

DEVELOPING AN INSTRUMENT TO MEASURE STUDENTS'
UNDERSTANDING OF THE NATURE OF SCIENCE

by

Nathan Porter

A senior thesis submitted to the faculty of

Ithaca College

in partial fulfillment of the requirements for the degree of

Bachelor of Science

Department of Physics

Ithaca College

May 2011

Copyright © 2011 Nathan Porter

All Rights Reserved

ITHACA COLLEGE

DEPARTMENT APPROVAL

of a senior thesis submitted by

Nathan Porter

This thesis has been reviewed by the senior thesis committee and the department chair and has been found to be satisfactory.

Date

Dr. Matthew Price, Advisor

Date

Dr. Dan Briotta, Senior Thesis Committee Member

Date

Dr. Matthew C. Sullivan, Senior Thesis Instructor

Date

Dr. Beth Ellen Clark Joseph, Chair

ABSTRACT

DEVELOPING AN INSTRUMENT TO MEASURE STUDENTS' UNDERSTANDING OF THE NATURE OF SCIENCE

Nathan Porter

Department of Physics

Bachelor of Science

An essential part of student learning in science is for students to become scientifically literate. An important aspect of scientific literacy is for students to learn and understand the ideas and concepts behind the nature of science (NOS). NOS can be defined as science as a way of knowing, the values and beliefs as a part of scientific knowledge and development, and the processes and products of science. To increase education opportunities in learning about NOS we need to first be able to reliably and accurately measure students understanding of NOS. Being able to measure students understanding of NOS will allow for assessment of teaching practices in the sciences leading to better methods and techniques. We want to measure students understanding of NOS. To do this we are constructing an instrument to measure the students understanding of NOS. In constructing this instrument we first have developed a large amount of NOS questions that we want to narrow down. Narrowing

down the questions involves piloting the instrument with the full amount of questions. After writing the questions we sent them to professionals to review. We only received 4 responses from the professionals. Therefore, we were unable to do statistical analysis to narrow down the questions. Using the responses from the 4 professionals we did a case study analysis and narrowed down the question list from 61 questions to 39 questions.

ACKNOWLEDGMENTS

I first want to thank Matt Price and Michael Rogers for all of their help while working on this project. I would also like to thank Matthew C. Sullivan for dealing with me and my witty humor throughout my 4 years at Ithaca College. A big thank you goes to my parents for supporting me for my 4 years at Ithaca College. I need to thank Kevin Hurley for keeping Sarah and I in line inside and outside of the classroom. Finally, I want to thank Sarah Burleson for her smile and support. Oh, and Sarah, I won the pencil game!

Contents

Table of Contents	vii
List of Figures	ix
1 Introduction	1
1.1 What is the Nature of Science?	3
1.2 Instrument Design	6
2 Theory	7
2.1 NOS Instruments	8
2.1.1 MPEX, EBAPS, and CLASS	8
2.1.2 VOSTS and VASS	10
2.1.3 VNOS	11
2.2 Developing a Trusted Instrument	12
2.2.1 Validity	12
2.2.2 Reliability	14
2.3 Looking at Favorable and Unfavorable Responses	18
3 The Instrument	21
3.1 Developing a New Instrument	22
3.1.1 Writing the Individual Questions	22
3.1.2 Putting Together the Preliminary Instrument	23
3.2 Giving the Preliminary Questions to Professionals	24
3.2.1 What is asked of the Professionals	25
4 Analysis	27
4.1 Professional Quantitative Analysis	27
4.2 Professional Qualitative Analysis	28
4.3 Who Responded to the Instrument	29
4.4 Case Study Analysis	30
4.4.1 Theories and Laws	35
4.4.2 Creativity	36
4.4.3 Scientific Method	36
4.4.4 Society	36

4.4.5	Uncertainty	38
4.4.6	Subjective and Tentative Nature of Science	38
5	Conclusions	39
5.1	Future Work	40
	Bibliography	41
	Appendix A NOS Survey	45
	Appendix B Consent Form	51
	Appendix C Sample Survey	55

List of Figures

2.1	A plot showing how students change from pretest to post test.	20
4.1	Example histograms for case study analysis.	32

Chapter 1

Introduction

An essential part of student learning in sciences is scientific literacy. Scientific literacy is important for students to grasp not so that they can become scientists, but so that they can make educated judgements about science. Preparing scientifically literate students has been increasingly advocated by science educators and major scientific education organizations such as the American Association for the Advancement of Science (AAAS) [1] and the National Research Council (NRC) [1]. The NRC states, “Scientific literacy enables people to use scientific principles and processes in making personal decisions and to participate in the discussions of scientific issues that affect society. A sound grounding in science strengthens many of the skills that people use every day, like solving problems creatively, thinking critically, working cooperatively in teams, using technology effectively, and valuing life-long learning.” [2]

Although there is not a single agreed-upon definition of the nature of science (NOS), it generally refers to the values and assumptions existing in scientific knowledge [3]. Using this idea, NOS is most often connected to “science as a way of knowing and the values inherent in the development of scientific knowledge.” [3]

There are many aspects of NOS, but there are an agreed set of foci: “(a) Scientific

knowledge is tentative (subject to change), (b) empirically based (based on and/or derived from observations of the natural world), (c) subjective (theory laden), (d) necessarily involves human inference (subjective), imagination, and creativity (involves the invention of explanations), (e) necessarily involves a combination of observations and inferences, and (f) is socially and culturally embedded.” [4] Two other main ideas behind NOS are the differences between observation and inference, and the differences between scientific theories and laws.

A large issue is measuring students’ understanding of NOS. There have been attempts in the past such as: The Maryland Physics Expectations Survey (MPEX) developed by the Edward F. Redish, Jeffery M. Saul, and Richard N. Steinberg is a survey asking subjects how they view their beliefs about themselves as learners in physics [5]. The Epistemological Beliefs Assessment for Physical Science (EBAPS) developed by Barbara White, Andrew Elby, John Frederiksen, and Christina Schwarz measures students’ epistemology using an agree/disagree scale [6]. The Views on Science-Technology-Society (VOSTS) developed by Glen Aikenhead, Alan Ryan, and Reg Fleming measures how students view the social nature of science and how science is conducted [9]. The Views about Science Survey (VASS) developed by Ibrahim Haloun and David Hestenes surveys subjects about their views on learning and knowing science and how it connects to an understanding of science and course achievement [7]. The Colorado Learning Attitudes about Science Survey (CLASS) developed by W. K. Adams, K. K. Perkins, N. S. Podolefsky, M. Dubson, N. D. Finkelstein, and C. E. Wieman measures students beliefs about physics and their beliefs about learning physics [8]. The Views on the Nature of Science (VNOS) developed by Norman Lederman, Fouad Abd-El-Khalick, Randy Bell, and Renee S. Schwartz is a questionnaire that allows subjects to give written responses to NOS questions.

All of these instruments have their own strengths and weaknesses, but it is because

of their weaknesses that we need to develop a new instrument. The weakness with the MPEX, EBAPS, and CLASS is that they all have some questions related to NOS, but they are more focused on personal epistemology. They are more about the subjects' beliefs, values, and attitudes about themselves as learners. We are looking for the larger picture of science, the process and products of science as a whole. When we look at the VOSTS and the VASS we realize that they are more intended for the classroom and are not broad based like we would like. The VOSTS and VASS have also not been well researched in the design phase of the instrument. This brings us to the VNOS. The VNOS is a great NOS instrument that measures a lot of what we want to measure. The problem with the VNOS is that in college science general education classes that the class size is usually no smaller than fifty students. The VNOS, because it asks for written responses to questions about NOS, takes a very long time to analyze. It took me personally ten weeks to code and analyze the data from about 150 students in an astronomy class alone.

1.1 What is the Nature of Science?

NOS education research has been around for over 100 years. NOS and inquiry teaching started to develop in the middle of the nineteenth century [10]. Scientists began to argue that science was different than the other school subjects because it involved inductive logic. This was different than mathematics and grammar because they both involved clear logic rules. On the other hand science involved detailed observations, which then lead to general principles. This difference lead to students having to learn how to make observations and from these observations use inductive reasoning to draw conclusions.

A good place to start is to first ask: “what is science?” The simplest answer to this

is that science is a body of knowledge, meaning that it is both a method or process and a way of knowing [11]. Some have attempted to codify NOS all models are different but they have similarities. Important topics in NOS are: inference vs. observation, theory and law, cultural impacts on science, scientific knowledge changing due to new technologies, science biased by personal beliefs, and creativity in science. NOS is based on individuals' values and beliefs surrounding scientific knowledge.

One of the first ideas behind NOS is understanding the difference between an inference and an observation. An observation is something that can be directly detected by the senses that can be easily confirmed by several other observers. Observations are simply descriptions of an event explaining it exactly as it appears. An inference is something that is more than just something viewed with the senses. An inference may be used to explain an observation. For example, you wake up in the morning and walk outside. You observe that there are puddles on the ground and that the sky is gray. You can make an inference based on these observations that it may have rained while you were sleeping.

Another issue, which is closely related to the idea of the difference between observations and inferences, is the difference between scientific laws and theories. The first thing that comes to mind about scientific laws and theories is that there is a hierarchical relation between them. Most think that scientific laws are "better" or more important than scientific theories. Most also think that theories, if "proven" become laws. Both of these ideas are incorrect; theories and laws are different types of knowledge. It is important to realize that theories and laws are both equally important in the scientific world. A scientific law is a "statement or description of the relationships among observable phenomena" where a scientific theory is "inferred explanation for observable phenomena" [11]. We can see that laws are more closely related to observation while theories are closely related to inferences. A great example described

by Lederman uses Boyle's law to clarify the difference. Boyle's law explains how pressure of a gas relates to its volume at a constant temperature. This is clearly a description of a relationship about something that is observable. It simply says what happens when the volume or pressure of a gas changes, but does not explain why. The kinetic molecular theory, on the other hand, provides an inferred explanation of what is happening in Boyle's law [11].

The next discussion is about scientific knowledge. As expected, scientific knowledge is derived from observations about the natural world, but observations are not the entirety of scientific knowledge. Scientific knowledge also involves imagination and creativity. This means that scientists need to use their creativity to imagine or invent their inferences and explanations of observations.

Scientific knowledge is subjective and/or theory-laden [11]. Theory-laden science means that all of science makes the assumption that it depends on previous theories for new science. Everything in a scientist's past forms a way of thinking that affects the problems and experiments that he or she conducts and investigates. All past work on theories, their beliefs, previous knowledge and education, their experiences, everything in the scientist's past affects their future work. Observations and inferences are usually made to solve a problem. To different scientists there may be different avenues to solve this specific problem, and because scientists have different backgrounds and different motivations to solve the problem. This leads to different mindsets and ways of thinking about a problem, which is how scientists can come to different conclusions and discoveries based on the same observations. Along with past scientific experience, a scientist's culture affects what he or she does. Everyone is a product of their culture, so various products of culture such as social normalities, politics, economics, and religion, all change how different scientists think and act. As a result, science in general is affected by cultural factors.

Scientific knowledge is never certain or absolute: all scientific knowledge, including laws, theories, facts, etc., all of it is subject to change if new evidence arises. New evidence can come from advances in new theories and advances in technology. Scientific knowledge is always changing because of the many different elements that create it. It is made up with observations, inferences, creativity, and cultural and personal backgrounds. With all of these aspects and new evidence through new technology and theories scientific knowledge is always subject to change.

1.2 Instrument Design

The first question we have to ask is “How do we measure understanding of the Nature of Science?” To accomplish this we developed an instrument that is a multiple choice questionnaire asking the subject to give a letter grade to statement about NOS topics. The subject group that we are targeting in our questionnaire are college students who are non-science majors participating in general education science classes. This questionnaire makes strong statements about NOS topics such as, “Societal values and expectations determine what science is conducted and accepted.”

Chapter 2

Theory

Our main goal when developing this instrument is to be able to effectively and accurately measure students' understanding of NOS, which will allow us to better instruct students on the concepts and ideas of NOS. As we have previously mentioned, there have been several attempts to design instruments to measure NOS. These instruments have brought good ways to measure NOS, but they have not achieved measuring NOS in the way that we would like to. The previous instruments have had many good qualities, but there are cons to each instrument that we would like to point out and improve upon.

In the upcoming sections we will describe these previous instruments. Then we will outline how we would like to improve upon the instruments. Next, we will discuss how to ensure that our instrument has both validity and reliability. This will ensure that the instrument is testing what we have written it to test, and that it can provide the same results over and over again. Finally, we finish with some of the other ideas behind our instrument such as looking at unfavorable responses.

2.1 NOS Instruments

To develop our instrument to measure students' understanding of NOS we need to first look at instruments that have already been developed. All of these instruments have their own advantages and disadvantages. We want to build on their advantages and make sure we steer clear of the disadvantages. In the following sections we will describe what some of the previous instruments set out to do and what we can learn from them.

2.1.1 MPEX, EBAPS, and CLASS

The first three instruments that we want to discuss are the MPEX, EBAPS, and the CLASS. These three instruments are well researched and have been put together thoughtfully. The Maryland Physics Expectations Survey (MPEX) is a 34-item Likert-scale (5 item agree-disagree scale) survey that asks students about their attitudes, beliefs and assumptions about physics. The survey typically takes 30-40 minutes for most people to complete. One of the pros about the MPEX is that it is a Likert-scale survey, so it is easy to code and analyze. Another valuable item about the MPEX is that it is very well written and tested. It was tested for over four years in more than 15 universities and colleges. The creators also validated that the students understood what they were being asked with over 100 hours of videotaped student interviews. In the conducted interviews students were asked to describe their interpretations of the statements on the survey and to explain why they answered the way they answered the survey. They also gave the survey to five different calibration groups to make sure what they defined as 'expert' or 'novice' responses were correct. We can use many of these pros to implement into our instrument to make it the best possible way to measure students understanding of NOS.

The Epistemological Beliefs Assessment for Physical Science (EBAPS) is another Likert-scale survey but it has 30 items compared to the 34 of the MPEX. The EBAPS typically takes 15-22 minutes to complete. The EBAPS is aimed at high school and college students taking introductory physics, chemistry or physical science. It is best used in algebra-based courses. The EBAPS is intended to focus on epistemology of students. It also looks into the students' epistemological knowledge to see how that may affect the student's learning behavior. Just like the MPEX, the EBAPS is easy to code and analyze because it is a Likert-scale survey, and we can use many of its good qualities to adapt to our instrument.

The last instrument we need to discuss in this sections in the Colorado Learning Attitudes about Science Survey (CLASS). The CLASS, just like the two previous instruments, is a Likert-scale survey, but has 42 items compared to 34 and 30. The goal of the CLASS is to measure students beliefs about physics and learning physics. Like the MPEX, the CLASS uses interviews, reliability studies, and extensive statistical analysis to validate the survey. The CLASS was written to try to improve upon the MPEX. The CLASS is intended to address a wider variety of issues that instructors consider important to learning physics, and it also makes clearer, more concise statements so that the survey can be completed in ten minutes or less. The CLASS took some of the ideas of the MPEX and implemented it into its instrument, as we intend to do with our instrument.

As shown above, all of these instruments are important and useful tools in their own sense. We want to use some of the ideas and approaches that they use with our instrument. However, these three instruments are mainly about personal epistemology. They are targeting the students' beliefs about themselves as learners in science. They are trying to find the students' beliefs, values, and attitudes towards science. An example from the CLASS is: "After I study a topic in physics and feel that I un-

derstand it, I have difficulty solving problems on the same topic” [8]. This is asking the students’ opinions on how they learn the material. In contrast, we are trying to measure the students’ understanding of NOS. This is asking more of what a scientist is and what a scientist does. We are trying to measure the students’ understanding of the process and products of science. What is scientific process, and what does it produce? This is the main idea that we want to see if students understand.

2.1.2 VOSTS and VASS

The Views on Science-Technology-Society (VOSTS) is a group of 114 multiple choice questions that involve the topics of science, technology, and society. It was developed over six years with 11th and 12th grade students in Canada. The VOSTS was developed because the instruments before it assumed that all students either agreed or disagreed with scientists about a certain question. The VOSTS points out that a student could answer these questions but not know what the question is actually asking. Therefore, the students response was graded either correct or incorrect although the student did not understand the question. The authors of the VOSTS asked students to write about the questions before they turned them into multiple choice questions, ensuring that the students understood what the question was asking about first.

The Views about Science Survey (VASS) is constructed of 33 statements. Each statement is followed by two contrasting alternatives and then the respondent selects one of eight options. The VASS was created to assess student views about knowing and learning science to see how these views relate to the students views of achievement in science courses [7]. The VASS is based for high school classroom use and are targeting non-science major undergraduates.

The VOSTS and VASS do ask NOS questions similar to what we would like, but do not always target the questions that we want to ask. Another problem with these

two instruments is The VOSTS and VASS were developed to be used in just the classroom setting. We want our instrument to be more broad based, meaning that our instrument can work both in and outside of classroom settings. In addition, these two instruments are older and much more research has come out since their creation in both testing procedures and instrument development. Finally, these instruments have not been well researched in design. This is in stark contrast to the MPEX, EBAPS, and CLASS, all of which explain exactly how they came up with the statements and questions that they use, and the revision process for creating these statements and questions.

2.1.3 VNOS

The Views on the Nature of Science (VNOS) is an instrument that targets exactly what we want to measure. The VNOS “aims to provide meaningful assessment of learners’ NOS views” [3]. The VNOS is an opened ended instrument. It asks students questions and has them write out their answers. The students are also asked to justify and explain their answers. Our concern with the VNOS is that we are trying to measure the students’ understanding of NOS in a general education class. Most of these science classes are very large, usually with a minimum of 50 students. Because the VNOS has each participant given written answers to its 7 questions, it gathers an enormous amount of data, requiring very long coding and analysis. We hope to follow the ideas and some of the questions in the VNOS, but alter them so they can be answered on a Likert-scale. This will allow for much quicker data analysis to measure the understanding of a large class.

2.2 Developing a Trusted Instrument

One of the more important parts of writing an instrument is proving that the instrument can be trusted. To make sure that an instrument can be trusted, one must determine its validity and reliability. If an instrument is both valid and reliable it means that it measures what it was developed to measure, and that it can continue to produce the same results.

2.2.1 Validity

When creating an instrument, one of the first things that has to be looked into is making sure that this instrument actually measures what it is intended to measure. This is referred to the instrument's validity. The more valid that the instrument is, the more likely that the instrument is providing the desired measurement. The validity of an instrument is not either valid or invalid, but validity exists along a continuum, from high to low, in varying degrees [12]. Another important aspect of validity to consider is that neither the set of items, or the questions that are asked in the instrument, or the scores from the set of items are valid or invalid. It is the interpretation from the author that is valid or invalid [13]. There are many different ways to determine validity, but during this process we need to remember, does the instrument measure what it was created to measure?

The first type of validity that we need to look into for our instrument is construct validity. Construct validity is when both the authors and respondents of the instrument interpret the construct of the instrument the same way [12]. In our case, the construct is NOS. This is difficult to show because the goal of our instrument is to measure this understanding of the construct. To test construct validity for this instrument we first have to administer it to professionals. The professionals that we

will give the instrument to will be subjects that we know have a high understanding of both NOS and science. They will be either college science professors or scientists that have an educational research background. When administering the instrument to students testing construct validity will consist of ensuring that they understand what each question is asking. If the students understand what each item is asking, the instrument will successfully measure their understanding of NOS.

Next, we want to look into the content validity of the instrument. Content validity means that the content of the test reflects appropriately upon the important aspects of the content [13]. Again, this means that the content of our test reflects upon the understanding of NOS. There cannot be irrelevant content in the instrument. An example of an irrelevant question on our instrument could be something along the lines of, “Darwin was highly contended when he developed the theory of evolution,” instead of “Darwin’s theory of evolution is unimportant because it is just a theory, and theories are not proven.” The first statement is asking about Darwin and the history of his theory, while the second statement is asking more about theories and students’ thoughts about theories and what a theory means. we also want to ensure that none of the content in the instrument is underrepresented. To combat this we started by developing a list of 60 items for the instrument. These items are separated into 6 groups of 10 questions. The 6 groups represented are: theory vs. law, creativity in science, the myth of the scientific method, the social and cultural pressures of science, uncertainty measurements in science, and the subjective and tentative nature of science. These 60 questions will be narrowed down after pretesting with professionals, making sure that each grouping is equally represented.

Finally we want to consider face validity, or how closely related the content of the instrument looks like the construct to a nonprofessional. Face validity has no direct bearing on the empirical or theoretical quality of the test because it is based

on the nonprofessionals' opinions [13]. Despite this, it is an important part of the validity of the instrument. This is because if the test taker is reading through the test and at face value does not think that the test reflects upon the construct of the test, then they will take the test differently. The test taker may answer randomly or simply guess at our instrument because they do not believe the test as a whole and the individual items have a connection. To make sure there is face validity it is important to make sure that the test takers know what they items on the test are asking and to make sure that they believe that it connects to the construct of NOS understanding.

2.2.2 Reliability

Now that the proper steps have been made to ensure that the instrument measures what it is intended to measure, the next step is to make sure that it produces the same results over and over again. Reliability can be defined as the extent to which an instrument produces the same information at a given time or over a certain period of time [12]. Another definition of reliability is given as the extent to which differences in respondents' observed scores are consistent with differences in their true scores [13]. This is saying that reliability comes from an observed score, a true score, and an error measurement. An observed score is what is actually measured. An example of this is when a student measures the length of a a certain distance. The true score is the actual length of that certain distance. Ideally observed scores are good estimates of true scores. So reliability for a measurement comes from the differences in the respondents' observed scores on the measurement attributed to the differences in their true scores [13]. Because there is nothing that completely and entirely reliable, it is our job to make our instrument as reliable as possible.

The degree of reliability of the instrument depends on the differences in which

the test scores vary because of the individual, and the differences that the test scores vary because of measurement error. It is assumed that measurement error occurs as if it is random [13]. Due to this we can say two more things about measurement error. Generally, measurement error as an average cancels itself out. So over a large group of test subjects the inflating and deflating measurement errors will even out making an average measurement error that should be near zero. The next thing that comes out of the assumption that measurement error is random is that the error is uncorrelated with the respondents' true score.

We have said that reliability depends on the differences in the tests scores that vary because of the individual and the differences of the test scores that vary because of error measurement. Keeping this in mind we can represent this as

$$X_o = X_t + X_e \quad (2.1)$$

where X_o is the observed test score, X_t is the true test score, and X_e is the amount of error causing the difference between the observed and true scores. Because we have said that error is random it can change test scores which can cause inconsistencies among all of the respondents'. Another way to approach this is to look at the variances in the data, which in our case will be the answers to the individual questions we ask to the professionals. The variance for the errors scores can be calculated as

$$s_e^2 = \frac{\sum(X_e - \bar{X}_e)^2}{N} \quad (2.2)$$

this is where s_e^2 is the variance in the error, \bar{X}_e is the average error measurement, and N is the number of respondents'. The variance in the error is the degree to which error affected different people in different ways [13]. The higher the degree of variance represents poor reliability in the instrument. Eq. 2.2 can be adjusted slightly to find the variance of the observed test score and true test score. The total variance of the observed score is equal to the sum of the variance of the true score, the variance of

the error score, and the covariance of these terms. The covariance is the correlation of the individual variances. This term comes from Eq. 2.1, or the observed score which is the sum of the true score and error score terms. So the sum of the variances is:

$$s_o^2 = s_t^2 + s_e^2 + 2r_{te}s_t s_e. \quad (2.3)$$

Previously we have described that our error is independent of the true score. This makes the covariance term go to zero which leaves us with:

$$s_o^2 = s_t^2 + s_e^2. \quad (2.4)$$

One way to express reliability is to look at the proportion of the variances of the observed score and the true score. The proportion of observed score variance to the true score variance, also called the reliability coefficient, is given by:

$$R_{xx} = \frac{s_t^2}{s_o^2} \quad (2.5)$$

where R_{xx} is our reliability coefficient. These values will be between 0 and 1. The higher the R_{xx} the better the quality of the instrument. Generally, a reliability between .70 and .80 are satisfactory for most research purposes [13]. Another way to look at the reliability of an instrument is the lack of error measurement. The smaller the error in your measurement the better the reliability. To find the coefficient of reliability using error variance we first take Eq. 2.5 and substitute in Eq. 2.4 for s_t^2 in the numerator. This gives:

$$R_{xx} = \frac{s_o^2 - s_e^2}{s_o^2}. \quad (2.6)$$

Now we simplify this equation and get:

$$R_{xx} = 1 - \frac{s_e^2}{s_o^2} \quad (2.7)$$

Again, when R_{xx} will range from 0 to 1 with the better reliability being closer to 1. When there is a small degree or error variance, or the s_e^2/s_o^2 term is close to 0, it means that the respondents scores are only slightly varied by error measurement.

Just like there are two measurements for reliability with proportion measurements, there are two measurements of reliability with correlations. The first relationship is the squared correlation between observed scores and true scores. We start with the covariance term of the observed scores and the true scores:

$$c_{ot} = \frac{\sum(X_o - \bar{X}_o)(X_t - \bar{X}_t)}{N}. \quad (2.8)$$

We multiply out the nominator of Eq. 2.8 and get:

$$c_{ot} = \frac{\sum X_t^2 + \bar{X}_t^2 + X_e X_t - X_e \bar{X}_t - 2X_t \bar{X}_t}{N}. \quad (2.9)$$

We notice that the first part of the sum is the variance of the true score and the second half is another covariance term between the true scores and the measurement error scores. We again remember that there is no correlation between the error measurements and the true scores so the covariance term goes to zero. This gives us:

$$c_{ot} = s_t^2. \quad (2.10)$$

The correlation between observed scores and true scores is [13]:

$$r_{ot} = \frac{c_{ot}}{s_o s_t} \quad (2.11)$$

Substituting Eq. 2.10 into Eq. 2.11 gives:

$$r_{ot} = \frac{s_t^2}{s_o s_t}. \quad (2.12)$$

Simplifying and squaring Eq. 2.12 we find:

$$r_{ot}^2 = \frac{s_t^2}{s_o^2}. \quad (2.13)$$

This is the squared correlation between the observed and true scores. It is equal to our reliability which again is between 0 and 1 with the closer the value to 1 meaning the instrument is more reliable. So we finish with:

$$R_{xx} = r_{ot}^2. \quad (2.14)$$

The last technique to measure reliability is the lack of a squared correlation between observed scores and error scores. This is very similar to the approach taken in the correlation between observed scores and true scores and the lack of variance proportion. We first want to find r_{oe}^2 , or the correlation between observed scores and error scores, which is given by:

$$r_{oe}^2 = \left(\frac{s_e^2}{s_o^2} \right) \quad (2.15)$$

This is found following the same steps that we found r_{ot}^2 . The value of r_{oe}^2 is supposed to be small when it is a reliable instrument. The reliability coefficient for this correlation is what we have already found in Eq. 2.7 but now we write it as:

$$R_{xx} = 1 - r_{oe}^2. \quad (2.16)$$

2.3 Looking at Favorable and Unfavorable Responses

Our instrument is created to seek out unfavorable responses. By seeking out unfavorable responses with this instrument we can determine the components of NOS that subjects do not understand. Because the instrument is looking at unfavorable responses it is also easier to see movement from the unfavorable response to a more favorable response from pretest to post-test. With this instrument subjects will have either unfavorable, neutral, or favorable responses to the given question. We want to look at unfavorable responses because when pretest and post-test data are compared the percentage of favorable responses can stay the same while the percentages of neutral and unfavorable responses can change. Previously developed instruments have primarily examined the change of the favorable responses percentage. This instrument can and will look at how unfavorable responses on a pretest can change to more neutral responses on the post test. This means that in some way after the class that the student who responded unfavorably on the pretest has in some shape

or form changed his or her thinking in a more favorable direction.

We can visualize this by plotting the percentage of favorable responses on the y-axis and unfavorable responses on the x-axis. This can be seen in Fig. 2.1. When plotted, the data point does not always, and usually does not, add up to 100% because there are students who answered neutral. We want the data points from the pretest to move in either the upward or up and left direction. If they move straight upwards the percentage of favorable responses increased. This can mean one of two things: that some of the respondents changed from a neutral response to a favorable response, or that the same percentage of unfavorable responses that changed to either neutral responses or favorable responses were cancelled out by either changing neutral or favorable responses (hopefully not favorable responses changing to unfavorable) to keep the same percentage of unfavorable responses. The data point moving up from pretest to post-test is a positive change because more students answered favorably. If the data point moves to left, this means the the favorable percentage stayed the same but the unfavorable percentage decreased. This is also an improvement as there are less students responding unfavorably: they are starting to understand the ideas behind NOS. When the data point moves up and to the left, the percentage of favorable responses increased and the percentage of unfavorable responses decreased, which is the ultimate goal. Finally, what we do not want is for the data points to move either down, to the right, or down and to the right, all of which mean that the percentage of unfavorable responses are increasing.

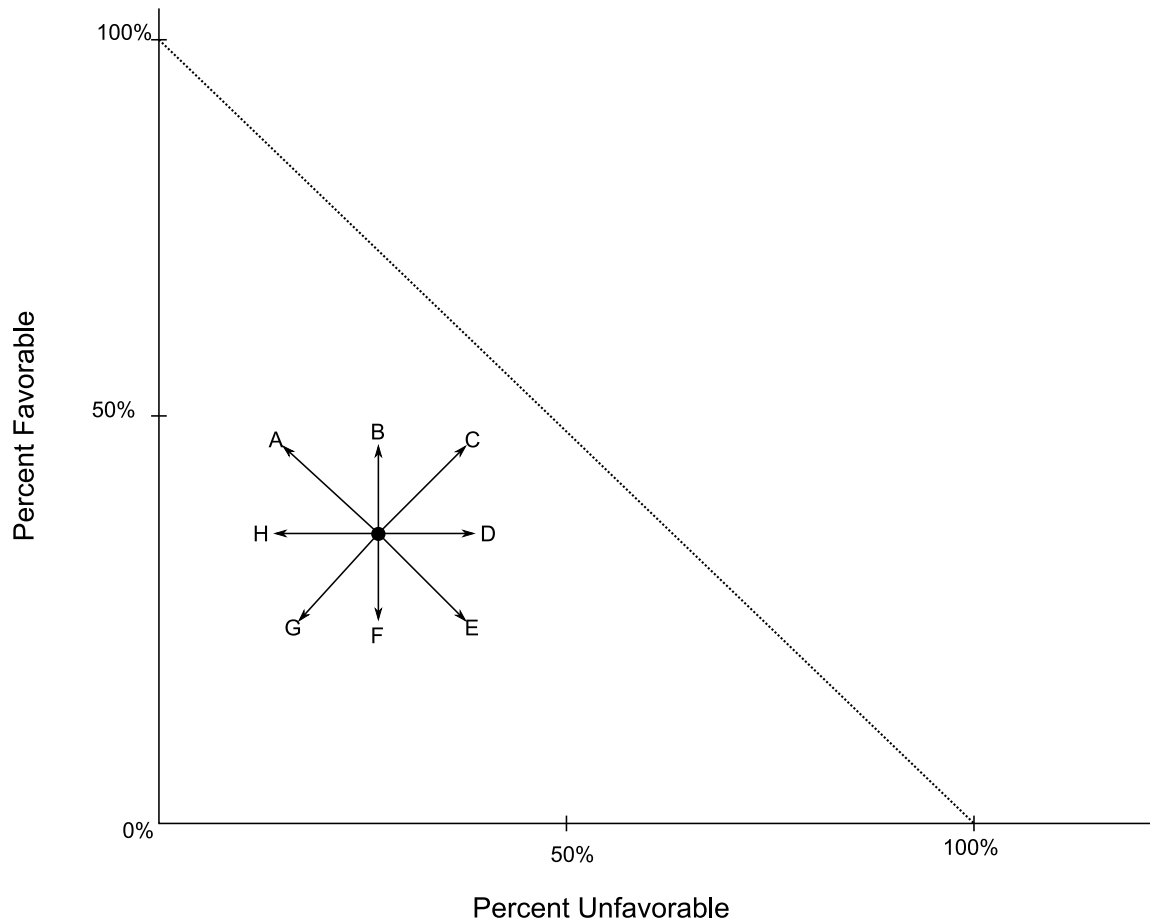


Figure 2.1 How students' responses can change from pretest to posttest. (A) The students have more favorable responses and less unfavorable responses. (B) The students have more favorable responses and the same amount of unfavorable responses. (C) The students have more favorable and unfavorable responses. (D) The students favorable responses stayed the same amount while the unfavorable responses increased. (E) The amount of favorable responses decreased and the amount of unfavorable responses increased. (F) The amount of favorable responses decreased and the amount of unfavorable responses remained the same. (G) The amount of favorable and unfavorable responses decreased. (H) The amount of favorable responses remained the same and the amount of unfavorable responses decreased.

Chapter 3

The Instrument

Our instrument is going ask respondents to give a letter grade to individual statements about NOS. As previously mentioned this means that the participants are given a statement and they are then asked to decide what type of letter grade they would give to the statement if it was a response on a quiz or test. The general process of creating this instrument is:

- Research Question
- Write individual questions
- Group individual questions
- Send instrument to professionals
- Analyze professional responses to eliminate poorly written questions
- Pilot revised instrument

3.1 Developing a New Instrument

To start the development of this instrument, we needed to start with a main research question. Our question was “How do we build a reliable tool to measure understanding of the Nature of Science?” Some questions that went along with the research questions were:

- “What is NOS?”
- “What makes the instrument reliable?”
- “How do you measure an understanding of science?”

The difficult question that we have to answer throughout the entire process of creating this instrument is how do we actually measure understanding of science? We try to answer this throughout the developmental process of the instrument through testing and revising our question set. In this process we started by creating goals which will guide us to writing the instrument.

3.1.1 Writing the Individual Questions

Initially we decided to use a Likert-style (Agree/Disagree) scale on our survey. After many discussions we felt that a Likert-style scale would not be an effective means to measure students’ understanding of science. This was changed because we originally wanted to have professionals select the response that they would NOT want their students to select on the Likert scale. We wanted to select the answer they would not want their students to choose because we felt there would be more agreement on the wrong answer as compared to what they believe is the correct selection. This lead us to take a different approach to what our responses would be. Instead, we are now asking the professionals to give a letter grade to each statement assuming

that each statement is an answer to a question that they asked about that particular section of NOS on a quiz or test. We feel that this is a clearer indication of what the respondents' understanding of science is. This unconventional approach will allow us to both see what the professionals understanding of science and what they would or would not want students to say about science.

When writing the statements we made sure that each one was a strong or absolute statement. Strong statements force the students to pick whether they agree or disagree with the question compared to when soft statements are used. This is because it is much easier to get lost in the language and wording of the question when the statements are weaker. For example, one of our statements is "Scientific theories are subject to testing and revision." To change this into a weaker statement we could state, "Scientific theories may be subject to testing and revision." The first statement forces the student decide whether or not he or she believes that theories are subject to testing and revision. On the other hand, the second statement could have students respond with "sometimes." The student could get confused because this statement does not specify whether it is all, some, or just one scientific theory that is subject to testing and revision.

3.1.2 Putting Together the Preliminary Instrument

We started by creating a list of 30 questions about NOS. This list of 30 questions was not in any order or organized in any specific way. The point of this first draft of questions was to start thinking about how to write these questions and link them to NOS ideas to evaluate students on. Next, we reviewed the first list of compiled questions. Next, we went through several iterations of the instrument to ensure:

- Concepts are directly related to defined subscales

- All statements are strong
- Languages is not distracting or confusing

The subscales were based on the work by Lederman et al. The subscales for our instrument are:

- Theory vs. Law
- Creativity
- Myth of the Scientific Method
- Societal Pressures
- Uncertainty
- Subjective and Tentative Nature of Science

Each of our preliminary questions fit into one of the six subscales. The subscales were expanded to 10 or 11 questions each giving us a total of 61 questions to send to professionals for comments. This list of questions can be see in Appendix A.

3.2 Giving the Preliminary Questions to Professionals

The first step in reducing a 61 question instrument into a reliable 30 questions instrument is to give the give the question set to a selection of professionals in various fields and have them evaluate the questions. The professionals were chosen because they have educational background and work in various levels of post-secondary education. This allows us to get a general understanding of how valid and reliable our

instrument is, and at the same time allow us to see if the professionals believe that our instrument is measuring NOS understanding.

Before the professionals reviewed our instrument we created an informed consent form for all of the professional volunteers participating in our study. The informed consent form can be seen in Appendix B. Next, we created an example of what the professionals will be using to evaluate and make comments on our instrument. This form can be seen in Appendix C. Finally, we compiled a list of professionals who would be willing to participate in our survey. These professionals are emailed about what they will be doing with the consent form and sample survey attached. Then we give them a link to the actual survey that we have constructed on a survey website. The website that we will use www.surveymonkey.com.

3.2.1 What is asked of the Professionals

The professionals are asked first to read the informed consent form before continuing. We make sure to let them know that if at any time they wish to discontinue their participation in the survey that they can just close their browser. The when we give the professionals the survey they will be asked some demographic information about the institution they teach at and the classes they teach. These questions are can be seen in Appendix C. The professionals will then be asked to go through and answer each of the 61 NOS questions we have on our survey. At the beginning of each section of NOS ideas we give a statement stating that they should assume that they have asked a question on a quiz or test about NOS. Each section is edited so it is a question about that particular aspect of NOS. For example, in the Theory vs. Law section we state “For this page assume you asked your students to write a statement that describes their understanding of theories or laws. Below are 10 student statements that you have received, please select the grade that you would give to each statement.”

With all NOS surveys there is always some room for argument about what the one single correct answer is. By looking at our survey responses we feel that we will have more agreement between the professionals on what the wrong answers are. After each question on the survey there will be a comment box that the professionals will have a chance to leave comments in. This allows the professionals to give us comments on questions they may have particularly liked or disliked.

Chapter 4

Analysis

The first set of data that we have and need to look at is the responses that we have received from the professionals. This data set will give us demographic information about the institution they teach at and the classes that they teach. The demographic information will allow us to better look at the results that our respondents give us. We will then look at these responses both quantitatively and qualitatively to allow us to find which questions need to be discarded or reworded.

4.1 Professional Quantitative Analysis

To analyze the data quantitatively we first need to sort all of the question responses. Each question can then be given its own “GPA” based on the responses that we have received from the professionals. The GPA will be calculated just like a semester GPA, with an ‘A’ being a 4.0 and an ‘F’ being a 0.0. Then we average the numbers. This gives us an average for what response was given to each question, but does not give us any idea about the amount that the respondents agree on a certain grade for each statement. This means that a questions could receive a GPA of a 2.0, but could

have been graded as an A, B, C, D, and F by five different respondents. To show a correlation that the respondents have graded a particular question the same, we need to see what percentage each letter grade got for the particular statement. This shows if a certain letter grade, or a certain range of letter grades are selected more often than the other letter grades. This is a correlation study of the questions.

Next, we have to use this data to discard some questions. We want to be able to use this data to discard some questions to help us narrow down which statements are the best ones to ask. To decide which questions we want to discard, we will use the data on which questions have the most agreement. The more agreement the question has the better. This shows that the professionals are not confused by the question and can easily decide their opinion on the statement. The questions that do not have a high amount of agreement need to be discarded, or edited. The disagreement may come from the wording of the question, disagreement between disciplines of the professionals, or just a disagreement of the statement in general.

4.2 Professional Qualitative Analysis

The quantitative analysis tells us statistically how much the professionals agree with each question, but just because the professionals agree statistically, it does not necessarily mean that the question is a well written question that should move onto the next round of testing. Statistically, the question may agree almost 100% but we also need to look in the comments box of each question to get the complete data set for the questions. For example, one question could have near 100% agreement that 'F' is the correct response, but then we read the comments for this questions. We may receive comments like "this is poorly written, therefore I was forced to select 'F'," "this statement did not make sense to me so 'F' was the logical choice," or "re-write

this question.”

To decide qualitatively which questions we want to keep for the next round of testing we first looked at the amount of comments that each question receives. If there are no comments for the question at all it is a question that we would want to keep for more testing. the questions that are commented on by fewer than 20% of the respondents we can assume that this is a well written question that does not need to be discarded because of negative comments. A question that receives more than the 20% comments we then evaluate what the comments are saying. If most of the comments are positive, or offer suggestions to make the statement a better written comment than we can still keep this question. It is the negative comments that we need to seek out because these are the questions that we need to discard.

Throughout our survey we also have general comment areas. These are after each section that we have created on survey, and on general comment box about the entire survey. These comment boxes give respondents an area to comment on each section. These comments may be about the section as a whole. Some comments that may be received are comments about whether or not they believe that we are asking about an important aspect of science, or whether the respondent may feel different about the ideas that we present in that area. The general comment box will give us an idea, based on the comments, if people are willing to adopt and use this instrument.

4.3 Who Responded to the Instrument

We sent a formal email out to 30 different professionals. Of the 30 professionals that we sent our instrument to we received 6 responses. 6 responses may not seem like a high number of responses, but this is 20% of our targeted audience. This is not a bad response amount when you consider it in terms of a non-captive audience. What we

mean by a non-captive audience is that the professionals that we send this instrument to do not have to complete the survey and they do not have an incentive to complete the survey. The responses were entirely volunteer based.

4.4 Case Study Analysis

Of the 6 responses that we received only 4 of the respondents actually finished the entire survey. 2 of the 6 respondents only filled out the demographic information. This is troubling because it now skews the demographic data when using it to look at the instrument data because all of the responses to the survey are anonymous.

The small number of respondents changes the analysis tools that we have available. In order to perform reliability analysis as described previously, we need the mean of the data to be stable under small changes. With only four respondents one person changing their answer will swing the mean through extremes. In addition validity studies need respondents to give comments on their interpretations of each question. In many cases only one respondent answered any of the questions making validity difficult. However, we can look at the individual responses in order to identify invalid responses.

To attempt to show some of the validity of the instrument we need to look at the professional responses to see which questions they have the greatest agreement on. To find this we did a case study analysis. For our case study we will look at every question individually and look for an emerging response for each question. Again, we want to look for agreement from the respondents, so by an emerging response we want one letter grade for each question or one end of the letter grade scale to be selected most often. For example if we have a question that has 3 responses or 75% 'A' and 1 response or 25% 'B' than this is an emerging response showing us that there is a

high agreement that this is a good statement about science. Further, this means that this is a good statement for us to keep on the instrument. Another example of this could be if 50% selected 'D' and 50% selected 'F' for a question. We can see example histograms in Fig. 4.1.

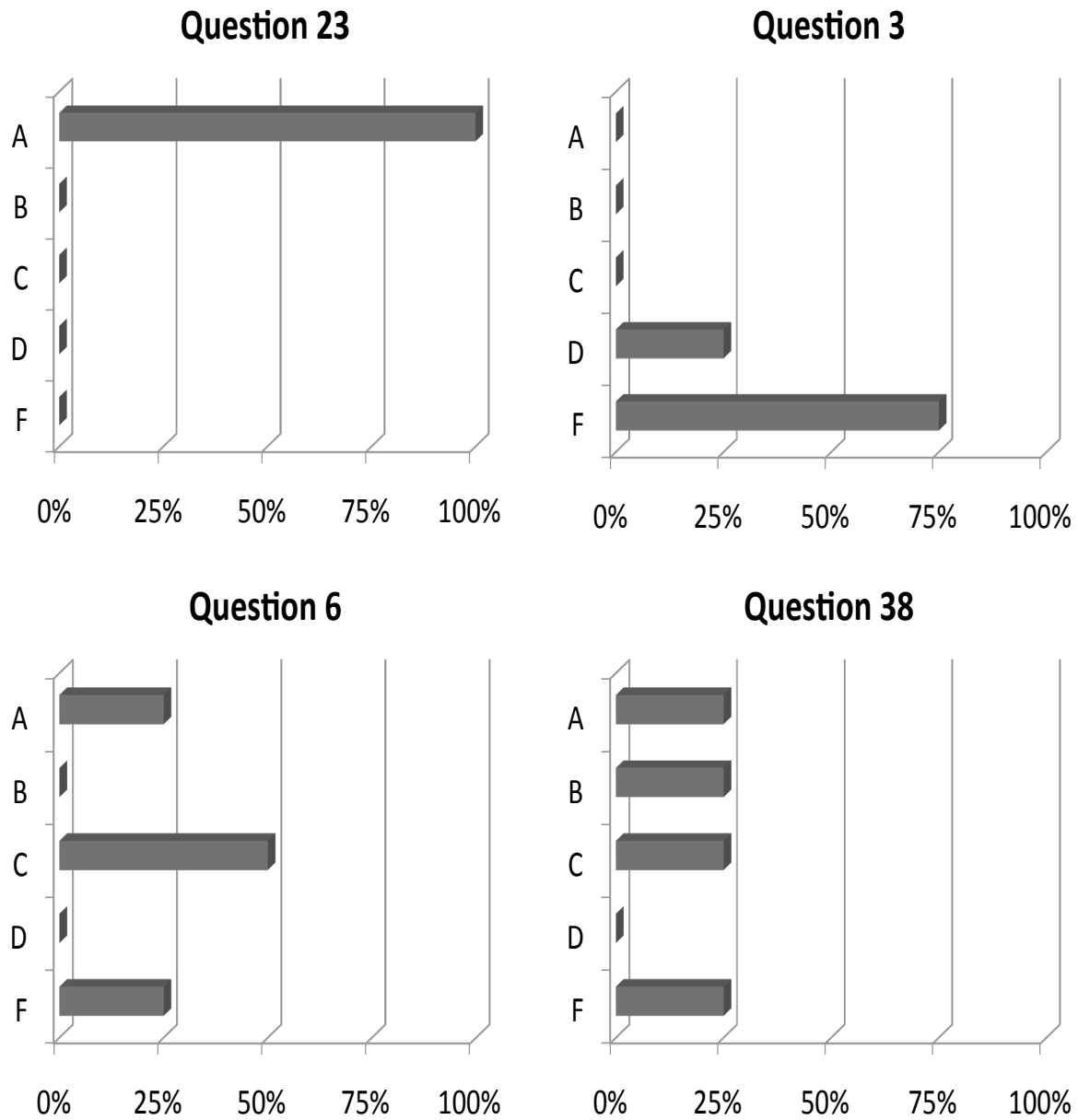


Figure 4.1 Four example histograms with the top two histograms representing emerging responses and the bottom two showing questions that were discarded or edited.

Again, this shows that the professionals are agreeing that this is not a good statement about science. For the instrument, it doesn't matter if the professionals agree that it is a good or bad statement about science, just that they agree. It doesn't matter if it is a particularly good or bad statement about science because we will then want to see if students can tell if it is a good or bad statement about science.

In Table 4.1 we can see each question and the responses that the professionals gave. Each question number refers to the NOS questions in Appendix A. Table 4.1 shows us which questions need to be eliminated because of the responses that we received from the professionals. We will start by looking at each grouping of questions and eliminating the questions that did not have emerging responses from the professionals.

Table 4.1 Responses to each instrument question

Question	'A' Responses	'B' Responses	'C' Responses	'D' Responses	'F' Responses
1	0	0	0	2	2
2	0	1	1	0	2
3	0	0	0	1	3
4	0	0	1	0	3
5	0	1	1	0	2
6	1	0	2	0	1
7	0	0	1	0	3
8	0	1	0	0	3
9	1	0	0	0	3
10	0	1	0	0	3
11	3	0	0	0	1
12	3	0	0	0	1
13	2	0	1	0	1
14	0	0	1	0	3
15	0	0	0	1	3
16	2	0	0	0	2

17	0	0	0	1	3
18	0	0	1	0	3
19	2	1	0	0	1
20	1	1	0	0	2
21	0	0	2	0	2
22	0	1	1	1	1
23	4	0	0	0	0
24	4	0	0	0	0
25	0	0	0	2	2
26	0	0	1	1	2
27	0	0	0	1	3
28	0	0	1	1	2
29	4	0	0	0	0
30	2	1	1	0	0
31	0	1	0	1	2
32	2	1	1	0	0
33	0	0	1	1	2
34	1	3	0	0	0
35	0	0	1	0	3
36	0	2	2	0	0
37	1	1	2	0	0
38	1	1	1	0	1
39	1	2	1	0	0
40	2	1	1	0	0
41	1	0	2	0	1
42	0	0	1	0	3
43	0	0	1	0	3
44	1	2	0	0	1
45	0	0	0	0	4
46	0	3	0	0	1

47	3	0	1	0	0
48	0	0	0	0	4
49	0	0	0	0	4
50	0	0	0	0	4
51	3	1	0	0	0
52	2	0	0	0	2
53	0	0	0	1	3
54	4	0	0	0	0
55	0	0	1	1	2
56	0	0	0	1	3
57	3	1	0	0	0
58	4	0	0	0	0
59	4	0	0	0	0
60	0	0	0	1	3
61	4	0	0	0	0

4.4.1 Theories and Laws

In the theory and law grouping we were able to eliminate questions 2, 5, and 6. All of these questions had three different letter grades selected and the grades were spread out so that there was no emerging response for the question. Because of this we were able to eliminate these questions from our survey. The interesting thing that came up while looking at this section is that all of the questions had a majority of the responses being ‘F’ answers. In fact one of the comments we received in the general comments box was “I think there needs to be some statements that might earn an ‘A’ from the professor.” We agree with this comment and will have to rewrite a few of these questions so that they could earn an ‘A’ grade. This may require us to rewrite a few of the questions as opposite to what they read, or to write some entirely new

questions.

4.4.2 Creativity

In the creativity subsection we immediately eliminated questions 13, 16, 19, and 20 because they did not have emerging responses. For questions 11 and 12 which both had 3 ‘A’ responses and 1 ‘F’ response we decided that we would keep them although the 1 ‘F’ answer seems to cause confusion to what the emerging response is. In a larger sample of responses from professionals we expect that this ‘F’ response would become an outlier.

4.4.3 Scientific Method

The scientific method section was pretty straightforward with our analysis. We eliminated questions 21, 22, 26, 28, and 30. We saw our first questions that got 100% of the same answers in questions 23, 24, and 29. We had a good balance of ‘A’ and ‘F’ emerging responses with 3 of the questions we are keeping emerging as an ‘A’ response and 2 emerging as an ‘F’ response.

4.4.4 Society

In the society subsection we ran into some problems. The way we were doing our case study we would eliminate questions 32, 33, 37, 38, 39, and 40. This would only leave us with 3 questions for the section. If we look at Table 4.1 we can see that question 34 had 1 ‘A’ response and 3 ‘B’ responses. This shows an emerging ‘B’ response, but because it is neither an ‘A’ or ‘F’ response it shows us that it is not a strong statement. This also applies to question 36 that had 2 ‘B’ and ‘C’ responses. Therefore we only were only left with 1 question out of the entire set of 10 that we

started with basically eliminating the entire subsection of questions. We still wanted to have societal questions on our survey because we view it as an important subsection of understanding of science. To fix this we had to decide which questions to eliminate and which questions needed to be rewritten. We eliminated questions 31, 33, 38, 39, and 40. Questions 32, 34, 35, 36, and 37 were rewritten as follows:

Question 32

- **Was:** Societal values and expectations determine what science is conducted and accepted.
- **Now:** Societal values and expectations *always* determine what science is conducted and accepted.

Question 34

- **Was:** Societal values and expectations determine how science is conducted and accepted.
- **Now:** Societal values and expectations *never* determine how science is conducted and accepted.

Question 35

- **Was:** All cultures conduct science in the same fashion.
- **Now:** All cultures *always* conduct science in the same *exact* fashion.

Question 36

- **Was:** Societal pressures determine what science research is conducted.
- **Now:** Societal pressures *sometimes* influences what scientific research is conducted.

Question 37

- **Was:** Politics influence the scientific research.
- **Now:** Politician always determine what scientific research will take place.

4.4.5 Uncertainty

We only found two questions to eliminate in the uncertainty subsection. We eliminated questions 41 and 44 from this section. Like the theory and law section a majority of the questions received ‘F’ responses, but there was a better balance than the theory and law section so we do not feel that we need to change any of the remaining questions in this section.

4.4.6 Subjective and Tentative Nature of Science

In this subsection we only eliminated questions 52 and 55. Of the remaining questions we had 6 of the questions emerge to be an ‘A’ response and 3 of them to be ‘F’ responses giving a a good balance of both sides of the of the grading scale.

Chapter 5

Conclusions

An important part of undergraduate learning, especially at a liberal arts college, is for students to become scientifically literate, meaning that students understand the processes and products of science, or the Nature of Science. To ensure that students are understanding the processes and products of science we needed to create an instrument that would be able to measure students understanding of NOS.

To create our instrument we first need ask our research question, “How do we build a reliable tool to measure understanding of the Nature of Science?” Although the development of the instrument is ongoing, we have started to answer this question. We started by writing a large list of NOS questions. Next, we put these questions into groupings that we felt were important aspects of science that non-science majors should know and understand. To make sure the instrument is reliable we then sent the question list to a group of professionals. The responses from the professional allowed us to narrow down the question list from 61 questions to the current list we have now of 39 questions.

We were able to conduct a case study analysis on the professional responses. In this study we were looking for emerging responses. With this we were able to eliminate

some questions getting us closer to our goal of 30 questions for the final instrument. We found some interesting results in the society subscale. In that subscale we needed to eliminate most of the preliminary questions that were written. We still felt that the society subscale was an important part of our instrument. To solve eliminating most questions we had to rewrite some of the questions. When rewriting the questions, we made sure that they were strong or absolute statements that could not be answered with a “sometimes.” response.

To finish answering our research question we need to continue to develop the instrument. In the future the new instrument, with 39 questions, needs to be revised and sent out again for more testing. As of right now, our survey is still open to be answered by the first group of professionals. If more professionals answer the first pass we can then do statistical analysis and revise the current list of questions that we have. If we do not receive anymore professional responses we do currently have a pilot instrument. We can now give the pilot instrument to students who were taught NOS principles. With this data we can then revise the pilot instrument. We continue this process until we have our final instrument to use with our students.

5.1 Future Work

- Statistical analysis on future responses from first group of professionals
- Pilot instrument with students
- Send out revised instrument to second group of professionals
- Revise instrument again from larger sample of responses
- Use completed instrument to measure students understanding of science!

Statistical analysis will require larger numbers of respondents. These numbers will come from student responses to the pilot instrument. The pilot instrument will be given to the students in introductory majors and non-majors science courses. Students will be administered the instrument in the same manner as the professionals. The survey will ask the students to imagine that they are a teacher of a science class and each of the questions was a response on a quiz or test from one of their students that they need to grade. The students will be given the same grading scale, A-F, that the professionals were given.

Once we have to student responses we can then do our statistical analysis for validity and reliability of the instrument. This will again allow us to eliminate troublesome questions eventually narrowing down the list of questions. The ideal amount of questions for the final instrument will be around 30.

Bibliography

- [1] F. Abd-El-Khalick, R. Bell, and N. Lederman, “The nature of science and instructional practice: Making the unnatural natural,” *Science Education*, vol. 82, no. 4, pp. 417–436, 1998.
- [2] N. R. Council, *National science education standards*. National Academy Press Washington, DC, 1996.
- [3] N. Lederman, F. Abd-El-Khalick, R. Bell, and R. Schwartz, “Views of nature of science questionnaire: Toward valid and meaningful assessment of learners’ conceptions of nature of science,” *Journal of Research in Science Teaching*, vol. 39, no. 6, pp. 497–521, 2002.
- [4] N. Lederman, “Teachers’ understanding of the nature of science and classroom practice: Factors that facilitate or impede the relationship,” *Journal of Research in Science Teaching*, vol. 36, no. 8, pp. 916–929, 1999.
- [5] E. Redish, J. Saul, and R. Steinberg, “Student expectations in introductory physics,” *American Journal of Physics*, vol. 66, no. 3, pp. 212–224, 1998.
- [6] A. Elby, “Helping physics students learn how to learn,” *American Journal of Physics*, vol. 69, p. S54, 2001.

-
- [7] I. Halloun, "Views about science and physics achievement: The VASS story," in *AIP Conference Proceedings*, vol. 399, p. 605, 1997.
- [8] W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein, and C. 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*, vol. 2, no. 1, p. 10101, 2006.
- [9] G. Aikenhead and A. Ryan, "The development of a new instrument: Views on Science-Technology-Society (VOSTS)," *Science Education*, vol. 76, no. 5, pp. 477–491, 1992.
- [10] G. Deboer, "Historical perspectives on inquiry teaching in schools," *Scientific inquiry and nature of science*, pp. 17–35, 2004.
- [11] N. Lederman, "Nature of science: Past, present, and future," *Handbook of research on science education*, pp. 831–879, 2007.
- [12] D. Colton and R. Covert, *Designing and constructing instruments for social research and evaluation*. Jossey-Bass Inc Pub, 2007.
- [13] R. Furr and V. Bacharach, *Psychometrics: an introduction*. Sage Publications, Inc, 2008.

Appendix A

NOS Survey

NOS Questions before professional review:

Theory/Law

1. Scientific theories become scientific laws.
2. Scientific laws are more important than scientific theories.
3. If a scientific theory is proven it becomes a scientific law.
4. Until a scientific idea is called a law, it is just a theory.
5. Unlike theories, scientific laws are not subject to change.
6. Scientific theories explain scientific laws.
7. Scientific theories are not as important as scientific laws because they are not proven.
8. Scientific laws are scientific theories that are proven.
9. Scientific theories are the first step towards finding scientific laws.

10. The difference between scientific laws and scientific theories is that laws are proven true and theories are not proven.

Creativity

11. When a scientist is setting up an experiment, she will be creative.
12. When a scientist is analyzing data, he will be creative.
13. When a scientist is collecting data from an experiment, she will be creative.
14. Scientists do not use creativity because this conflicts with their logical reasoning.
15. Scientists do not use creativity because this interferes with objectivity.
16. Scientists need to use creativity.
17. Creativity is a negative influence on science.
18. If a scientist uses creativity while collecting data he or she will have bad data.
19. Science needs scientists to use creativity.
20. Scientists must use creativity in scientific research.

Myth of the Scientific Method

21. All scientists follow the same step-by-step scientific method.
22. When scientists follow the scientific method correctly, they always get accurate results.
23. Scientists use different types of methods to conduct scientific investigations.
24. Experiments are not the only means used in the development of scientific knowledge.

25. The scientific method is the only way to conduct an experiment.
26. When conducting experiments all scientists follow the same sequence of steps.
27. All scientific experiments are conducted in the same order of procedure.
28. The scientific method is a step-by-step process that all scientists use in order to accurate results.
29. There is more than one way to conduct scientific research.
30. The process followed in scientific research varies from researcher to researcher.

Societal Pressures

31. All societies conduct scientific research the same way.
32. Societal values and expectations determine what science is conducted and accepted.
33. Scientific research is not influenced by society because scientists are trained to conduct pure, unbiased studies.
34. Societal values and expectations determine how science is conducted and accepted.
35. All cultures conduct science in the same fashion.
36. Societal pressures determine what scientific research is conducted.
37. Politics influence the scientific research.
38. Scientific research will vary when conducted in different cultures.
39. Religion influences scientific research.

40. Economics influences scientific research.

Uncertainty

41. After analyzing data scientists always have uncertainty measurements, these are from errors made by the scientists in the data collection.

42. When scientists give uncertainty measurements this is because they have made a mistake in their experiment.

43. When scientists give error estimates on a measurement this is because they have done something wrong during the measurement.

44. Uncertainty measurements are how much a calculated value may differ from the true value.

45. Error bars on plots show how much of an mistake was made during the experiment.

46. Error bars on plots show how much the calculated value can differ from the true value.

47. No measurement is exact.

48. All measurements scientists make are exact.

49. All measurements scientists make are exact and because they are exact error estimates are from a mistake in the measurement.

50. If the same block is measured by different scientists with the same ruler all of the scientists would measure the exact same length.

Subjective and Tentative Nature of Science

51. Scientific theories change over time.
52. A scientific theory is just a "best guess" and not actually a proven fact.
53. When a scientific idea gets the name theory, like the theory of relativity, it will never change.
54. Scientific theories change because of new research.
55. Scientists observations about the same event will be the same because scientists are objective.
56. Scientists observations about the same event will be the same because observations are facts.
57. Scientists may make different interpretations based on the same observations.
58. Scientific theories may be completely replaced by new theories in light of new evidence.
59. Scientific theories may be changed because scientists reinterpret existing observations.
60. Scientific theories based on accurate experimentation will not be changed.
61. Scientific theories are subject to testing and revision.

Appendix B

Consent Form

INFORMED CONSENT FORM Development of Ithaca College Measure of Student Understanding of Science

1. Purpose of Study: The purpose of our study is to create a likert-type instrument that will be able to successfully measure students understanding of the Nature of Science (NOS), and more specifically the understanding of non-science majors in general education science classes.

2. Benefits of the Study: This study will allow us to better select and narrow down the amount of survey questions to give to students on our instrument. For you this will allow you to add your own opinion on which NOS survey questions are appropriate and well written to allow measurement of students understanding. For the scientific community this will add an instrument that was uniquely created to measure students understand of NOS to then bring in better teaching practices for these students.

3. What You Will Be Asked to Do: We have compiled 61 NOS questions for our instrument. We would like to narrow down this list. For this survey you will be asked about the normal class size you teach, the type and size of the institution you teach

at (but not the institution name), the type of non-major science class you teach (if any), and how often you teach that course. We will then ask you to go through the 61 survey questions and choose the answer that you would NOT want to see your student give. You will then be given room to comment on questions that you found particularly good or bad. The amount of time that you will need to dedicate to this survey will be based on the level of interaction you have with the questions. We expect that this should not take you much longer than an hour.

4. Risks: We will not be asking for any personal data about you, but as previously stated information about your institution (size and type), and the classes that you teach to non-major science students. Data collected from this survey on the internet will not have identification attached to it. This email is our only identification of who we have asked to participate. There is minimal risk to your confidentiality by participating in this study.

5. Compensation for Injury: If you suffer an injury that requires any treatment or hospitalization as a direct result of this study, the cost for such care will be charged to you. IF you have insurance, you may bill your insurance company. You will be responsible to pay all costs not covered by your insurance company. Ithaca College will not pay for care, lost wages, or provide other financial compensation.

6. If You Would Like More Information about the Study: You should contact Nathan Porter at nporter2@ithaca.edu or Matthew Price at mprice@ithaca.edu.

7. Withdraw from the Study: If you would like to withdraw from the study you may close your browser at anytime to exit the survey. Your results will not be saved if you close your browser before submitting your answers to the survey.

8. How the Data will be Maintained in Confidence: Data obtained in this study will be on a secure password protected file server. Access to this system is only granted to the leader of the research team, Matthew Price.

I have read the above and I understand its contents. I agree to participate in the study. I acknowledge that I am 18 years of age or older.

Appendix C

Sample Survey

This is a model for what the participants of our survey will see.

Purpose of Study: The purpose of our study is to create a likert-type instrument that will be able to successfully measure students understanding of the Nature of Science (NOS), and more specifically the understanding of non-science majors in general education science classes.

Benefits of the Study: This study will allow us to better select and narrow down the amount of survey questions to give to students on our instrument. For you this will allow you to add your own opinion on which NOS survey questions are appropriate and well written to allow measurement of students understanding. For the scientific community this will add an instrument that was uniquely created to measure students understand of NOS to then bring in better teaching practices for these students.

What You Will Be Asked to Do: We have compiled 61 NOS questions for our instrument. We would like to narrow down this list. For this survey you will be asked about the normal class size you teach, the type and size of the institution you teach at (but not the institution name), the type of non-major science class you teach (if any), and how often you teach that course. We will then ask you to go through the

61 survey questions and choose the answer that you would NOT want to see your student give. You will then be given room to comment on questions that you found particularly good or bad. The amount of time that you will need to dedicate to this survey will be based on the level of interaction you have with the questions. We expect that this should not take you much longer than an hour.

If You Would Like More Information about the Study: You should contact Nathan Porter at nporter2@ithaca.edu or Matthew Price at mprice@ithaca.edu.

Withdraw from the Study: If you would like to withdraw from the study you may close your browser at anytime to exit the survey. Your results will not be saved if you close your browser before submitting your answers to the survey.

To answer these survey questions please select the response that you would FAIL a student for giving.

Question 1: If a scientific theory is proven it becomes a scientific law. a) I strongly agree b) I agree somewhat c) I am not sure if I agree or disagree d) I disagree somewhat e) I disagree strongly

Place leave comments or questions you may have about this particular question.

There will be a total of 61 questions

Demographic Information:

1. My institution is: a) A public research institution b) A private research institution c) A community college d) other
2. The size of my institution is: a) Over 20,000 students b) 10,000-20,000 students c) 6,000-10,000 students d) 2,000-5,000 students e) under 2000 students
3. Do you teach non-major science classes a) yes b) no
4. If yes how many students: a) over 200 students b) 100-200 students c) 50-100 students d) 25-50 e) under 25 students
5. How often do you teach this or these classes (if you teach a non-major science

course) a) 1 per year b) 2 per year c) every other year d) more than 2 years between teaching e) once only

