers to identify clusters not only of answer choices, but of students as well.

As with factor analysis, one possible modification to modularity maximization would be to eliminate the correct answers from the graph, as in Brewe et al. [2016]'s implementation of MAMCR.

5.10.2.1 Improvements to my MAMCR implementation

As mention in section 5.7.4.3, MAMCR generally failed to find significant features in simulated data. Because MAMCR gave human-interpretable results in the analysis of Brewe et al. [2016], I suspect this points to a problem with my implementation. One option would be to change the network backbone extraction to match Brewe et al. [2016]'s method. (I didn't originally do this because I didn't find an R package with the LANS method built in.) Another is to revisit the InfoMap implementation I'm using. Some sources suggest igraph contains an old implementation of InfoMap that gives poor results [inf].

One other key difference between my implementation of MAMCR and Brewe et al. [2016]'s is that they discarded correct answers from their network whereas I did not. My expectation is that if this was the cause MAMCR's poor performance on my simulations, then MAMCR should return a single cluster for all the correct answers in every run, but this never happened.

Chapter 6: Conclusions

6.1 Metaphors be with you

In summer 2013, about a year before I moved to the University of Maryland to study PER, I was teaching a physics course to talented high schoolers at a summer camp.

I thought a particular homework assignment needed one more easy problem, so I asked for a free body diagram for a car driving down a road at constant velocity.

I was shocked by the results. Nearly every student had exactly the same answer: there was a downward force from gravity and an upward normal force from the road. Then there was a wind resistance force pushing the car backwards. This made it clear there must be a force pushing the car forward, to cancel the wind drag. On homework after homework, I saw the students had drawn a forward-pointing arrow labeled "the force of the engine". Not a single student had noted that a friction force from the road pushes the car forward.

I wasn't surprised by incorrect answers; I knew that if your homework assignments came back with all correct answers, they weren't challenging enough to give students a chance to grow. But how had almost every single student in the class come to the same wrong answer? I Googled for free body diagrams of cars, and found they weren't alone. Even a teaching website published by the BBC showed friction with the road slowing the car down, and a big, yellow "driving force" pushing the car forward [BBC, 2014], as did many other sites.

It wouldn't be possible for so many sources to independently come to the same wrong conclusion unless there were some reason behind it, so I began to wonder what the reason was. "Force", I realized, might be defined as a push or a pull in physics class. Does the road push on the car? When I push on things, I get tired. I need to use energy. If the road is pushing the car, why do I even have to pay for gas?

In everyday language, we say, "I'm being forced to work an extra shift" or "don't speak so forcefully to your elders" or "you don't know the power of the dark side [of the Force]". Force is about agency. It is "the thing that makes stuff happen". The engine is what makes the car go. The force that pushes it forward is the force from the engine. Of course.

Before this, I had viewed student mistakes as fundamentally ineffable. They were like Sandpeople. You could startle them off with a nice takedown of some misconception, but those mistakes would be back, and in greater numbers.

Perhaps you could predict mistakes a bit by modeling in your head what things student "can do" and "can do not" (there is no try), but the idea that those mistakes come from a completely sensible place, and that we could study the processes behind how students thought, was beyond anything I'd really considered. But the force awakened.

I used that homework as a chance to alter my view on wrong answers. I took up the idea that they were, in fact, to be expected. I gathered up all the bits of knowledge I had about cognition into an essay on learning physics and solving problems, which I published on Quora [Eichenlaub, 2013].

Writing that essay, I cited more authors than I read. One of those authors was George Lakoff, and he would later start me down the path to exciting ideas about cognition with *Metaphors We Live By* [Lakoff and Johnson, 2008].

Now, after four years with the UMD PERG group, I look back on that essay to see how much my understanding has changed. I realize that force, the mysterious driving force, is still with me.

6.2 Phenomenal Primitives

The question that originally drove me to PER was something like this:

> People are animals of the savanna. We evolved in medium-sized groups of hunter-gatherers. Our great innovation was to walk upright. Why should such an animal, whose concerns are so mundane, ever be able to understand something like quantum mechanics, or the Pythagorean theorem, or the number 6×10^{23}? How could primitives do things so phenomenal?

The summer before beginning the Under/Over project, Deb Hemingway and I read DiSessa [1993]'s foundational and challenging work on phenomenological primitives. These are small, irreducible bits of generalizable reasoning that can apply to a wide variety of situations, like *closer is more*, the closer you get to something, the more of it there is. Hence, things look bigger when you get closer, noises get louder when you approach their source, and it's hotter in the summer because we're closer to the sun.

This was among the first views I had of the ideas that had been developed into the resource framework. It was a grand vision - an over-arching picture of how people thought. We had mental resources, small little tools for thinking, and we learned to activate them together in patterns that helped us tackle big problems. Epistemologies, ontologies, misconceptions, metacognition, and the rest could all be understood in terms of resources.

One of the important intellectual ancestors of the resource framework is Minsky [1991]'s "society of mind". A mind is complicated. It can do all sorts of complex things. Where does that complexity come from? It's the same phenomenal primitive question. Minsky's answer is a mind is made of many individual "agents", each of which can only perform a simple task. But the agents exist in a hierarchy. There are agents for very low level things, like "determine the angle of flexion in your elbow right now", but there are high-level executive agents doing things like deciding whether Fermi's Golden Rule applies here, which they do by calling an army of sub-agents, which call on sub-agents, etc. Minsky used this architecture to build early artificial intelligence systems.

All of which is to suggest we could build theories around resources. Lakoff's work, for instance, detailed how a process of metaphor involved taking resources we developed for one process and using them to understand another. To put it another way: the agents in our society of mind can be re-used for new purposes.

Energy levels of a physical system are "high" and "low", and we can "raise" them, and then we'll draw them physically higher on a graph. It's via our resources for reasoning about the concrete, everyday experiences related to "up" and "down" that we can understand something as abstract as energy. Standing upright made for phenomenal primitives after all.

And then we could build these ideas further. Energy isn't just about high/low. Dreyfus et al. [2015] points that that it's when we pour energy into the system that energy levels go up. We mix metaphors - sometimes energy is a substance (how else would you pour it?). Sometimes it's levels. Sometimes, it's both, in a complicated way that we can explicitly map out. These were exciting ideas on the path to understanding that driving question. I spent a lot of afternoons in Joe's office talking about them, even if I never read all the books he gave me on the subjects.

6.3 Mind of the Society

Then I was introduced to some even wilder ideas, like Hutchins [1995]'s distributed cognition. The canonical example is that no one knows how to operate a battleship (or a star cruiser). There's a captain, who knows the ship's capabilities, its mission, and at least his direct subordinates among the crew. There are engineers who know how different systems work mechanically, and gunners and helmsman who know how to operate them, and a cook who knows how to make the food. But viewed as a whole, making and executing decisions about the battleship isn't done by any one person. It's distributed among the crew.

For that matter, I don't know how to tie my shoes. I tried to make the correct motions in the air just now, and I couldn't figure out if I was doing it right. I can only tie my shoes when I have my shoes with me to tie. The knowledge of how to tie my shoes is distributed across me and the shoes themselves. Even the crew alone isn't enough to understand how to run a battleship. It's distributed across the crew and the battleship.

This theory is one I never dove into deeply, but saw from time to time in classes and conversations. It has clear implications for physics, where the cognition might be distributed across students and the representations they make, or between groups of students, or the apparatuses they interact with in labs, etc. But it also got me to start thinking about a broader context than simply cognition inside a single individual as what's important in physics education. Beyond the society of mind in an individual is a society of minds.

6.4 To see a grain of sand in a world

Speaking of student mistakes, I think was one in the name of William Blake. The "e" was two spots two far back. *Songs of Innocence and Experience* paints learning more about the world as dreadful and dreary, since you'll wake up to nothing but the corruption and soullessness of modernity. (Is it a surprise? If reading Blake were a positive experience, why would his most famous poem be a cathode?)

I suppose it would be possible to take this sort of view when learning about education as well. If there's anything drearier than an overly-romantic poet, it's an economist. Caplan [2018]'s recent book is titled *The Case Against Education:*

Why the Education System is a Waste of Time and Money. He believes that most of education produces no learning of value. In economic terms, getting a degree is "about 80%" just signaling, in a very expensive way, that you are the type of person who can put your nose to the grindstone, do what you're told, and get a college degree. There's no shortage of evidence on his side, if all you want to see is test results.

But the time I've spent in PER has been the opposite of Blake in two ways. The first is that, in doing research, I don't "see a world in a grain of sand". In fact, I've been progressively more narrowed, from problem solving to epistemologies, to a tenuous link from ontological metaphors to epistemic games.

The second is that, in all my other activities becoming a member of this community, the world has become much broader, and infinitely more fascinating. PER has shown me the importance of so many things beyond trying to map out how an individual is thinking. Problem solving is only one of many problems to solve.

Graduate students at UMD in recent years have studied the development of student identities as learners and physicists. They've pointed out the moments in classrooms that build a culture, and how this has implications for how everyone there learns, but also how their peers and instructors respond to them. They've looked at how faculty learn to become better teachers, and studied quantum mechanics, computer simulations, and biophysics.

What I left out of that essay four years ago wasn't the resource framework or the piles of theories that go around it. In PER, I've learned about epistemologies, ontologies, metacognition, framing, the language of physics, hidden curricula, and many others. I didn't have the same names for them as I do now, but the passages I quoted from Feyman on problem solving were extolling students to adopt a new epistemological framing. The passages I quoted from Steven Pinker were on ontologies, and I described the benefits of immersing yourself in physics culture to absorb its language, and the rest of the hidden curriculum. I didn't know cognitive theories' details, but I knew they were there.

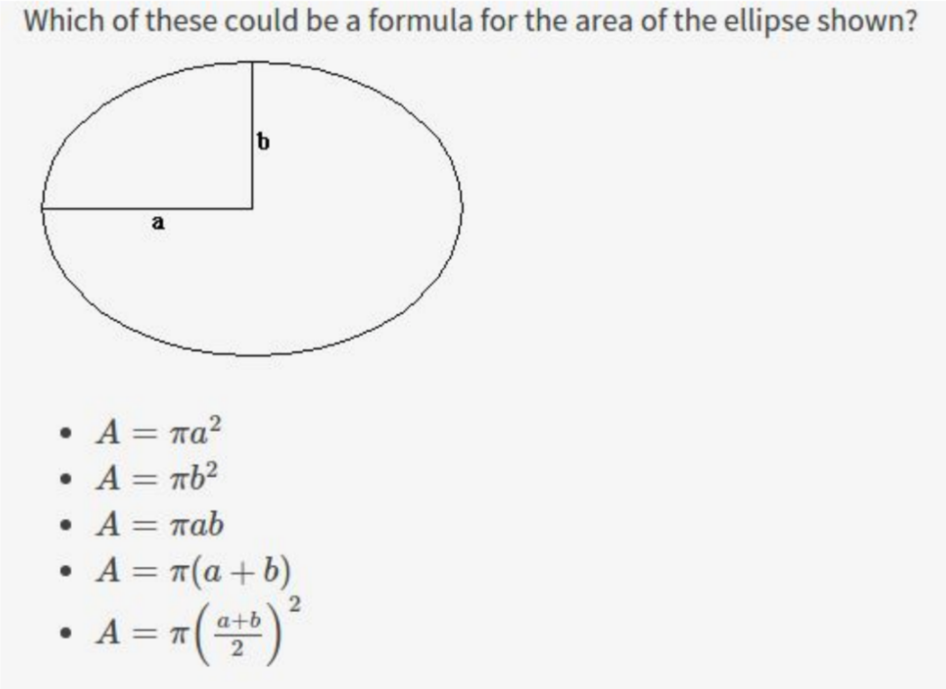
What working in the UMD PER group during this dissertation has given me is a new and far broader view of education. This thesis is not where I covered it, but while working on this thesis, that's what I learned. It's changed the way I teach, the way I talk about education, and what I value in all areas of life.

In PER, there's a concept of the "grain size" of your analysis. Looking at smaller and smaller moments in finer and finer detail is moving to a smaller grain size. No, I didn't look at a *small grain sand*. But move the 's' over one spot. But I looked at *small grains and* I glimpsed a whole world, still ripe to explore.

Appendix A: Interview Protocols

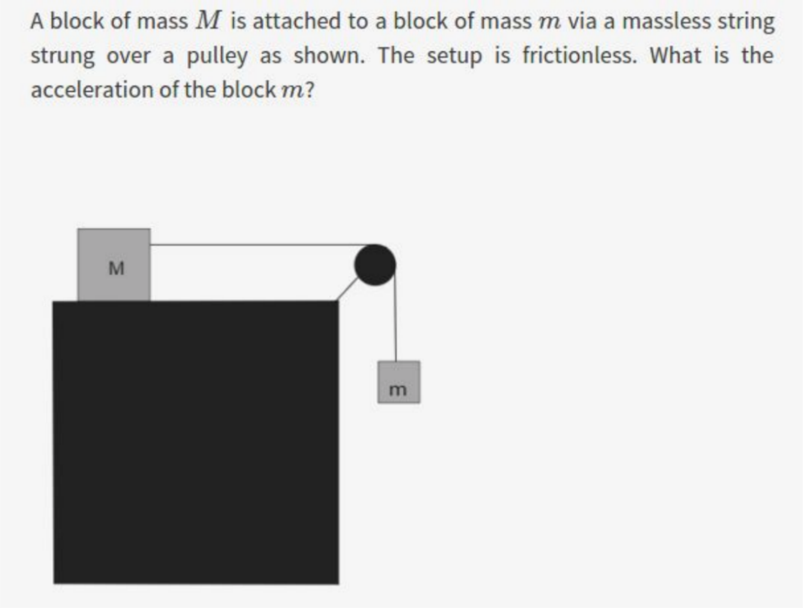
This appendix presents the problems given to students in problem-solving interviews.

A.1 Ellipse and half-Atwood machine



Which of these could be a formula for the area of the ellipse shown?

- $A = \pi a^2$
- $A = \pi b^2$
- $A = \pi ab$
- $A = \pi(a + b)$
- $A = \pi\left(\frac{a+b}{2}\right)^2$



A block of mass M is attached to a block of mass m via a massless string strung over a pulley as shown. The setup is frictionless. What is the acceleration of the block m?

A.2 Ring of charge, springs, ellipse (group interview)

Group Interview, Physics 131 students

Charge

Suppose you take a charge Q and spread it out into a circle of radius R. You can think of this as a very large number of very small point charges spread evenly around the circle, and the total charge of all the little charges adds up to Q. The setup looks like this:

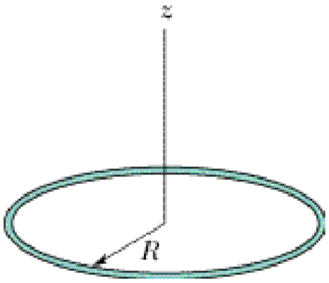


Figure 1:

z is the axis that points up out of the loop. (The x and y directions are in the plane of the loop.)

Plot

If a small charged particle were on the z-axis, it would feel a force from the ring of charge. The force would depend on how far up the z-axis the charge was placed. First, let's try drawing a plot of the force versus distance along the z-axis. What features can you find on the plot?

Calculation

Next, calculate the magnitude of the electric force on a charge q on the z-axis at a height h above the loop. If you get stuck, write down the best guess you have so far and move on.

Center of the loop

What is the magnitude of the electric force on q if it is directly in the center of the circle? How do you know?

Far away from the loop

Imagine that you are very far away from the loop, so h is much, much greater than R. Can you find an approximate answer for the magnitude of the electric field in this case?

How do the approximate answers you found in the last two steps fit in with the exact calculation? If you haven't finished the calculation, try again, keeping your approximate answers in mind.

Springs

Spring constants

With your group, make sure you understand what a spring constant is. What does it mean when the spring constant is very high or very low?

Series Springs

Imagine two springs connected end to end, then connected from a wall to a box, like this

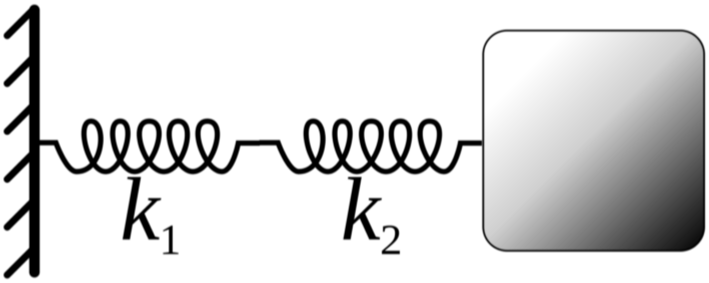


Figure 2:

This is equivalent to connecting a single spring to the box, as long as the spring has a new spring constant, k_{series}. Find a formula for k_{series} in terms of k_1 and k_2.

How can you test this formula to see if it makes sense?

Parallel Springs

Next imagine two springs connected both connected directly from a wall to a box, like this

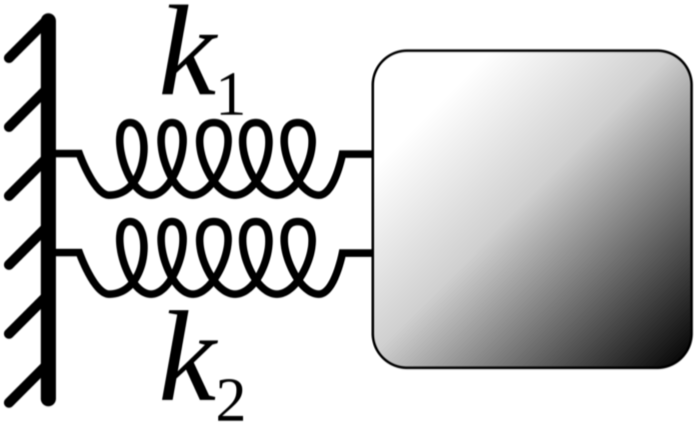


Figure 3:

Again this is equivalent to connecting a single spring with spring constant $k_{parallel}$. Find a formula for $k_{parallel}$. How can you tell whether the formula makes sense?

A.3 Terminal velocity (group interview)

A.3.1 Terminal velocity

When a sphere falls through a fluid, there are three forces on it:

- a gravitational force whose magnitude is mg
- a viscous force whose magnitude is $6\pi\mu rv$
- a drag force whose magnitude is $\frac{1}{2}c_D\rho Av^2$
- a buoyant force whose magnitude is ρgV

The symbols are:

- m: mass of the sphere
- g: gravitational acceleration
- A: cross-sectional area of the sphere
- μ: viscosity of the fluid
- r: radius of the sphere
- v: velocity of the sphere through the fluid
- ρ: density of the fluid
- V: volume of the sphere
- c_D: drag coefficient of the sphere

The viscosity of air is about $2 * 10^{-5} kg \cdot m^{-1} \cdot s^{-1}$ and the viscosity of water is about $10^{-3} kg \cdot m^{-1} \cdot s^{-1}$.

The density of air is about $1 kg \cdot m^{-3}$ and the density of water is about a thousand times greater.

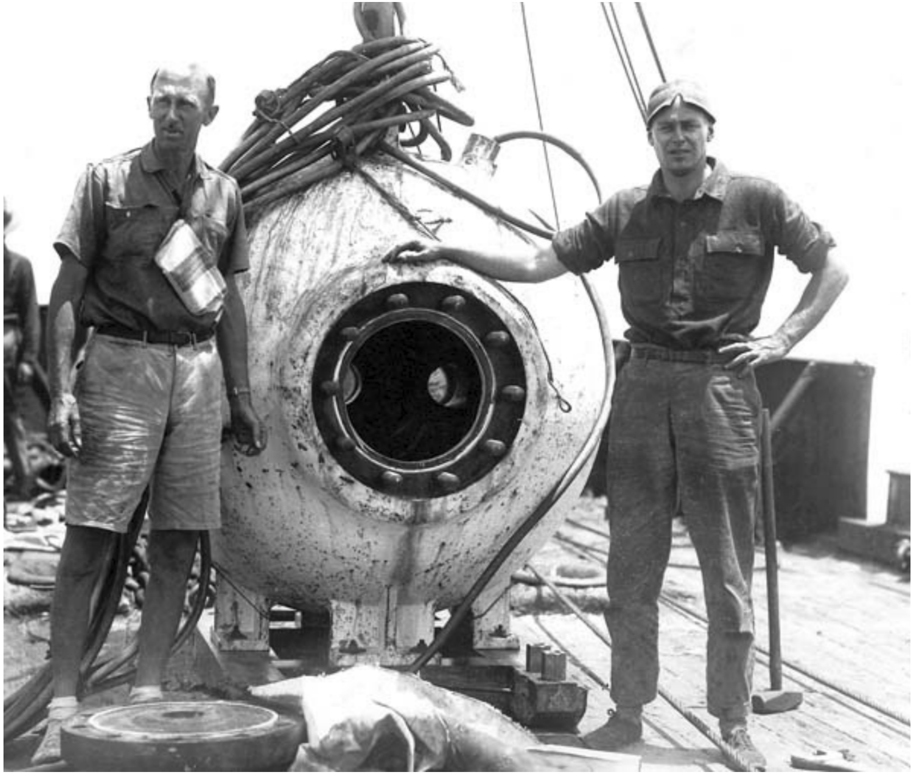
The drag coefficient of a sphere varies somewhat, but is usually around one half.

A.3.1.1 Question 1

What is the terminal velocity of the sphere in terms of the other variables?

A.3.1.2 Question 2

The Bathysphere was an approximately sphere-shaped, deep sea submersible, pictured here:



The Bathysphere's mass (with people inside) was approximately 2000 kg. It was lowered into the ocean, where it fell under gravity to the bottom (and was later pulled up by a cable). Estimate how long it took the Bathysphere to descend.

A.3.1.3 Question 3

After the 1883 eruption of Krakatoa (a volcano in Indonesia), people reported especially vivid sunsets around the world for roughly three years. About how small would the volcanic ash from Krakatoa have to have been to affect sunsets for that long?

A.3.1.4 Question 4

When drag can be ignored (because viscosity is much larger than drag), the equation for the velocity of a falling sphere over time is

$$v(t) = v_t(1 - e^{-t/\tau})$$

v_t is the terminal velocity.

Which curve has the highest value of τ? Which has the lowest?

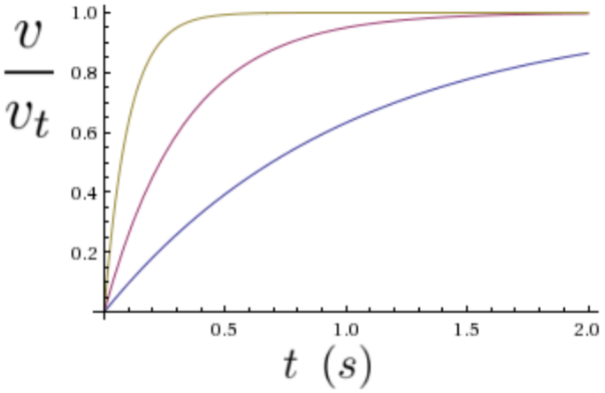


Figure A.1: terminal velocity plot

Estimate τ for each of the three velocity curves shown below.

How is τ related to g?

Appendix B: List of Epistemic Games

This appendix supplements the list of epistemic games in Tuminaro [2004] and follows the format there for describing and identifying e-games. For further discussion of these two e-games, please see chapter 2. Both of these games fall under Tuminaro's "quantitative sense making frame"

B.1 Extreme or Special Cases e-game

This e-game is a an example of the "sanity check" e-game described in chapter 2. It is also an example of the "mapping from mathematics to meaning" e-game. However, I analyze this e-game separately because examining extreme cases is a specific instructional target, and has been shown to have specific effects on how students think about physical scenarios, for example promoting vivid visualization [Clement and Stephens, 2009].

B.1.1 Description

Students use particular values of variables or parameters to evaluate whether a given equation is likely to be a good description of a physical system.

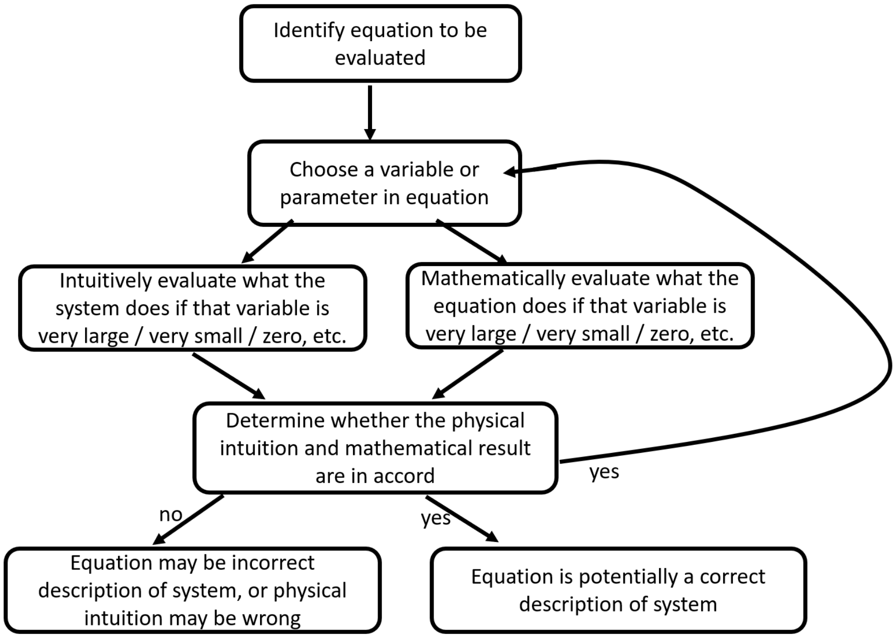

Identify equation to be evaluated
Choose a variable or parameter in equation
Intuitively evaluate what the system does if that variable is very large / very small / zero, etc.
Mathematically evaluate what the equation does if that variable is very large / very small / zero, etc.
Determine whether the physical intuition and mathematical result are in accord
yes
no
yes
Equation may be incorrect description of system, or physical intuition may be wrong
Equation is potentially a correct description of system

Figure B.1: Moves in the extreme/special cases epistemic game

B.1.2 Identification

Analysis involves formal mathematical expressions. Students reference particular symbols in the expression and mention specific values of the variables.

B.1.3 Moves

See figure B.1.

B.1.4 Knowledge Base

Mathematical resources, especially limits and the ”prop+” and ”prop-” symbolic forms. All resources relevant to physical intuition of systems.

B.1.5 Epistemic forms

The equation being evaluated, a list of variables and parameters in that equation. Students may sometimes draw up and down arrows next to symbols as they play this game.

B.1.6 Notes

In a second version of this game, students begin by examining the physical extreme cases, and use this to determine what mathematical equations for the scenario should look like. This is a specific example of Tuminaro's "mapping from meaning to mathematics" game, but I have coded both extreme/special cases games as the same e-game in this thesis because my expectation is that the epistemic frame and resources associated with playing these two e-games are similar.

B.2 Dimensional analysis

B.2.1 Description

Using the dimensions of the symbols in an equation to construct one test of whether that equation is correct.

B.2.2 Identification

Student reference length, mass, and time, or meters, kilograms, and seconds, usually putting them together these into more complicated combinations. They may also

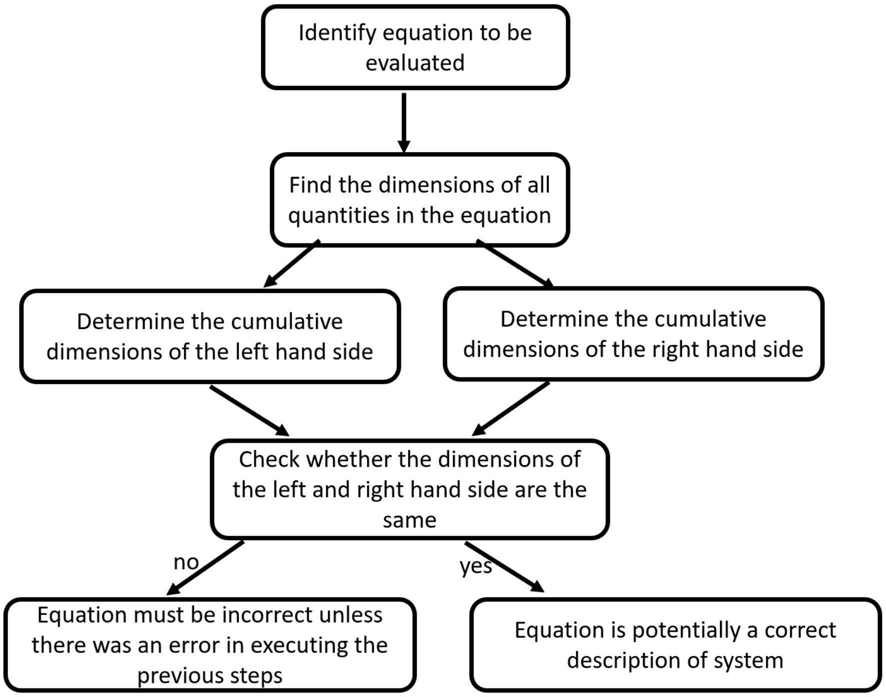

Identify equation to be evaluated
Find the dimensions of all quantities in the equation
Determine the cumulative dimensions of the left hand side
Determine the cumulative dimensions of the right hand side
Check whether the dimensions of the left and right hand side are the same
no
yes
Equation must be incorrect unless there was an error in executing the previous steps
Equation is potentially a correct description of system

Figure B.2: Moves in the dimensional analysis epistemic game

reference canonical examples of dimensions, e.g. "these would both be areas".

B.2.3 Moves

See figure B.2.

B.2.4 Knowledge Base

The dimensions associated with various physical quantities, rules for combining dimensions.

B.2.5 Epistemic form

The equation being evaluated. In some cases, a new equation that is built containing only M, L, and T (or kg, m, and s) in place of the variables in the original equation.

B.2.6 Notes

There are many uses of dimensional analysis, but the e-game described here applied without significant modification to the uses we coded as dimensional analysis in our data set.

Appendix C: Codebook for ontological metaphors for equations

This appendix describes each ontological metaphor code I used to code references to equations, see chapter 3 for details.

For each code, I give a description and three examples, each with brief commentary explaining why I coded the example with that ontology, and sometimes suggesting a connection to epistemic framing.

C.1 Equation as whole

C.1.1 Description

refers to or indicates an entire equation without distinguishing different parts of the equation

C.1.2 Example 1

> this one makes sense to me [points to $A = \pi \left(\frac{a+b}{2}\right)^2$]

“this one” refers to an entire equation at once

C.1.3 Example 2

> I also know it wouldn't be three [points at third answer choice on multiple choice quetion]

'"three" is a stand in for an entire formula

C.1.4 Example 3

> okay, so like if the formula was A equals pi times a plus b squared

although individual symbols are read out, they aren't reasoned about separately, and instead are all "the formula".

C.2 Equation as parts

C.2.1 Description

refers to or indicates a term, variable, or group of variables indepedently from the rest of the equation

C.2.2 Example 1

> there's no four, one fourth, in the original equation, and it's not taking into account the a

student is referencing just the number 4 in the equation $\frac{(a+b)^2}{4}$ while using only that number to draw inferences

C.2.3 Example 2

> I would have to manipulate part of this overall equation [points to a specific term in an equation, then makes a circling motion around entire equation] for it to come out to the answer I got

uses equation as parts and equation as whole one after the other. also an example of equation as mutable - agentive

C.2.4 Example 3

> if it was a plus b over two gives the average radius

discusses a single part of full equation $A = \pi \left(\frac{a+b}{2}\right)^2$. This is not grouping because there's no indication that the symbols referenced are a single entity, although a single conclusion has been drawn from all of them

C.3 Equation as mutable - agentive

C.3.1 Description

states or implies that equations can change (e.g. terms dropped or added, exponents adjusted, one equation becoming another related equation) due to the intervention of a specific person, group of people, or other agent. This does not include rewriting the same equation to solve for a different variable or otherwise finding an equivalent algebraic expression.

C.3.2 Example 1

> Let's just do m a equals ten m times... ten m minus m a. I'm just messing with it.

The agent is the student. "Let's just do" implies they are making a choice; many equations must be possible. "Messing with it" is actively changing the equation. (The student wasn't engaged in algebraic manipulation, but was instead creating equations not previously written.)

C.3.3 Example 2

> I'll just use deltas for now. Delta x over delta t. That's delta x over delta t, is zero.

The equation is mutable because the deltas exist "for now", implying that the equation may be different in the future. It is agentive because the speaker is making the decision about what form the equation takes.

C.3.4 Example 3

> we need to get rid of this [points to a in an equation] acceleration

"We" serves as an agent. The equation is mutable because it is possible to "get rid of" a symbol in it

C.4 Equation as mutable - non-agentive

C.4.1 Description

states or implies that equations can change (e.g. terms dropped or added, exponents adjusted, one equation becoming another related equation), but no specific person, group of people, or other agent is implied. This does not include rewriting the same equation to solve for a different variable or otherwise finding an equivalent algebraic expression.

C.4.2 Example 1

> the mass times acceleration becomes the overall force

an equation changes from ma to f_{net}, but it simply "becomes", no agent is identified

C.4.3 Example 2

> it would approach the acceleration of ten meters pers second squared

this example is also "equation not differentiated from phenomena". In context, "it" may mean the equation or the half-Atwood machine. There is change because it would "approach" a new quantity. No specific agent is identified as enacting this approach

C.4.4 Example 3

> I think in that case then it would just be k-one over k-two divided by two, but I'm not a hundred percent sure.

The equation depends on the "case", implying that under different cases, the equation changes to something different.

C.5 Equation as immutable

C.5.1 Description

explicitly states, or it can be reasonably inferred, that it is not possible for the equation to change (in the sense described in "equation as mutable" tags)

C.5.2 Example 1

> ...cause the force of the weight is equal to mass times gravity. And this is Newton's first law. It's just a law f equals m a. And let's just replace f with f w because that's the weight force.

The first half is immutable. The weight simply is something, and the reason is a law. As "just a law" it is epistemologically an unquestionable axiom, but ontologically it is incapable of change. The second half serves as a contrast. "Let's replace. . . " is mutable, because the terms are being changed, and agentive, with the agent being "us".

C.5.3 Example 2

> but I feel like finding the area of an ellipse should have the same formula no matter what the dimensions are

the formula is always the same, so it cannot be changed (e.g. to accomodate a special circumstance)

C.5.4 Example 3

> I'm putting that down. Our final answer.

Because the equation referenced is "final", it is no longer allowed to change.

C.6 Grouping

C.6.1 Description

refers to several terms or variables in an equation as a single entity, either with verbal cues, gesture, or writing. grouping is often an example of "equation as parts", but requires some extra evidence that the symbols being grouped are ontologically a single entity, rather than referencing several individual parts at once. Equation as whole may fit this definition of group, but I have set it off to a separate category.

C.6.2 Example 1

> They're saying this whole thing [draw parentheses around $\frac{M}{m}g$ in the equation $m\frac{M}{m}g$] is equal to a.

separates out three symbols from four and refers to them as a single entity, a “thing”

C.6.3 Example 2

> a equals v over t and g equals meters over second squared. those [draws a vertical line from $a = \frac{v}{t}$ to $g\frac{m}{s^2}$ and retraces it back] equal each other. v over t and meters over second squared. same dimensions.

When “those” equal each other, there are two groupings going on. One is the right hand side of the equation $a = \frac{v}{t}$. The second is symbols $\frac{m}{s^2}$ used to show dimensions, but considered part of a larger equation.

C.6.4 Example 3

> if the mass is m plus M

two symbols in the denominator of an equation, m and M, are grouped into a single conceptual entity, the total mass of the system

C.7 Equation / phenomena not differentiated

C.7.1 Description

it is ambiguous whether a verbal utterance or gesture indicates the mathematical equation in question or the physical or geometric object, system, or phenomena it describes

C.7.2 Example 1

> you have all the same equations. the vectors are the same and our free body, everything was the same. So if we just change a big block versus a smaller block. Same idea. These are all equal to each other [points to three diagrams of the half-Atwood machine].

when “changing the big block for the smaller block” we might be changing them in the equation mentioned at the start of this quote, or in the diagrams pointed to at the end

C.7.3 Example 2

> f net has to equal the weight of gravity

“f net” references a symbol, f_{net}, used extensively in class, and given that it must be “equal” to something it may be thought of as part of an equation here. But the other side of the equation is described as “the weight of gravity”, a physical phenomenon, instead of, for example, mg, an algebraic expression, so “f net” may refer to the physical net force

C.7.4 Example 3

> it would just be saying the area would be a circle with double radius a which wouldn’t make sense because like, then the radius would just be like a much bigger object

“the radius” is an “object”, suggesting it is geometrical, but it is also “double a”, suggesting it refers to the symbol for an ellipse axis as well

C.8 Symbol as parameter

C.8.1 Description

reference to a symbol as having a specific value that is constant in a physical system across its time evolution or different initial conditions, but which may change by, e.g. switching out for a new apparatus of the same design but larger scale

C.8.2 Example 1

> so you have a really small m. okay so you just have, kind of ten m over M

the equation $a = \frac{10m}{M+m}$ is simplified to $a = \frac{10m}{M}$ not in general, but because the specific value for m allows it under the current consideration

C.8.3 Example 2

> What’s the acceleration? Would be negative nine point eight meters per second squared. That’s gravity.

“gravity” takes a specific set value and doesn’t change and doesn’t have to be found. It is a reference point for the acceleration (a symbol treated as variable)

C.8.4 Example 3

> if this mass is super big it's gonna be moving in the up direction which wouldn't make sense

the general behavior of the system depends on the value of the mass, which is fixed when we consider the motion, and changes between different setups of the system

C.9 Symbol as variable

C.9.1 Description

reference to a symbol as having a specific value at a given time or space, but that value changes as the system evolves with time, or holds different values at different places

C.9.2 Example 1

> I don't know if acceleration is constant at all yet

indicates that a certain symbol, a, referenced as "acceleration", may or may not change over time

C.9.3 Example 2

> since we know that they have to be accelerating at the same time, then we just solve for a from this equation

the symbol a is identified with physical acceleration, which is a function of time

C.9.4 Example 3

> When it's far away, the h would be really big, so this denominator would be increasing in size.

h is acknowledged to be a variable that changes based on physical position considered. (The next part of the utterance exemplifies grouping.)

C.10 Equation as one form of a relationship

C.10.1 Description

refers to a given equation as only one way of expressing the functional relationship between variables, implying that solving for a different variable or otherwise rewriting the same relationship between variables in a different way is a potential mathematical move

C.10.2 Example 1

> then pi a plus b over two squared would be pi a squared plus a b plus b squared over four, which pi is that... I don't think that can equal pi a b or pi r squared.

the student expanded out the form $\pi \left(\frac{a+b}{2}\right)^2$ algebraically, using only the expanded form to make a conclusion comparing to other formulas. The conclusion still referenced the original form, so the equation was seen as one form of a more general relationship to reason about

C.10.3 Example 2

> so I could do algebra. Solving for a

recognizes that using algebra to transform the equation into a new form will give insight into the physical acceleration

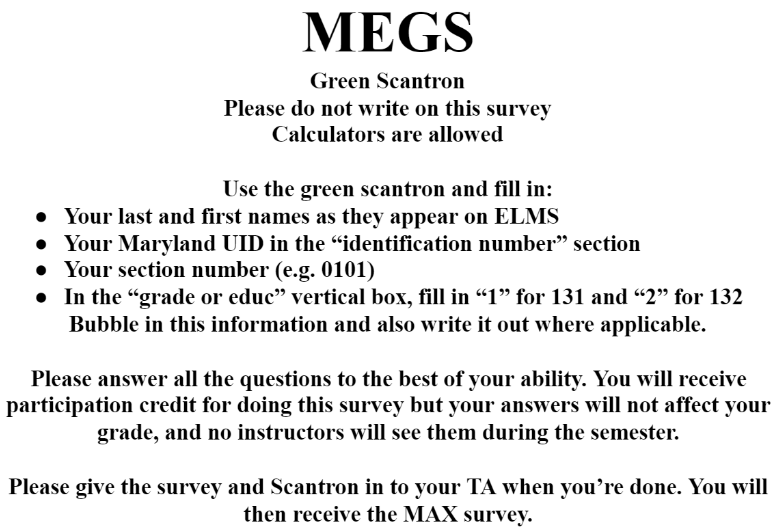
C.10.4 Example 3

> I'm trying to decide on what the form of the equation we need is.

Implicitly states that equations have different forms with varying usefulness, since there is one form "we need"

Appendix D: Math Epistemic Games Survey

Below are images showing the MEGS survey, v1.1.



MEGS

Green Scantron
Please do not write on this survey
Calculators are allowed

Use the green scantron and fill in:

- **Your last and first names as they appear on ELMS**
- **Your Maryland UID in the "identification number" section**
- **Your section number (e.g. 0101)**
- **In the "grade or educ" vertical box, fill in "1" for 131 and "2" for 132**

Bubble in this information and also write it out where applicable.

Please answer all the questions to the best of your ability. You will receive participation credit for doing this survey but your answers will not affect your grade, and no instructors will see them during the semester.

Please give the survey and Scantron in to your TA when you're done. You will then receive the MAX survey.

Student Performance Using Math:

The Math Epistemic Games Survey (MEGS) V.1.1

1. There are many old English units for volume, only a few of which are familiar today. For example:

   1 dram = 60 minim
   1 teaspoon = 80 minim
   1 pony = 6 dram
   1 tablespoon = 3 teaspoon

   How many drams were in a tablespoon?
   a. 2.25
   b. 4
   c. 6
   d. 15
   e. 14,400

   For questions 2 and 3: When you inject a small blob of a dye in a fluid, it will start spreading out due to the random jiggling at the molecular scale. If *x* is the radius of the blob and *t* is time, the radius is approximated by the equation:

   $$x^2 \approx 6Dt$$

   where D depends on the particular type of dye and fluid.

2. After 1 minute you observe the once-small drop has now expanded into a blob with a radius of 0.10 mm. What will the radius of the blob be after 100 minutes?
   a. 0.01 mm
   b. 0.10 mm
   c. 1.0 mm
   d. 10 mm
   e. 100 mm

3. How does the the blob's expansion rate change as you increase *D*?
   a. It increases
   b. It decreases
   c. It stays the same
   d. None – the expansion depends on *t*, not *D*

4. Which expression could represent the surface area of a solid object? Variables A, B, and C represent lengths, such as the length of the side of an object or the diameter of a circular object.
    a. $2(AB + A\sqrt{C^2 - A^2} + BC)$
    b. $\sqrt{A^2 + B^2 + C^2}$
    c. $\frac{\sqrt{2}}{2}A^2B$
    d. $\frac{3AC}{2B}$
    e. None of these could be a surface area

5. The endoplasmic reticulum has a highly-convoluted structure in order to maximize its surface area. Phillips, Kondev, Theriot, and Garcia estimated its surface area with the equation

$$A = \frac{8\pi}{3d}({R_{ER}}^3 - {R_{nuc}}^3)$$

where d is the distance between folds of the endoplasmic reticulum, R_{ER} is the radius of the endoplasmic reticulum, and R_{nuc} is the radius of the cell nucleus (which the endoplasmic reticulum surrounds). When would $A = \frac{8\pi}{3d}({R_{ER}}^3)$ be a close approximation to this estimate?
    a. When the endoplasmic reticulum is about as large as the nucleus
    b. When the endoplasmic reticulum is more than 1 micron
    c. When the endoplasmic reticulum is much larger than the nucleus
    d. When the endoplasmic reticulum is much smaller than the nucleus
    e. Never
    f. Always

6. Estimate the thickness of a page in a typical textbook.
    a. 10^1 m
    b. 10^{-2} m
    c. 10^{-4} m
    d. 10^{-6} m
    e. 10^{-8} m

7. Individual, single-celled *Dictyostelium discoideum* amoeba sometimes combine to form a small slug, typically about 500 μm by 60 μm by 60 μm . Scientists estimated the number of amoeba per a slug, assuming that the radius of a typical amoeba was 5 μm , but they later found that the radius of a typical amoeba is actually 2.5 μm . How far off was the original estimate for the number of amoeba in a slug?
    a. It was too big by a factor of 8
    b. It was too big by a factor of 4
    c. It was too big by a factor of 2
    d. It was too small by a factor of 2
    e. It was too small by a factor of 4
    f. It was too small by a factor of 8

8. 1 cubic centimeter of water has a mass of 1 gram. What is the mass of 1 cubic meter of water?
    a. 10^{-6} kg
    b. 10^{-3} kg
    c. 1.0 kg
    d. 100 kg
    e. 1000 kg
    f. 10,000 kg

9. If we redefined the length of an hour so there were 10 hours in a day, how would we need to change speed limit signs, assuming we don't want to change how fast we actually drive (and assuming we update all speedometers correctly)?
    a. The signs should be changed to higher numbers of miles per hour
    b. The signs should be changed to lower numbers of miles per hour
    c. The signs should stay the same
    d. There is not enough information to decide.

10. Cassabanana plants defensively extrude a waxy substance when attacked by aphids. The amount of waxy substance extruded for different numbers of aphids has been measured over a range from 5 to 20 aphids, and the rate of extrusion is known to be linear in that range.

    When 5 aphids attack, a particular plant extrudes 0.50 ± 0.05 grams/day. For 20 aphids, the rate is 2.1 ± 0.1 grams/day, and for 35 aphids, 2.5 ± 0.1 grams/day. Does the rate wax is extruded continue to follow a linear trend through 35 aphids?

    a. No, the amount extruded at 35 aphids is significantly less than a linear trend predicts
    b. Yes, a linear trend held within the measurement accuracy
    c. No, the amount extruded at 35 aphids is significantly more than a linear trend predicts
    d. There is not enough information to decide.

11. Approximately how many breaths does an average person take in their lifetime?
    a. one thousand
    b. one million
    c. one billion
    d. one trillion

12. The surface area of a cylinder of radius r and length l is $2\pi r(r+l)$. Which of these would be the best approximation to the surface area of a long, thin cylinder?
    a. $2\pi r^2$
    b. $2\pi l$
    c. $2\pi rl$
    d. $2\pi l^2$

Use for the following three questions:
Neuroscientists have found that an enzyme, denoted enzyme A, increases neural activity and another, denoted enzyme B, inhibits neural activity. The rates of production of A and B depend on the concentrations of various precursors, denoted P_1, P_2, and P_3. Specifically, the rate of production of enzyme A is proportional to $P_1 * P_2 / P_3$. The rate of production of B is proportional to $P_2 * P_3$.

13. If P_1 is increased, holding the other P_i constant, is the neural activity increased or decreased?
    a. Increased
    b. Decreased
    c. Remains the same
    d. There is not enough information to decide

14. Same question for P_2, holding the other P_i constant.
    a. Increased
    b. Decreased
    c. Remains the same
    d. There is not enough information to decide

15. Same question for P_3, holding the other P_i constant.
    a. Increased
    b. Decreased
    c. Remains the same
    d. There is not enough information to decide

16. Bob and Fred have the same body proportions and body density, but Bob is 5’0” tall and Fred is 6’0” tall. Bob weighs 100 pounds. How much does Fred weigh?
    a. 100 pounds
    b. 120 pounds
    c. 155 pounds
    d. 173 pounds
    e. 200 pounds
    f. There is not enough information to decide.

17. A certain gene is expressed at a nominal rate in an in vitro preparation of cells. When the cells are treated with a certain agent, they first increase the expression of the gene and then decrease the expression. The plot below shows the measured rate of expression with treatment.

How does the total amount of gene expression over 24 hours with treatment compare to total gene expression at nominal rate over the same time?

a. It is lower
b. It is the same
c. It is higher
d. There is not enough information to decide

18. A patient ingests a radioactive substance before a PET scan. The equation for the amount of radioactive substance remaining, S, at a time t, is

$$S(t) = S_0(1/2)^{t/\tau}$$

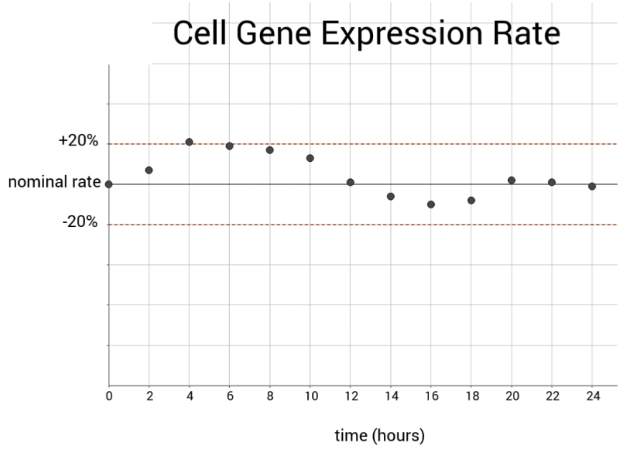
If you want the amount of radioactive substance to diminish very little over the course of an hour-long period, how should you choose which radioactive substance to use?

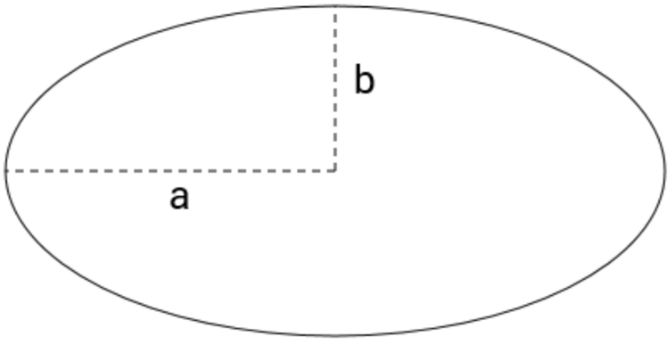
a. Choose the substance with the highest value of τ
b. Choose the substance with the lowest value of τ
c. Choose the substance with $\tau = t$
d. Choose the substance with $\tau = 1/2$

19. Which of these is closest to how fast an average person's hair grows?

a. $5*10^{-11}$ cm/s
b. $5*10^{-9}$ cm/s
c. $5*10^{-7}$ cm/s
d. $5*10^{-5}$ cm/s
e. $5*10^{-3}$ cm/s

Consider the ellipse below for questions 20 and 21.





20. Which of these is the formula for the area of the ellipse?
    a. πa^2
    b. πb^2
    c. $\pi(a+b)$
    d. πab
    e. $\pi(\frac{a+b}{2})^2$

21. How does the area of the ellipse change as you increase b?
    a. It increases
    b. It decreases
    c. It stays the same
    d. None - the area depends on *a*, not *b*

22. You buy 0.26 pints of olive oil for two dollars at the farmer's market. You plan next week to buy P pints of olive oil. Which expression gives how much this will cost?
    a. $(2*P)/0.26
    b. $P/0.26
    c. $(2*0.26)/P
    d. $(P*0.26)/2
    e. $0.26/(2*P)

For questions 23 and 24, consider the following statement:

"There are twelve times as many students as professors." Some students were asked to write an equation to represent this statement, using s for the number of students and p for the number of professors. Four of the students wrote the following:

Student 1 wrote: 12s/p
Student 2 wrote: 12s = p
Student 3 wrote: 12s + p
Student 4 wrote: s = 12p

23. Which student(s) is (are) correct?
    a. Only Student 1
    b. Only Student 2
    c. Only Student 3
    d. Only Student 4
    e. Students 1 and 2 are both correct
    f. Students 1 and 4 are both correct
    g. None of the students are correct

24. If the ratio of students to professors remains the same, how does the number of students vary as we increase the number of professors?
    a. It increases
    b. It decreases

c. It stays the same

d. None - this question does not make sense

25. This question may be used for survey validation purposes. Please select answer "d" to have your responses validated.

26. Here are some items and their monetary values per kilogram from most to least:

Item	Value per kilogram
Gem-quality diamond	$5,000,000
Printer Ink	$5,000
Silver	$450
Crude Oil	$0.20

Where would US $100 bills fit on this chart?

a. below crude oil

b. between crude oil and silver

c. between silver and printer ink

d. between printer ink and gem-quality diamond

e. more than gem-quality diamond

27. You step on a moving sidewalk moving forward at speed *s*. After going a little way, you realize you dropped your wallet before stepping on, so you turn around and run back to the beginning of the sidewalk. Your running speed is *r*. How fast would an observer standing on the ground next to the sidewalk see you moving?

a. $r + s$

b. $r - s$

c. $r * s$

d. s / r

28. The reduced mass of a two-body system is $\mu = \frac{m_1 m_2}{m_1 + m_2}$. If m_1 represents the mass of the earth and m_2 represents the mass of a small satellite, which of these would be the best approximation for μ?

a. m_1

b. m_2

c. $m_1 + m_2$

d. $m_1 m_2$

29. In mitosis, a single cell divides into two. If the combined volume of the daughter cells is the same as the volume of the parent cell, is the combined area of the daughter cells' membranes more or

less than the area of the parent cell's membrane? Assume the daughter cells have the same shape as the parent cell.

a. They would have half as much combined membrane area as the parent.
b. They would have less combined membrane area, but more than half as much.
c. They would have the same combined membrane area as the parent cells.
d. They would have more combined membrane area, but less than twice as much.
e. They would have twice as much combined membrane area as the parent cells.

30. A certain model for a small spherical object moving through a fluid says that the sphere experiences two resistive forces that tend to slow it down. These are inertial drag force, which is proportional to the square of the object's speed (v), and the viscous drag force, which is directly proportional to the object's speed. These are each represented by the equations

$$F_{inertial} = \tfrac{1}{2} C_d \varrho \pi R^2 v^2$$
$$F_{viscous} = 6\pi R \mu v$$

where C_d, ϱ, μ, π, and R can be treated as constants for a given object. Is there ever a speed when these two forces have the same magnitude?

a. Yes
b. No
c. Maybe - it depends on the values of the constants.
d. There is not enough information to decide.

31. How much effort did you put into this test?

a. I gave it my best effort
b. A lot
c. A medium amount
d. Only a little
e. No effort

32. What percent of your answers do you think are correct?

a. 80% to 100%
b. 60% to 80%
c. 40% to 60%
d. 20% to 40%
e. 0% to 20%

Appendix E: MEGS test administrations

This appendix summarizes the administrations of the MEGS.

All UMD test administrations had approximately 250 students for on-sequence courses (Fall 131, Spring 132) and approximately 125 students for off-sequence courses.

S is Swarthmore, MC is Montgomery College. About 30 students at these school took the MEGS. Their results were very similar to UMD students. These students did not do a post-test and we haven't included their data in our analysis beyond noting it here.

Additionally, we administered a free response version of the MEGS for validation in spring 2018.

"Policy" describes how students received credit for taking the MEGS. "participation" means all students earned a small amount of credit for taking the MEGS, regardless of their performance. "Score" means the MEGS was scored and counted as a student grade like a normal assignment. "Mixed" means that students in different sections may have received different instructions.

Administration	Pre/Post	Campus	Course	Dates	MEGS version	poli
Year 1						
fall 2015	Pre	UMD	131	Aug 2015	0.8	parti
fall 2015	Post	UMD	131	Dec 7 - 10, 2015	0.8	parti
spring 2016	Pre	UMD	131	Jan 2016	0.9	parti
spring 2016	Post	UMD	131	May 2 - 5, 2016	0.10	parti
spring 2016	Post	UMD	132	May 2016	0.10	parti
Year 2						
fall 2016	Pre	UMD	131	Aug 2016	0.10	parti
	Pre	UMD	132	Aug 2016	0.10	parti
fall 2016	Post	UMD	131	Dec 2016	0.10	parti
	Post	UMD	132	Dec 2016	0.10	parti
spring 2017	Pre	UMD	131	Jan 2017	1.1	parti
	Pre	UMD	132	Jan 2017	1.1	mixe
	Pre	S	131	Jan 2017	1.1	parti
	Pre	MC	131	Jan 2017	1.1	parti
spring 2017	Post	UMD	131	May 2017	1.1	score
	Post	UMD	132	May 2017	1.1	score
fall 2017	Pre	UMD	131	Aug 2017	1.1	parti
	Pre	UMD	132	Aug 2017	1.1	parti
	Post	UMD	131	Dec 2017	1.1	mixe
	Post	UMD	132	Dec 2017	1.1	parti

Table E.1: Locations, dates, and versions of MEGS administrations

Bibliography

Physport: Assessments. `https://www.physport.org/assessments/`. Accessed: 2018-02-28.

`https://stackoverflow.com/questions/20364939/community-detection-with-infomap-algorithm-producing-one-massive-module`.

Wendy K Adams, Katherine K Perkins, Noah S Podolefsky, Michael Dubson, Noah D Finkelstein, and Carl E Wieman. New instrument for measuring student beliefs about physics and learning physics: The colorado learning attitudes about science survey. *Physical review special topics-physics education research*, 2 (1):010101, 2006.

John Airey and Cedric Linder. A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes. *Journal of Research in Science Teaching*, 46(1):27–49, 2009.

Russell G Almond, Robert J Mislevy, Linda S Steinberg, Duanli Yan, and David M Williamson. *Bayesian networks in educational assessment*. Springer, 2015.

Abraham Arcavi. Symbol sense: Informal sense-making in formal mathematics. *For the learning of Mathematics*, 14(3):24–35, 1994.

Raquel Ataide and Ileana Greca. Pre-service physics teachers theorems-in-action about problem solving and its relation with epistemic views on the relationship between physics and mathematics in understanding physics. In Gesche Pospiech, editor, *Mathematics in Physics Education Research*. Springer, in press.

Anita Bandrowski, Ryan Brinkman, Mathias Brochhausen, Matthew H Brush, Bill Bug, Marcus C Chibucos, Kevin Clancy, Mélanie Courtot, Dirk Derom, Michel Dumontier, et al. The ontology for biomedical investigations. *PloS one*, 11(4): e0154556, 2016.

Lei Bao and Edward F Redish. Model analysis: Representing and assessing the dynamics of student learning. *Physical Review Special Topics-Physics Education Research*, 2(1):010103, 2006.

Michael J. Barber. Modularity and community detection in bipartite networks. *Physical Review E*, 76(6):066102, 2007. ISSN 15393755. doi: 10.1103/PhysRevE.76.066102. URL http://arxiv.org/abs/0707.1616http://dx.doi.org/10.1103/PhysRevE.76.066102.

Onofrio Rosario Battaglia and Claudio Fazio. Response patterns and knowledge conceptual dimensions in engineering freshmen answers to force concept inventory questions.

BBC. Forces and braking, 2014. URL http://www.bbc.co.uk/schools/gcsebitesize/science/add_aqa/forces/forcesbrakingrev1.shtml. [Online; accessed 20-May-2018].

Stephen J Beckett. Improved community detection in weighted bipartite networks. *Royal Society open science*, 3(1):140536, 2016.

Carl M Bender and Steven A Orszag. *Advanced Mathematical Methods for Scientists and Engineers I*. Springer Science & Business Media, 1999.

C Bernaards and R Jennrich. Gparotation: Gpa factor rotation. r package version: 2012.3-1, 2012.

Thomas J Bing and Edward F. Redish. Analyzing problem solving using math in physics: Epistemological framing via warrants. *Physical Review Special Topics - Physics Education Research*, 5(2):020108, dec 2009a. ISSN 1554-9178. doi: 10.1103/PhysRevSTPER.5.020108. URL http://link.aps.org/doi/10.1103/PhysRevSTPER.5.020108.

Thomas J Bing and Edward F Redish. Analyzing problem solving using math in physics: Epistemological framing via warrants. *Physical Review Special Topics-Physics Education Research*, 5(2):020108, 2009b.

Thomas J Bing and Edward F Redish. Epistemic complexity and the journeyman-expert transition. *Physical Review Special Topics-Physics Education Research*, 8 (1):010105, 2012.

S Brahmia, Andrew Boudreaux, and Stephen E Kanim. Obstacles to mathematization in introductory physics. *arXiv preprint arXiv:1601.01235*, 2016.

Eric Brewe, Jesper Bruun, and Ian G Bearden. Using module analysis for multiple choice responses: A new method applied to force concept inventory data. *Physical Review Physics Education Research*, 12(2):020131, 2016.

Percy Williams Bridgman. *Dimensional analysis*. Yale University Press, 1922.

David T Brookes and Eugenia Etkina. force, ontology, and language. *Physical Review Special Topics-Physics Education Research*, 5(1):010110, 2009.

Bryan Caplan. *The Case Against Education: Why the Education System is a Waste of Time and Money.* Princeton University Press, 2018.

Heidi B Carlone, Catherine M Scott, and Cassi Lowder. Becoming (less) scientific: A longitudinal study of students' identity work from elementary to middle school science. *Journal of Research in Science Teaching*, 51(7):836–869, 2014.

Michelene TH Chi and James D Slotta. The ontological coherence of intuitive physics. *Cognition and instruction*, 10(2-3):249–260, 1993.

Michelene TH Chi, Paul J Feltovich, and Robert Glaser. Categorization and representation of physics problems by experts and novices. *Cognitive science*, 5(2): 121–152, 1981.

Eugene Chiang. Astronomy 250: Order-of-magnitude physics. `http://w.astro.berkeley.edu/~echiang/oom/oom.html`. Accessed: 2017-05-04.

John Clement, Jack Lochhead, and George S Monk. Translation difficulties in learning mathematics. *The American Mathematical Monthly*, 88(4):286–290, 1981.

John J Clement and L Stephens. Extreme case reasoning and model based learning experts and students. In *Proceedings of the 2009 Annual Meeting of the National Association for Research in Science Learning*, 2009.

Allan Collins and William Ferguson. Epistemic forms and epistemic games: Structures and strategies to guide inquiry. *Educational psychologist*, 28(1):25–42, 1993.

National Research Council et al. *How people learn: Brain, mind, experience, and school: Expanded edition.* National Academies Press, 2000.

Matthew Gordon Ray Courtney and Matthew Gordon. Determining the number of factors to retain in efa: Using the spss r-menu v2. 0 to make more judicious estimations. *Practical assessment, research & evaluation*, 18(8):60, 2013.

Linda Crocker and James Algina. *Introduction to classical and modern test theory.* ERIC, 1986.

Jessica T DeCuir-Gunby, Patricia L Marshall, and Allison W McCulloch. Developing and using a codebook for the analysis of interview data: An example from a professional development research project. *Field methods*, 23(2):136–155, 2011.

Andrea a. DiSessa. Toward an Epistemology of Physics. *Cognition and Instruction*, 10(2):105–225, 1993. ISSN 0737-0008. doi: 10.1080/07370008.1985.9649008.

Andy A Disessa. Knowledge in pieces. 1988.

David P Doane and Lori E Seward. Measuring skewness: a forgotten statistic? *Journal of Statistics Education*, 19(2), 2011.

Carsten F Dormann, Jochen Fruend, Bernd Gruber, Maintainer Carsten F Dormann, and TRUE LazyData. Package bipartite. 2017.

Benjamin W Dreyfus, Ayush Gupta, and Edward F Redish. Applying conceptual blending to model coordinated use of multiple ontological metaphors. *International Journal of Science Education*, 37(5-6):812–838, 2015.

Nikk Effingham. *An introduction to ontology.* John Wiley & Sons, 2013.

Mark Eichenlaub. Why did people create complex numbers?, 2012. URL `https://www.quora.com/Why-did-people-create-complex-numbers/answer/Mark-Eichenlaub`. [Online; accessed 10-April-2018].

Mark Eichenlaub. Do grad school students remember everything they were taught in college all the time?, 2013. URL `https://www.quora.com/Do-grad-school-students-remember-everything-they-were-taught-in-college-all-the`
`answer/Mark-Eichenlaub`. [Online; accessed 03-May-2018].

Andrew Elby. Helping physics students learn how to learn. *American Journal of Physics*, 69(S1):S54–S64, 2001.

Paula V Engelhardt. An introduction to classical test theory as applied to conceptual multiple-choice tests. *Getting Started in PER*, 2, 2009.

Jerome Epstein. Report on basic skills test given to various student populations. 1993.

K Anders Ericsson and Herbert A Simon. Verbal reports as data. *Psychological review*, 87(3):215, 1980.

Richard P Feynman. *" Surely You're Joking, Mr. Feynman!": Adventures of a Curious Character: Adventures of a Curious Character*. WW Norton & Company, 2010.

Douglas C Giancoli. *Physics for scientists and engineers*, volume 3. Prentice hall Upper Saddle River, NJ, 2000.

Michelle Girvan. Phys 615: Nonlinear dynamics of extended systems. `https://umdphysics.umd.edu/images/syllabi/2015/phys615-girvan.pdf`. Accessed: 2018-05-10.

Erving Goffman. *Frame analysis: An essay on the organization of experience.* Harvard University Press, 1974.

Ayush Gupta and Andrew Elby. Beyond epistemological deficits: Dynamic explanations of engineering students difficulties with mathematical sense-making. *International Journal of Science Education*, 33(18):2463–2488, 2011.

Ayush Gupta, David Hammer, and Edward F. Redish. The Case for Dynamic Models of Learners' Ontologies in Physics. *Journal of Learning Sciences*, 19(3): 285–321, 2010a. ISSN 1050-8406. doi: 10.1080/10508406.2010.491751. URL http://arxiv.org/abs/0802.4278.

Ayush Gupta, David Hammer, and Edward F Redish. The case for dynamic models of learners' ontologies in physics. *The Journal of the Learning Sciences*, 19(3): 285–321, 2010b.

Ayush Gupta, Andrew Elby, and Luke D Conlin. How substance-based ontologies for gravity can be productive: A case study. *Physical Review Special Topics-Physics Education Research*, 10(1):010113, 2014.

Richard R Hake. Interactive Engagement vs. Traditional Methods: A Six-Thousand-Student Survey of Mechanics Test Data for Introductory Physics Courses. 1998a.

Richard R Hake. Interactive-engagement vs. traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. 1998b.

Paul R Halmos. *Naive set theory.* Courier Dover Publications, 2017.

David Hammer. Student inquiry in a physics class discussion. *Cognition and Instruction*, 13(3):401–430, 1995.

David Hammer. Student resources for learning introductory physics. *American Journal of Physics*, 68(S1):S52, jul 2000. ISSN 00029505. doi: 10.1119/1.19520. URL http://link.aip.org/link/?AJP/68/S52/1{&}Agg=doi.

David Hammer and Andrew Elby. Tapping epistemological resources for learning physics. *The Journal of the Learning Sciences*, 12(1):53–90, 2003.

David Hammer and Andrew Elby. Tapping Epistemological Resources for Learning Physics. (December 2014):37–41, 2009. doi: 10.1207/S15327809JLS1201.

David Hammer, Andrew Elby, Rachel E Scherr, and Edward F Redish. Resources, framing, and transfer. *Transfer of learning from a modern multidisciplinary perspective*, pages 89–120, 2005.

Sadri Hassani. *Mathematical physics: a modern introduction to its foundations.* Springer Science & Business Media, 2013.

Andrew F Heckler. Some consequences of prompting novice physics students to construct force diagrams. *International Journal of Science Education*, 32(14): 1829–1851, 2010.

Frank Heile. Why is the force between two charged particles eerily similar to the force between two large masses?, 2015. URL https://www.quora.com/Why-is-the-force-between-two-charged-particles-eerily-similar-to-the-force-betw answer/Frank-Heile. [Online; accessed 14-December-2015].

Patricia Heller and Douglas Huffman. Interpreting the force concept inventory: A reply to Hestenes and Halloun. *The Physics Teacher*, 33(8):503, 1995. ISSN 0031921X. doi: 10.1119/1.2344279.

David Hestenes, Malcolm Wells, and Gregg Swackhamer. Force concept inventory. *The Physics Teacher*, 30(3):141, mar 1992a. ISSN 0031921X. doi: 10.1119/1.2343497. URL `http://scitation.aip.org/content/aapt/journal/tpt/30/3/10.1119/1.2343497`.

David Hestenes, Malcolm Wells, Gregg Swackhamer, et al. Force concept inventory. *The physics teacher*, 30(3):141–158, 1992b.

Kathleen Hogan. Relating students' personal frameworks for science learning to their cognition in collaborative contexts. *Science education*, 83(1):1–32, 1999.

NG Holmes, Dhaneesh Kumar, and DA Bonn. Toolboxes and handing students a hammer: The effects of cueing and instruction on getting students to think critically. *Physical Review Physics Education Research*, 13(1):010116, 2017.

Leonardo Hsu, Eric Brewe, Thomas M Foster, and Kathleen A Harper. Resource letter rps-1: Research in problem solving. *American Journal of Physics*, 72(9): 1147–1156, 2004.

Douglas W Hubbard. *How to measure anything: Finding the value of intangibles in business*. John Wiley & Sons, 2014.

Douglas Huffman and Patricia Heller. What does the force concept inventory actually measure? *The Physics Teacher*, 33(3):138, 1995. ISSN 0031921X. doi: 10.1119/1.2344171. URL `http://eric.ed.gov/ERICWebPortal/recordDetail?accno=EJ500178`.

Edwin Hutchins. *Cognition in the Wild*. MIT press, 1995.

Eric Kuo, Michael M Hull, Ayush Gupta, and Andrew Elby. How students blend conceptual and formal mathematical reasoning in solving physics problems. *Science Education*, 97(1):32–57, 2013.

Eric Kuo, Nicole R Hallinen, and Luke D Conlin. How prompting force diagrams discourages student use of adaptive problem-solving shortcuts. In *The Physics Education Research Conference 2015*, 2015.

Mary Bridget Kustusch, David Roundy, Tevian Dray, and Corinne A. Manogue. Partial derivative games in thermodynamics: A cognitive task analysis. *Physical Review Special Topics - Physics Education Research*, 10(1):1–16, 2014. ISSN 15549178. doi: 10.1103/PhysRevSTPER.10.010101.

George Lakoff and Mark Johnson. *Metaphors we live by*. University of Chicago press, 2008.

Anton E Lawson. The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, 15(1):11–24, 1978.

Beth A Lindsey and Megan L Nagel. Do students know what they know? exploring the accuracy of students self-assessments. *Physical Review Special Topics-Physics Education Research*, 11(2):020103, 2015.

Lulu Liu and Mark Eichenlaub. Reading physics, 2015. URL `http://readingphysics.tumblr.com`.

Adrian Madsen, Sarah B McKagan, and Eleanor C Sayre. Best practices for administering concept inventories. *arXiv preprint arXiv:1404.6500*, 2014.

Adrian Madsen, Sam McKagan, and Eleanor Sayre. Addressing common concerns about concept inventories, 2017. URL `https://www.physport.org/recommendations/Entry.cfm?ID=93462`. [Online; accessed 14-March-2018].

Sandra P Marshall. Assessing problem solving: A short-term remedy and a long-term solution. *The teaching and assessing of mathematical problem solving*, 3: 159–177, 1988.

Jeffrey Marx and Karen Cummings. Development of a survey instrument to gauge students problem-solving abilities. In *AIP Conference Proceedings*, volume 1289, pages 221–224. AIP, 2010.

Timothy L Mccaskey, Andrew R Elby, Rebecca F Lippman, and Edward F Redish. The MPEX2 : Modification of a Survey Instrument The Story So Far. pages 1–17, 2003.

Same McKagan, Eleanor Sayre, and Adrian Madsen. Normalized gain: What is it and when and how should i use it?, 2017. URL `https://www.physport.org/recommendations/Entry.cfm?ID=93334`. [Online; accessed 14-March-2018].

Hugh Mehan. *Learning lessons*. Harvard University Press Cambridge, MA, 1979.

David E Meltzer and Valerie K Otero. A brief history of physics education in the united states. *American Journal of Physics*, 83(5):447–458, 2015.

George A Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

Marvin Minsky. Society of mind: a response to four reviews. *Artificial Intelligence*, 48(3):371–396, 1991.

Bahar Modir, Paul W Irving, Steven F Wolf, and Eleanor C Sayre. Learning about the Energy of a Hurricane System through an Estimation Epistemic Game. *Proceedings of the 2014 Physics Education Research Conference*, pages 3–6, 2014. doi: 10.1119/perc.2014.pr.044.

David Morin. *Introduction to classical mechanics: with problems and solutions.* Cambridge University Press, 2008.

R Nave. Coulomb's law, 2017. URL `http://hyperphysics.phy-astr.gsu.edu/hbase/electric/elefor.html`. [Online; accessed 10-May-2017].

James C Nearing et al. *Mathematical Tools for Physics.* Dover Publications, 2003.

Tristan Needham. *Visual complex analysis.* Oxford University Press, 1998.

Mark EJ Newman. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23):8577–8582, 2006.

Rob Phillips, Jane Kondev, Julie Theriot, and Hernan Garcia. *Physical biology of the cell.* Garland Science, 2012.

Sterl Phinney. Ph 101 order of magnitude physics. `https://www.its.caltech.edu/~oom/`. Accessed: 2017-05-04.

G Pospiech. *Mathematics in Physics Education.* Springer, in press.

G Raiche. nfactors: An r package for parallel analysis and non graphical solutions to the cattell scree test. *R package version*, 2(3), 2010.

Edward F Redish. A theoretical framework for physics education research: Modeling student thinking. *arXiv preprint physics/0411149*, 2004.

Edward F Redish. Oersted Lecture 2013 : How should we think about how our students think ? pages 1–19, 2013.

Edward F Redish and Todd J Cooke. Learning each other's ropes: negotiating interdisciplinary authenticity. *CBE-Life Sciences Education*, 12(2):175–186, 2013.

Edward F Redish and Eric Kuo. Language of physics, language of math: Disciplinary culture and dynamic epistemology. *Science & Education*, 24(5-6):561–590, 2015.

Edward F Redish, Jeffery M Saul, and Richard N Steinberg. Student expectations in introductory physics. *American Journal of Physics*, 66(3):212–224, 1998.

Edward F Redish, Chandralekha Singh, Mel Sabella, and Sanjay Rebello. Introducing students to the culture of physics: Explicating elements of the hidden curriculum. In *AIP Conference Proceedings*, volume 1289, pages 49–52. AIP, 2010.

EF Redish, C Bauer, KL Carleton, TJ Cooke, M Cooper, Catherine Hirshfeld Crouch, BW Dreyfus, B Geller, J Giannini, J Svoboda Gouvea, et al. Nexus/physics: An interdisciplinary repurposing of physics for biologists. *arXiv preprint arXiv:1308.4947*, 2013.

EF Redish, C Bauer, KL Carleton, TJ Cooke, M Cooper, Catherine Hirshfeld Crouch, BW Dreyfus, BD Geller, J Giannini, J Svoboda Gouvea, et al. Nexus/physics: An interdisciplinary repurposing of physics for biologists. *American Journal of Physics*, 82(5):368–377, 2014.

RW Robinett. Dimensional analysis as the other language of physics. *American Journal of Physics*, 83(4):353–361, 2015.

Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2009.

Rosemary S Russ, Janet E Coffey, David Hammer, and Paul Hutchison. Making classroom assessment more accountable to scientific reasoning: A case for attending to mechanistic thinking. *Science Education*, 93(5):875–891, 2009.

Rachel E Scherr. Gesture analysis for physics education researchers. *Physical Review Special Topics-Physics Education Research*, 4(1):010101, 2008.

Alan H Schoenfeld and Alan H Sloane. *Mathematical thinking and problem solving.* Routledge, 2016.

Terry F Scott and Dániel Schumayer. Conceptual coherence of non-newtonian worldviews in force concept inventory data. *Physical Review Physics Education Research*, 13(1):010126, 2017.

Terry F Scott and Dániel Schumayer. Central distractors in force concept inventory data. *Physical Review Physics Education Research*, 14(1):010106, 2018.

Terry F. Scott, Daniel Schumayer, and Andrew R. Gray. Exploratory factor analysis of a Force Concept Inventory data set. *Physical Review Special Topics - Physics Education Research*, 8(2):020105, 2012. ISSN 1554-9178. doi: 10.1103/PhysRevSTPER.8.020105. URL `http://link.aps.org/doi/10.1103/PhysRevSTPER.8.020105`.

M Ángeles Serrano, Marián Boguná, and Alessandro Vespignani. Extracting the multiscale backbone of complex weighted networks. *Proceedings of the national academy of sciences*, 106(16):6483–6488, 2009.

Raymond A. Serway and John W. Jewett. *Physics for scientists and engineers.* Thomson-Brooks/Cole, 6th ed edition, 2004. ISBN 9780534408428,0495142425,0534408427,9780495142423.

Sev. Why is negative times negative = positive?, 2010. URL `https://math.stackexchange.com/questions/9933/why-is-negative-times-negative-positive`.

Anna Sfard. On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational studies in mathematics*, 22(1):1–36, 1991.

Bruce L Sherin. How students understand physics equations. *Cognition and instruction*, 19(4):479–541, 2001.

Tiffany-Rose Sikorski, Gary D. White, and Justin Landay. Uptake of solution checks by undergraduate physics students. In *2017 PERC Proceedings*, pages 368–371. AAPT, 2017.

Nate Silver. *The signal and the noise: why so many predictions fail–but some don't*. Penguin, 2012.

Fred Sommers. On concepts of truth in natural languages. *The Review of Metaphysics*, pages 259–286, 1969.

Benjamin T Spike and Noah D Finkelstein. Design and application of a framework for examining the beliefs and practices of physics teaching assistants. *Physical Review Physics Education Research*, 12(1):010114, 2016.

John Stewart, Cabot Zabriskie, Seth DeVore, and Gay Stewart. Multidimensional item response theory and the force concept inventory. *arXiv preprint arXiv:1803.02399*, 2018.

Ronald K Thornton. Measuring and improving student mathematical skills for modeling.

Paul A Tipler and Gene Mosca. *Physics for scientists and engineers*. Macmillan, 2007.

Kenneth G Tobin and William Capie. The development and validation of a group test of logical thinking. *Educational and Psychological Measurement*, 41(2):413–423, 1981.

Stephen E Toulmin. *The uses of argument*. Cambridge university press, 2003.

John S Townsend. *A modern approach to quantum mechanics*. University Science Books, 2000.

Jonathan Tuminaro. *A cognitive framework for analyzing and describing introductory students' use and understanding of mathematics in physics*. PhD thesis, 2004.

Jonathan Tuminaro and Edward F Redish. Elements of a cognitive model of physics problem solving: Epistemic games. *Physical Review Special Topics-Physics Education Research*, 3(2):020101, 2007.

Steven Weinberg. A model of leptons. *Physical review letters*, 19(21):1264, 1967.

Larry Weinstein. Fermi questions, 2018. URL `https://aapt.scitation.org/topic/collections/fermi-questions`. [Online: accessed 20-May-2018].

Kasper Welbers. semnet: Semantic network analysis. https://rdrr.io/github/kasperwelbers/semnet/. Accessed: 2018-05-17.

Gary White, Tiffany-Rose Sikorski, and Justin Landay. Metacognitive gimmicks and their use by upper level physics students. In *APS April Meeting Abstracts*, 2017.

Wikipedia. Variation of information. https://en.wikipedia.org/wiki/Variation_of_information. Accessed: 2018-05-12.

Wikipedia. Yukawa potential — wikipedia, the free encyclopedia, 2016. URL https://en.wikipedia.org/w/index.php?title=Yukawa_potential&oldid=715463242. [Online; accessed 3-October-2016].

Wikipedia contributors. Function (mathematics) — Wikipedia, the free encyclopedia, 2018. URL https://en.wikipedia.org/w/index.php?title=Function_(mathematics)&oldid=841910002. [Online; accessed 20-May-2018].

Bethany R. Wilcox, Marcos D. Caballero, Daniel A. Rehn, and Steven J. Pollock. Analytic framework for students' use of mathematics in upper-division physics. *Physical Review Special Topics - Physics Education Research*, 9(2), 2013. ISSN 15549178. doi: 10.1103/PhysRevSTPER.9.020119.

Michael C. Wittmann and Katrina E. Black. Mathematical actions as procedural resources: An example from the separation of variables. *Physical Review Special Topics - Physics Education Research*, 11(2):020114, 2015. ISSN 1554-9178. doi: 10.1103/PhysRevSTPER.11.020114. URL http://link.aps.org/doi/10.1103/PhysRevSTPER.11.020114.

Michael C Wittmann, Virginia J Flood, and Katrina E Black. Algebraic manipulation as motion within a landscape. *Educational Studies in Mathematics*, 82(2): 169–181, 2013.

Jun-ichiro Yasuda, Naohiro Mae, Michael M Hull, and Masa-aki Taniguchi. Analyzing false positives of four questions in the force concept inventory. *Physical Review Physics Education Research*, 14(1):010112, 2018.

Anthony Zee. *Quantum field theory in a nutshell*. Princeton university press, 2010.