Early-Term Epistemological Optimism:

Summer 2012 Analysis

Timothy D. Brown

Department of Physics, the University of Tulsa, Tulsa, OK, 74104

Abstract

This document provides a summary of all analysis done on the Early-Term Optimism sub-project from the Rapid Assessment and Web Reports (RAWR) project completed during the summer of 2012. The document summarizes all pertinent parts of the analysis, including RAWR's background, the analysis done within the R programming environment, links to graphs and figures, and conclusions.

1 Project Background

1.1 Getting Started

1.1.1 What is RAWR?

The Rapid Assessment and Web Reports (RAWR) project is a joint effort between the physics education research groups at Kansas State University (KSU) and the Rochester Institute of Technology (RIT), led by Eleanor C. Sayre (KSU) and Scott V. Franklin (RIT). RAWR is essentially an online survey consisting of a selection of "tasks," which are short forced-choice surveys about some specific physics topic, such as electric fields or epistemological beliefs. The surveys are administered once a week to introductory physics students, allowing weekly data samples to be drawn. RAWR is meant to be a large-scale study over many students' responses, and so the survey format is chosen over the richer interview method since surveys are less time-consuming to administer and analyze.

The RAWR task questions can be accessed at https://rawr.rit.edu/rawr/, with **Username: Student, Password: webcurves**.

The RAWR system makes use of Dr. Sayre and others' "Response-Curve Method" for collecting data. The idea is that every student completes one of the tasks every week, and the tasks are randomly assigned. A between-students analysis is therefore used, with each week considered a separate group: this elimiates retesting effects due to a student taking a test more than once. The task score collected for each week is assumed to be representative of the population, allowing a score-time trend to be graphed, the response curve of the task scores. We expect this method to be sensitive to by-week effects, such as the topic being taught or whether or not there is a test.

1.1.2 What is Early-Term Optimism?

Early-Term Optimism refers specifically to an effect observed within the Epistemology 1 and 2 and Identity 1 and 2 tasks on RAWR. These task questions deal primarily with the quality of students' epistemological beliefs and expectations for a physics course. Multiple previous studies have suggested that students' epistemological beliefs become less "sophisticated" (defined in Previous Studies) when comparing beliefs between before and after an introductory physics course. Students are more optimistic about their epistemological beliefs early in the semester, hence "Early-Term Optimism." The purpose of this project is to look at epistemological beliefs with the response-curve method to see if additionally, data supports a general trend of decreasing sophistication throughout the semester.

1.1.3 What is R?

R is a statistical analysis environment. It may be downloaded free from the internet at http://www.r-project.org. R has its own object-oriented programming language, similar to Matlab. One may write functions and scripts in R to accomplish certain tasks, streamlining data analysis. R comes pre-loaded with many functions for statistical analysis, for example ANOVA tests and linear correlation tests.

1.2 Previous Studies

The reader will find the scientific papers "Student expectations in introductory physics" by E. F. Redish, et al., "Correlating Student Beliefs With Student Learning Using The Colorado Learning Attitudes About Science

Survey" by K. K Perkins et al., and "The Idea Behind EBAPS" by A. Elby, et al. useful to reference in understanding the project. These papers all outline some of the previous work done with epistemology surveys.

Redish introduced the Maryland Physics Expectations (MPEX) survey in his paper, with a means of quantifying student responses by classifying the "sophistication" of their answers as expert-like or novice-like. They did this by administering the survey to a large variety of groups varying from professors with phD's to high schoolers. The phD's and other physics practitioners tended to agree in their responses to the survey, and these responses were deemed expert-like; a similar correlation was drawn between those less experienced in physics and novice-like sophistication. This quantification of epistemological sophistication has been well-used in similar studies (including ours). Furthermore, the Redish study purported to find a tendency for student sophistications to become more novice-like after instruction than they were before instruction.

The Perkins paper introduced the Colorado Learning Attitudes about Science Survey (CLASS) as a similar diagnostic for probing student epistemological beliefs. The study found several interesting results, like those with higher sophistications were more inclined to continue in physics classes, and also found (like the MPEX) that students' belief sophistication tended to decrease when measured before and after the semester. But perhaps the most original result was the correlation between more expert-like epistemological sophistication and more conceptual gains throughout the class. This finding promoted additional research into students' epistemologies, since a more expert-like belief system evidently correlates with increased physics learning.

The Elby paper introduced the Epistemological Beliefs Asssessment for Physical Science (EBAPS) as an attempt to improve upon the MPEX. In the paper Elby offers a critique of the MPEX, mainly that the MPEX "conflate[s] student beliefs about knowing with course-specific expectations and goals." Elby's EBAPS is an attempt to remove the course-specific element in order to imporove the ability to probe general epistemological belief structures.

Given that our goal as physics education researchers is to improve students' ability to understand and use physics concepts and methodology, the CLASS paper offers a compelling reason to pursue a model of student belief changes. The first compelling result in studying student epistemologies was that epistemological sophistication tended to decrease between the beginning and end of an introductory physics course. However, the pre/post testing design meant that a lot of the within-semester data was overlooked. In order to get a more complete picture we need to take data multiple times during the semester. This is the gist of the RAWR project.

2 Data Set

The data comes from responses by students enrolled in introductory classes at RIT and the United States Military Academy at Westpoint (USMA) to the Epistemology 1 and 2 (E-1, E-2) and Identity 1 and 2 (I-1, I-2) tasks on RAWR. The data was collected for approximately two years, between Fall 2010 and Winter 2011, inclusive. For any given observation, relevant data such as the task question, student response to question, student major, student math class, student physics class, student gender and etc are collected. Each distinct student is assigned a numerical ID number purely for matching purposes. Similarly, each task question is assigned a numerical ID for matching purposes.

2.1 School Differences

There are several intriguing differences between the data from RIT and USMA students. Please check that the data reflects the things below before performing analysis: data has been known to get lost in translation from RIT to here.

2.1.1 RIT

RIT has a workshop (studio) based style of introductory physics class, which mixes elements of lecture and lab-based instruction into one format. RIT also has course quarters as opposed to the traditional semester, with each quarter lasting 10 weeks. The introductory physics sequence is Physics 1 (Newtonian mechanics), then Physics 2 (Rotational motion and wave phenomena), then Physics 3 (Electricity and magnetism). Furthermore, if a student fails either of Physics 1 or 2, they retake the "remedial" version of the same class; for example a student who fails Physics 1 will be enrolled next in Physics 1a. The remedial and non-remedial courses are recombined when entering Physics 3. The remedial physics class differs from its doppelganger only by an additional two hours of instruction per week.

There is data from all physics courses (Physics 1, 1a, 2, 2a, 3) and all quarters (Fall 2010, Winter 2010, Spring 2011, Fall 2011, Winter 2011) for the Epistemology 1 and 2 data. There is only data from Physics 2 and 2a from all quarters for the Identity 1 data. There is only data from Physics 2 and 2a from Fall 2010 for the Identity 2 data. For each task for each semester for each course, there are approximately between 200 and 1200 observations, with most having around 700 observations.

2.1.2 USMA

USMA has traditional lecture-lab-recitation instruction format, with sixteen weeks per semester. All USMA data comes only from the Physics 1 (Newtonian mechanics and rotational motion) course. There is no remedial class sequence at USMA.

There is no USMA data for the Epistemology 1 task. There is data from the Fall 2011 semester for Epistemology 2. There is similarly data from only the Fall 2011 semester for Identity 1 and 2. This USMA data for each task has greater than 10,000 observations.

2.2 Tasks

Although all 4 tasks (E-1, E-2, I-1, I-2) are meant to probe students' epistemological beliefs and physics expectations they are distinct. Under the epistemologcal framework theory, student epistemologies are coherent belief systems with several orthogonal axes or dimensions. Each task is designed to probe a specific epistemological axis. Every task question is a "Likert-style" question meaning the student is asked to respond to a statement with varying degrees of agreement. These responses are easily coded into a numerical response, with a unique number corresponding to each possible level of agreement.

2.2.1 Epistemology 1

Questions 68-76 on RAWR comprise the Epistemology 1 task. E-1 is essentially an excerpt from the Epistemological Beliefs Assessment for Physical Science (EBAPS) created by A. Elby and others, a five-level ("strongly disagree"-"disagree"-"neutral"-"agree"-"strongly agree") Likert survey. The questions comprise two epistemological axes: "Real-Life Applicability," referring to students beliefs' on how much science and experimentation actually relate to the real world, and “Source of Ability to Learn,” referring to students’ beliefs on whether scientific knowledge is something that owes more to “hard work” or “natural ability.” To grade, each response is coded to a score from 0-4 depending on how close each response is to most favorable. The document labeled “EBAPS scoring scheme” will be useful to the reader.

2.2.2 Epistemology 2

Questions 144-155 on RAWR comprise the Epistemology 2 task. E-2 is essentially an excerpt from the Colorado Learning Attitudes about Science Survey (CLASS) created by K. K. Perkins and others, also a five-level Likert survey. These questions also comprise two epistemological axes: “Problem Solving-General” and “Problem Solving- Sophistication.” These questions tend to probe the methods students use to solve problems, but also seem to overlap somewhat with the EBAPS axes above. To grade, each response is coded to a score from 0-2 depending on how close each response is to most favorable (strong disagree and disagree, strong agree and agree are collapsed). The annotated copy of the CLASS will be helpful to the reader.

2.2.3 Identity 1 & 2

Questions 170-195 and 196-221 comprise the Identity 1 and Identity 2 tasks, respectively. I-1 and I-2 are both excerpts from the SLIDe survey created by Sissi Li and others, an 11-level Likert survey. Dr. Li should be the main contact in any investigation into SLIDe. There are many subcategories in the SLIDe task, but there are enough questions that are unassigned that I have found it more helpful to not separate the questions by subcategory. These questions tend to have some overlap with both the Epistemology 1 and 2 tasks, but the SLIDe additionally probes how students view their relationship with their teacher, and the responsibility of each for learning to occur. To grade, each response is coded to a score from 0-10 depending on how close each response is to most favorable.

3 Things to Know for Analysis

3.1 Classical Regression Statistics

This information is taken from http://www.wadsworth.com/psychology_d/templates/student_resources/workshops/stats_wrk.html. The reader will also likely find *What is a p-value anyway?: 34 Stories to Help You Actually Understand Statistics* by Andrew Vickers useful in clearing up questions on classical statistics.

3.1.1 ANOVA Test

The analysis of variance (ANOVA) test is one way in which we can detect trends in sets of data. Essentially, given a data set correlated to one or more factors, the ANOVA test computes how likely it is that the data from those factors comes from the same population. The ANOVA test assumes a linear model to the data set: $x = \mu + \alpha + \epsilon$, where x is an individual score, μ is the population mean, α is distance to the group mean, and ϵ is the error (within-group variance). If α is non-zero, we reasonably claim that there is some likelihood that the data come from different groups determined by the factor. The output of an ANOVA test is an F-ratio of the mean-squares of between groups effects to the mean-squares of the within-groups effects. This F-ratio can then be related to a probability that such an F-ratio could occur by chance. This gives us an idea of statistical significance, for our study: if $p < 0.1$ the differences between the groups are [.] significant, if $p < 0.05$ the differences are [*] significant ($2 - \sigma$ significant), if $p < 0.01$ the differences are [**] significant ($3 - \sigma$) and if $p < 0.001$ the differences are [***] ($4 - \sigma$) significant.

The ANOVA test's primary purpose is to tell us if some factor has a statistically significant effect on a data set, and if we may reject the null hypothesis (we have no evidence for significant differences) for that factor. If we decide to reject the null hypothesis, the data set separated by the factor each represent a new population, and we may not bin those sets together in the analysis. Note that the ANOVA test does not tell us which sets are different or how they are different, only that they are different within a given certainty. It is bad form in statistics if $p > 0.05$ to "accept" the null hypothesis; instead we say that we "fail to reject the null hypothesis" and that "we have not yet found enough evidence to suggest a statistical difference." See *What is a p-value anyway?*.

3.1.2 Pearson Correlation Test

The Pearson correlation test computes an r value that tells us how closely a linear model describes the data set of $< x, y >$ ordered pairs that we are seeing. It is computed by computing the average of the product of the z-scores (related to the Gaussian distribution) over the data set. The closer r is to one, the closer a positive linear correlation fits the data set, the closer r is to negative one, the closer a negative correlation fits the data, and the closer r is to 0, the more we fail to reject the null hypothesis. The quantity r^2 is also an important indicator of fit called goodness-of-fit, which tells us how much of the variance in the data is explained by the linear model. The closer r^2 approaches one, the better the goodness-of-fit of the linear model to the data.

Note that, whereas ANOVA tells us whether the null hypothesis applies, the Pearson test tells us what kind of trend could possibly exist between the data.

3.2 Bayesian Analysis and Multilevel Modeling

A newer form of statistical analysis that is coming into use is Bayesian analysis. It is typically more computationally intense, but on a computer this is hardly an important factor. Whereas classical regression statistics looks at factors one-by-one and cannot deal with nested groupings very well, Bayesian analysis can look at all factors simultaneously, whether thay are grouped or not. Where classical regression tells us with what probability we can or cannot reject the null hypothesis of no effect, Bayesian analysis allows us to propose a model to fit the data, set constraints, and tells us to what degree we can believe our model is accurate. This makes the Bayesian method ideal for use in demographic analysis.

Our group has three books, *Data Analysis Using Regression and Multilevel or Heirarchical Models* by Gelman, *A Handbook of Statistical Analyses Using R* by Everitt, and *Bayesian Computation with R* by Albert. My forbear, Yifei Sun, also did more work with multilevel modeling than myself, and so I refer the reader to his folder on the dropbox for more information.

3.3 Using the Data

3.3.1 Reading the File

The data is sent as a .csv file, also known as a comma-separated-value file. In this file format, the fields are delimited into rows by carriage returns and separated into columns by commas. Beware: commas are also commonly used as elements of English grammar, but all commas will be read as delimiters. Delete or change any comma that is not a delimiter before reading the file. Fortunately, .csv files can be opened in microsoft excel directly. All excel files may also be saved in .csv format, if only one spreadsheet is used and no cells are combined. This is convenient because R can read .csv files to perform analysis on the data.

It is possible that future data may be sent as other file formats, for example tilde-separated or tab-separated value (.tsv) files. We have no application to correctly read these files, and they will be converted to .out files when downloaded. The solution is to save the file as a text .txt file and then open with excel. Excel will start up a wizard asking for what symbol to read as a delimiter, and allow you to save the file to .csv. Other than in this method it is almost never useful to open a .csv file in a word processor like notepad; use excel instead.

Initially the data sets were uncombined and so there exist files with names like "Epist1_68" within the "Initial Analysis" folder referring to the responses to question 68 within the E-1 task. Since then it has been found that combining all of the data into one really big .csv file is the most efficient way to store and use the data. If we desire to look at the data divided by separate groups (which we usually do) this can be accomplished within R with boolean constraints (see R-Scripts section and documentation within the R scripts) after the data has been imported. At the time of this writing, all of the raw data is stored within the file "28 JUN Analysis\Raw Data\ALL_OF_THE_DATA_06_27.csv." Contained within this file are students who did not release us to process their data. Such entries have NULL values for their demographics information such as gender. Another file, "ALL_OF_THE_DATA_06_27_nullGone.csv" has this data removed.

Excel has the useful ability to sort a file based off of a single column, under Data → Sort. It can also do a heirarchical sort by one column, and then within that sort by another column, etc. Use this often.

3.3.2 Getting the Data

The data is stored in a relational database at RIT. Elements of the database can be queried and then converted to a .csv or .tsv file. At the time of this writing, our contacts are Nathan Popham (nap7276@rit.edu) and Matthew Koontz (mjk3979@rit.edu) at RIT; Matthew is the current primary contact who makes SQL queries from the database for us.

3.3.3 Understanding the Data

As mentioned before, the data comes from a relational database and is in .csv file format. When opening any of these files in excel, there should be several column names in the first row: question information like "Question_SysID," "Answer_Number," "QuarterWeek," and "Start_Date;" demographics information like "Student_SysID," "Course_SysID," "Gender," and "Year;" and class information like "PriorPhys," "PriorMath," "CurMath," "PriorMathName," "CurMathName," "Major," and "Major_Name."

These are the names of the columns in the file, which tell us certain information made for every observation, for example which student answered which question during which week of the 10-week quarter or 16-week semester, and how they answered it. Every student also has associated demographics information, for example the physics course they are enrolled in (1 ← Phys 1 at RIT, 2 ← Phys 2 at RIT, 3 ← Phys 3 at RIT, 38 ← Phys 1a at RIT, 39 ← Phys 2a at RIT, 40 ← Phys 1 at USMA), their gender (0 ← female, 1 ← male), their prior and current math class, their year in school, etc. This basic information is fairly intuitive.

Over the course of the summer I have also added other columns to the data file"ALL_OF_THE_DATA_06_27_nullGone.csv." to assist in analysis. These columns are: "Task_Type," "Task_Name," and "Axis_Name," which tell us which task each question is from and which axis it is supposed to probe (derived from "Question_SysID" and the surveys themselves); "Score_Number" and "Ratio" give the number of points awarded to the student's response and the ratio of the points awarded to the points possible, respectively (derived from "Answer_Number" and the answer key to the EBAPS, CLASS, and SLIDe surveys); "School_Name" (derived from "Course_SysID"). "Binary" is a unique column that has value 1 for a remedial class (38, 39) and for a student enrolled in Physics 3, but previously enrolled in Physics 1a or 2a; it has value 0 otherwise: this column separates students into those who are currently or have been previously enrolled in a remedial class and those that have not (derived from "Course_SysID" and "PriorPhys"). "Course_SysID_2" has the same values as "Course_SysID" except that it assigns 33 to values where 3 → "Course_SysID" and 1 → "Binary:" this effectively separates the physics 3 class into physics 3 and 3a classes, to parallel the 1 and 1a, 2 and 2a pairs (derived from "Course_SysID" and "Binary.") The "School_Name" and "College_Name" columns tell whether the student is from RIT or USMA, and the college that their major belongs to (derived from "Course_SysID" and the school websites). "Week_Study" and "Semester_SysID" tell us during which week out of the 50 weeks in the study the question was answered, and in which semester out of the 5 semesters in the study the question was answered (both derived from "Start_Date" and "QuarterWeek").

These column names are important as they allow an entire column of data to be imported into R using the column name as a tag.

4 R-Scripts

R has a command line from which certain operations can be performed, including running a script/ function. The first step when loading scripts should be to change the directory from which R is loading scripts to the directory in which the scripts are located (File → Change Directory), for example, I used "T:\RAWR\Tim's Folder\R-Scripts\Used Scripts." Then load the script into R by typing "`source(''script_name.R'')`" in the command line. If the script is edited, the changes should be saved, and then you will have to re-source the script in order to run the edited script.

Most of the scripts used in the data analysis are actually functions. Functions are scripts that take certain parameters as input and produce an output. The header line of every function is

```
function_name=function(input1,input2,...,inputN)
```

To use the function, save the script (usually as "function_name.R"), source the script and then type in the command line

```
function_name(input1,input2,...,inputN)
```

Many of the scripts are documented to describe how they function. If a line has a "#" before it, it is documentation and R will not read what is written on that line before the "#."

The syntax of a function is

```
#Documentation

Header Line

{
Function code and #Documentation
return(output)
}
```

Assuming the script is written correctly, and the inputs are correct, R will return the desired output to the function.

4.1 Non-Used Scripts

The directory "R\R-Scripts\Non-Used Scripts" is full of scripts that were ultimately not used in the final analysis, due to being incorrect or due to a more flexible script being written later.

4.1.1 Yifei Stuff

There is one group within the Non-Used Scripts folder, "Yifei Stuff," that my successor may find useful. This folder contains my attempt to create an R-script for Bayesian analysis in R using Yifei's previous research. Although these codes do not work, the reader may find their structure and this information useful.

Firstly, the scripts in here with "Yifei" as part of the name were modified from his files on the dropbox. My main modification was to change the names of the files being loaded into R, since Yifei's script loaded them from his T-drive which is unaccessible to anyone but him. My successor will need to do something similar after downloading the files from the dropbox, in retitling the filenames in the script to where they are actually located on the computer.

Secondly, the script “BayesEx.R” works. I took it from a website created by Gelman for users of his textbook; it should give the reader an idea of how the scripts should work.

4.1.2 Bugs

In the course of doing Bayesian analysis, R will need to call on another program, Bayesian inference Using Gibbs Sampling (Bugs). Download Bugs from http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml and follow the site’s directions to make Bugs compatible with R; it mostly involves downloading a few packages.

As far as I can understand it, the flow of a Bayesian analysis program should be

1. Read in a file for analysis, set variable names in R
2. Run the function “bugs” within the script
3. Bugs uses the variables set in R, and then follows the directions in a .bugs file, which must be in the same directory from which R is sourcing code, and must be referenced in the R script
4. Bugs opens, completes analysis according to the .bugs file
5. Bugs closes, result displayed in R

Getting these scripts to work will likely be an important next step in doing demographics analysis on the data. See chapter 16 in the Gelman textbook for more help in writing R and bugs scripts.

4.2 Used Scripts

The directory “R\R-Scripts\Used Scripts” contains the scripts that were in one way or another used in the analysis of the Early-Term Optimism data set. These scripts all contain documentation (any non-escaped # character in an R-script denotes commentary that R does not process when running the script), but a brief summary and list of input/output follows for these scripts.

4.2.1 “correlation.R”

This function performs a Pearson correlation test for Ratio as a function of QuarterWeek for a single-course, single-axis, single-quarter data set.

Inputs: .csv file, Course_SysID_2 for course of interest

Output: Pearson correlation test for Ratio=func(QuarterWeek)

4.2.2 “Demographics.R”

FIRSTLY IT IS MANDATORY THAT THE DATA SET IMPORTED INTO THIS FUNCTION IS SORTED IN EXCEL BY “Student_SysID”!!!!! The RAWR survey questions are each administered once to every student; hence there can be multiple observations in the data set per student. This function takes a sorted data set and eliminates all duplicate student data, then writing the new non-duplicate data to the file “O:\Demographics.R” for further analysis. The non-duplicated data set can then be used in demographics analysis.

Inputs: A STUDENT-SORTED .csv file;

Output: File “O:\Demographics.R,” a data set without duplicate students’ information

4.2.3 “PrePlotMatrix.R”

This function takes a single-course, single-axis data set and returns a matrix that may be ported into “plotresponse.R” to make a graph of Ratio vs. Time. This function’s only purpose is to serve as an intermediate step between the data set and “plotresponse.R”

Inputs: .csv file, Course_SysID_2 for course of interest

Output: A matrix that serves as input to “plotresponse.R”

4.2.4 “responseplot.R”

This function takes the output to the PrePlotMatrix function and uses it to create a plot of Ratio vs. Time. Please check the word document “R\R-Scripts\Used Scripts\Graphing code.docx” for a template of how to make the plots. Generally the graphs are of Ratio vs. Week_Study and contain two plots that comprise a course pair (for example physics 1 and 1a appear on the same plot). In the response.plot function addme=T in order to add a line to the plot instead of starting a new plot. Also note that whereas the filename to use as source code is “responseplot.R” the function name is response.plot().

This function also requires that we load several packages in R’s memory: lattice, binom, bitops, caTools, gdata, gtools, grid, KernSmooth, gplots, psych, GPArotation, MASS, stats. These are located under Packages → Load Package, and must be reloaded every time R is closed. The first time the packages are loaded, they will likey need to be installed with Packages → Install Packages. Every time R is closed and opened, these packages will have to be loaded again before the script will work

Input: PrePlotMatrix output matrix

Output: Plot of Ratio vs. Time for one course-pair

4.2.5 “statistics.R”

This function does an ANOVA test on Ratio over some specified factor throughout some specified data set. The specifications are made within the script itself, so this is a very flexible function. Specifications of data set are made by parsing the original set of all data by enforcing boolean constraints (see source code). FACTORS MUST BE LABELED AS SUCH IF THEY ARE CODED NUMERICALLY!!! (see source code)

Inputs: .csv file

Output: Summary of ANOVA test

4.2.6 “StudentCounter.R”

Like statistics.R, this script is also fairly flexible in terms of the specifications for the data set imported. Given a specified data set, the function simply removes the duplicate students and counts the remaining students. This function is useful for quickly comparing distinct-student population sizes of different groups.

Input: .csv file *Output*: Number of distinct students in data set

4.2.7 “TableMaker.R”

This function imports a data set WITH NO DUPLICATE STUDENTS!!! (i.e. it should use a data set produced by “Demographics.R”) and creates a table that shows how a certain demographic is divided by the remedial and non-remedial type RIT classes.

Input: Datafile with no duplicate students *Output*: File “O:\tables.csv” containing a table of demographics that can be cut and pasted elsewhere

4.2.8 “UnpooledRegression.R”

I created this script (note: script, not function) in an attempt to do a demographics analysis despite the Bayesian analysis code failing to work. The function does an unpooled factor regression of the response ratio as a function of gender subject to a factor of the user’s choosing, and returns the value of each factor’s influence and its standard error. This script uses classical unpooled analysis methods and is probably not as useful as a Bayesian analysis would be. See source code.

4.2.9 Common Errors

These errors have been made commonly enough by myself that I thought it useful to include a section for correcting them:

- If a function is not performing as expected, first check that the correct file is being read in. If the file being read has been edited, these changes must be saved, and then the file can be correctly read in.

- Every time a function is edited, the script must be saved-over and the script must be newly sourced; otherwise R uses the version of the function that was last sourced.

- When editing the scripts, the name of a column within R must be “datafile$column_name,” not just “column_name.” “column_name” will be an unrecognized object to R.

- A single equals sign (=) is the assign operator. A double equals sign (==) is the boolean “is equal to” operator.

- Check spelling of all things being processed by R, e.g. column names like “E_one.” “E-one” is a different, undefined object that will result in an error.

5 Data Analysis

5.1 Initial Analysis

A lot of the initial analysis resulted from experimentation with using the R-scripts to analyze the data. During the initial analysis I did not have as good an understanding of what an ANOVA test was used for in terms of binning the data, so this analysis tended to be separated by question and binned by courses. It was later found that this binning choice did not really offer a meaningful interpretation of the data. Later analysis did the opposite, binning all questions within the tasks together, and separating by course. This also was not a useful binning option. After a couple weeks of experimentation with how to separate the data set, and of ensuring that the whole data set had been sent, a separation by Axis_Name and Course_SysID but binned by QuarterWeek was tried. This seemed to offer some preliminary results, that

- The epistemological sophistication did not seem to vary much at all throughout the course
- The classes classified as remedial seemed to have more volatile trends
- There also appeared to be some significant decreases in the score ratio around weeks 2, 6, and 8 on some axes, which may have implied testing effects

It was these effects that would be investigated in the later analysis. This initial analysis is under the folder "Initial Analysis" on the dropbox.

5.2 RIT Epistemology Task Analysis

5.2.1 Summary

Much of this section is taken from the Data Analysis section of the PERC paper I wrote during the summer over the effects above. This data set was the largest, and therefore offered the strongest significance in its conclusions. The analysis follows three main sections: first I did anova tests to determine which data sets may be binned together into a single population, second I ran Pearson correlation tests for each axis each course each semester and visually confirm with graphical analysis, last I ran a demographics analysis on how remedial and non-remedial classes differ.

This analysis is under the folder "28 JUN Analysis."

5.2.2 ANOVA Analysis

This analysis is under "28 JUN Analysis\ANOVA_Testing_06_27.xlsx."

I conducted a single-tailed analysis of variance test (ANOVA) to determine if differences in the means of the ratio of expert responses separated across other parameters were significant. It is common practice that statistically distinct data sets should not be binned together. The ANOVA test found there to be significant differences in data sets from different task subsets (Epistemology 1 or 2) with $p < .001$ of belonging to the same population. Similarly, the anovas concluded that the epistemological axis of the question, the physics course of the student (i.e. Physics 1 or 2), the week of the quarter in which the task was first available, and the course remedial status all comprised different populations with $p < .001$ in all cases. I therefore separated the data set by the epistemological axes and physics courses which contained those separations. Note that these groups have a very high level of statistical difference: in most cases p is approximately 2×10^{-16}, the lowest probability R can compute.

With the data set separated correctly I could begin to test the main hypothesis that epistemological sophistication tends to decrease during the semester. A decreasing trend in epistemological sophistication independent of quarter (I expected the epistemological sophistication to decrease in the same way regardless of which course was being taught or when it was taught) implied that the week of the quarter was the largest determining factor, and that the effect should be invariant between different quarters. This meant that the interaction effect between the quarter week and quarter should be small. To test this implication I performed a double-tail ANOVA test on each of the 24 Axis-Course pairs. Out of the 24 pairs, 19 had a significant interaction effect with $p < 0.1$, and so I additionally separated the responses by quarter. Below is the summary of the anova tests for ratio as a function of quarter week and term quarter.

This last separation also caused the average sample size to decrease from about 300 to 50 responses for each Axis-Course pair. Despite this, I assumed a well-defined normal distribution for each sample and continued analysis.

Although I had determined that the epistemological development trends differed between quarters, this did not exclude the possibility that within every quarter the response fractions generally decreased. So I ran a second double-tail ANOVA on the quarter-sorted data to verify that the response fractions varied with the week of the study, and also to further investigate the effect of course remedial status on epistemological sophistication. Out of 12 (remedial and non-remedial combined) Axis-Course pairs, nine pairs showed a significant ($p < 0.01$) and six pairs showed a very significant ($p < 0.001$) interaction effect between the week of the study and the course's remedial status. This confirmed that each of these factors, quarter week and remedial status (Binary), separately had a significant main effect on the expert response ratio.

5.2.3 Pearson Correlation Tests

This analysis is in "28 JUN Analysis\ANOVA_Testing_06_27.xlsx."

I next determined the nature of the variation within each quarter of the response fractions with quarter week. A negatively-correlated linear fit seemed the best model to test for since previous studies suggested a decrease in epistemological sophistication between the beginning and end of a course. I calculated the Pearson product-moment correlation coefficient r for each quarter. Out of 120 Axis-Course-Quarter triplets, 87 triplets had small correlation values of $|r| < 0.1$; furthermore, for 48 of the 120 triplets $r \geq 0$, suggesting that there was not a negative correlation during those quarters, and the decreasing trend did not hold for every quarter as anticipated.

5.2.4 Graphical Analysis

Despite this, it was still possible that some meaningful (although not generally decreasing) variation existed between the epistemologies and the quarter week, for example pre/post testing effects. A qualitative graphical analysis suffices to decide this. There are twelve different data sets when separated by the schema set forth in the ANOVA analysis above: 4 epistemological axes times six separate courses divided by two courses (one remedial and one non-remedial) per set.

The twelve Ratio vs. Week graphs for RIT can be found in "28 JUN Analysis\Epistemology."

From these graphs observe three points. First, the general trend of response fractions within the quarter varies by quarter; there is no trend of decrease through every quarter. This suggests that introductory classes are not damaging student epistemological development every quarter. Second, there is no periodic variation in the response fraction with the week of the quarter across the quarters; the minima and maxima occur at different times within each quarter. Since testing in all of these classes is closely aligned, this suggests that test effects are not the only major contributor to within-quarter epistemology changes. Third, for each graph the remedial classes appear be more dynamic than the corresponding non-remedial class; the response fraction has more changes and more extreme changes in the expert-response fraction. This together with the previous ANOVA test finding significant differences in the populations made up of remedial and non-remedial classes, suggests that being enrolled in a lower-level physics course correlates with a more volatile epistemological sophistication.

5.2.5 Demographics Analysis

The last observation prompts an investigation into why there should be differences in response fraction trends between courses that differ only by remedial status. Given that the courses have no significant instructional differences, I next checked for effects due to demographic compositions. Although I thought it likely that the demographics would differ by factors like gender and major, we instead found more sigificant differences according to the students' prior physics class and prior and current math class.

Demographics histograms can be found in the different sheets of "28 JUN Analysis\Epistemology\Demographics\RIT_Demographics_Table_EP.csv."

These histograms are created by dividing the students into remedial and non-remedial groups, counting up the number of students in each group belonging to some other group within, and dividing by the total stdents in each group to obtain a "normalized representation" of the groups within the classes. There is very little apparent difference between the two classes for factors like gender and major. Instead observe that the most significant difference is that a large majority of non-remedial students were previously enrolled in "project-based" calculus, and similarly for remedial physics and non-project based calculus. The demographic composition between the two class types on the other math courses (Differential Equations, for example) are much more similar.

5.2.6 Implications

The analysis indicates that there is no correlation between expert-like response ratios and time as a quarter at RIT progresses, at any rate not for ther Epistemology task. The data does not appear to have any periodic variation from quarter to quarter, which suggests there is no trend (decreasing or otherwise) that occurs during every semester. The decreasing trend is not supported by the data: the linear correlation fit is much too low to conclude anything other than the null hypothesis. Furthermore, the data is noisy within each quarter, and the probability of changes in epistemological sophistication resulting from testing effects or any other known within-semester effects is low. The variations also can't have anything to do with the physics topic being taught at the time, since these variations are observed for the all the different curricula of physics 1, 2 and 3.

We also find an effect in that the remedial and non-remedial students make up different populations. This is shown graphically in how remedial students' expert-like response ratio fluctuates more widely and more often. A demographics analysis shows that this may possibly be correlated to which prior and current math students have taken.

5.2.7 Next Steps

The analysis resulted in results that are well worth a deeper exploration.

As mentioned above, separation by axis name and course resulted in a decrease of sample sizes to 50. Some sources say a good general rule of thumb for analysis is a sample size of 100, greater than ours. To make the results much stronger, there would need to be more participants in the study. Performing the study at a state school might well give us the larger sample sizes we need to verify the results.

The study also suggests that it is possible that there is a correlation between how expert-like a student's epistemological beliefs are, and the kind of math experience they have. This could be interesting to explore further. Since this study was done over summer. it was difficult to get in touch with teachers to obtain schedules for the math classes. A successor could try to find differences between how project-based and non-project based calculus courses at RIT are taught. If there are differences, those differences could be used, perhaps in an interview format, to determine if there is a likely correlation between math experience and epistemological sophistication.

5.3 Identity Task Analysis

5.3.1 Summary

After a visual confirmation graphical analysis I attempted a Bayesian demographics analysis. Having technical difficulties with this, I tried a classical unpooled analysis with the UnpooledRegression.R script.

5.3.2 Graphical Analysis

Both the RIT and USMA data for both the Identity 1 and 2 tasks were shown graphically to be extremely invariant with time. The PERC paper I submitted had more than enough to handle in just describing the RIT data for the Epistemology tasks, and so this Identity data was excluded from the paper. Furthermore, the data is so unchanging with time that no analysis was really needed to confirm its stasis. We decided to bin the data together across time since it was unchanging and do a demographics analysis to see if any of the subgroups were scoring significantly better than any of the others.

The reader can also verify the static behavior of the data over time: the graphs are located in "28 JUN Analysis\Identity"

5.3.3 Bayesian Analysis

Using Yifei's notes and code on dropbox and the example code from the internet, I attempted to modify and merge the two to create a script that could run a Bayesian demographics analysis on the data. As described above, my attempts to create my own code or to use my predecessor's were largely unsuccessful. When I tried to run the code, the model would initialize in bugs, but I would always get an error that "DICset() is greyed out" and the program would get stuck until I forcibly closed out of Bugs.

5.3.4 Classical Unpooled Regression for Identity 1: RIT

This unpooled regression analysis can be found in my folder on the dropbox, "28 JUN Analysis\Identity\Response ratio by demographics analysis.csv".

Failing being able to make the Bayesian Analysis scripts operational, I decided to do a simpler classical factor regression on the Identity demographics for RIT. I ran the UnpooledRegression.R script several times using the demographic categories College_Name, PriorMathName, CurrentMathName, PriorPhysName, thinking these to be most likely to cause any discrepancies in response fraction scores. The test found that to the extent to which the model was valid, certain factors in each area caused scores to be more or less above the mean. Unfortunately, the correlation coefficient for the models depending on each of these factors was exceedingly low, $r^2 \leq 0.01$.

At this point I changed my documentation to focus on the r^2 value of the unpooled models to try to find some factors that were more influential than $r^2 = 0.01$, testing all separation categories instead of just the demographics ones. I finally found significant factor correlation between the response fraction and the Student_SysID with $r^2 = 0.08$, Question_SysID with $r^2 = 0.29$ and Axis_Name with $r^2 = 0.05$. This still leaves 60% of the variance unexplained, owing to random between-observations fluctuations. A random variance of this high is extremely unlikely given our large sample sizes. We thus abandoned this line of statistical analysis, figuring that something was probably wrong in the statistical unpooled model.

5.3.5 Implications

The analysis suggests that either there are no statistically significant differences between the different demographics groups that take the Identity task on RAWR, or alternatively, that something about an unpooled regression analysis makes it inappropriate to use in analyzing our data set.

5.3.6 Next Step

The next step would most likely be to try to get an operational Bayesian analysis script to runon the data to see if it confirms or conflicts with the null result given by the classical unpooled analysis.