RASCH ANALYSIS OF STUDENT RESPONSES TO THE COLORADO LEARNING ATTITUDES ABOUT SCIENCE SURVEY

by

Xi Tang, BS

A thesis submitted to the Graduate Council of
Texas State University in partial fulfillment
of the requirements for the degree of
Master of Science
with a Major in Physics
August 2016

Committee Members:

David Donnelly, Chair

Wilhelmus Geerts

Donald Olson

FAIR USE AND AUTHOR'S PERMISSION STATEMENT

Fair Use

Duplication Permission

DEDICATION

For my father and my mother.

ACKNOWLEDGEMENTS

In the process of writing this thesis, I received help from different people. I would like to thank them for making this thesis possible.

First, I would like to give my greatest thanks to my thesis advisor Dr. David Donnelly from the Department of Physics at Texas State University. He suggested this intriguing thesis topic to me. When writing this thesis, he offered me liberty and freedom in deciding what to study about Rasch Analysis and students' epistemological beliefs. He also provided me with enlightened guidance when I was lost in exploration. Without his continuous concerns for my thesis, and his enthusiasm, encouragement, and support, this project could hardly have been completed. In addition, I am deeply grateful to Dr. John M. Linacre, the developer of Winsteps, which is the Rasch software used in this project. Dr. Linacre is also the Research Director of Winsteps.com, where he replies to users' questions. His help with Winsteps has been essential in finishing this work.

This project received the Thesis Research Support Fellowship from the Graduate College. Without this fellowship, this project would have been impossible. Furthermore, I would like to thank the Physics Education Research group in the Department of Physics. The weekly meeting held in 2015 and the discussion with the group members outside the meeting have provided me with important ideas about how to conduct my research.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF ABBREVIATIONS

Abbreviation	Description
CLASS	Colorado Learning Attitudes about Science Survey
JMLE	Joint Maximum Likelihood Estimation
ICCs	Item Characteristic Curves
DIF Test	Differential Item Functioning Test
VASS	Views about Science Survey
MPEX	Maryland Physics Expectation Survey.
EBAPS	Epistemological Beliefs Assessment for Physical Science
LEPS	Learning Physical Science
PSET	Physical Science and Everyday Thinking
NOS	Nature of Science Survey
SI	Scientific Inquiry
TIMMS	Trend in International Mathematics and Science Study
STEM	Science, Technology, Engineering and Mathematics
SACS	STEM Awareness Community Survey
SD	Strongly Disagree
D	Disagree
N	Neutral
A	Agree
SA	Strongly Agree

ABSTRACT

The Colorado Learning Attitudes about Science Survey (CLASS) was developed by University of Colorado, Boulder. It investigates students' beliefs about physics and learning physics[1], and has become one of the most popular attitude surveys used in physics.

Beginning in 2010, Dr. David Donnelly, Eleanor Close and Hunter Close started using CLASS to investigate students' change in attitude during introductory physics courses at Texas State University. They used the standard analysis developed by the creators of CLASS. In their study, students' attitudes did not display significant change during any semesters. However there was significant change during the winter break. This effect is called Winter Break Effect[1].

The credibility of a survey conclusion is closely related to the quality of the survey and its analysis method. By examining the CLASS and its analysis method, these researchers quickly noticed some problems: the standard analysis only partially counts a student's responses, and the difficulty of each statement is not weighted when determining this student's score.

This study reanalyzed the data from the aforementioned study by using the Rasch model, which makes full use of the responses from all five survey categories. The results obtained support the existence of the Winter Break Effect, and also found that an instructor significantly influenced students' attitudes toward physics during her class. In

addition, this study compared the Rasch Analysis and the standard CLASS analysis. The Rasch Analysis exhibited advantages over the standard analysis. Moreover, this study evaluated the design of CLASS and provided suggestions to improve its quality.

I. INTRODUCTION

In the past few decades, physics education researchers have found that students' views, expectations, beliefs and attitudes (these terms will be referred to as "epistemological beliefs") toward physics affect students' physics learning in a fundamental and profound way[2,3,4,5]. Measuring their epistemological beliefs is, therefore, necessary.

However, measuring beliefs is not easy. Since they are students' latent traits, they cannot be measured directly. They have to be inferred by students' words and actions[6]. Researchers have conducted interviews and observations to learn about students' epistemological beliefs, and they have also created several student self-report surveys, including the Views about Science Survey (VASS), the Maryland Physics Expectation Survey (MPEX), the Epistemological Beliefs Assessment for Physical Science (EBAPS), and the Colorado Learning Attitudes about Science Survey (CLASS) (these surveys will be referred to as "epistemological surveys"). However, the reliabilities of the conclusions obtained by these surveys are related to the quality of the surveys and the methods used to analyze student responses. Therefore, it is necessary to think about and try to optimize the quality of these surveys and the correspondent analyses.

Beginning in 2010, Dr. David Donnelly, Eleanor Close and Hunter Close from Texas State University started using the CLASS to assess students' change in attitudes. In this study, the researchers investigated an introductory course which is two semesters in duration. They did the pre-course and post-course comparison for each semester and they did not find any favorable shift of students' attitudes (i.e., attitudes becoming more expert-like) during semesters. However, by comparing the students' attitudes before and

after a winter break, some statistically significant favorable shifts have been found. This effect is referred to as the "Winter Break Effect". However, the standard CLASS analysis used in the aforementioned study shows an insufficient usage of students' responses, which makes the existence of this effect unclear.

CLASS is a Likert scale survey which has five categories for each statement. These five categories are "Strongly Disagree (SD)," "Disagree (D)," "Neutral (N)," "Agree (A)" and "Strongly Agree (SA)". As discussed by the CLASS developers, there are two ways of analyzing the Likert scale in CLASS, regarding it as an ordinal scale (Chapter II, 2.3.1 in this thesis) or an interval scale (Chapter II, 2.3.1). An interval scale is a more meaningful scale when comparing it with an ordinal scale[7] (this will be discussed in Chapter II, 2.3.1). However, the standard analysis chooses the former way to analyze the CLASS data. For an ordinal scale, the differences between adjacent categories are not the same and are unmeasurable. It is, therefore, impossible to assign values for these categories. As a result, except counting the numbers of students' responses in each category, there is hardly anything that can be done to use the information from each category. In addition, when regarding the Likert scale as an ordinal scale, every category has an ordinal meaning. Through the interviews with students, however, the researchers found it is not reasonable to assume that SA and A have different ordinal meanings (the same for SD and D.) because students have different interpretations to the term "strongly" (this is in-depth discussed in Chaoter II, 2.3.1 and Chapter III, 3.2). To the same extent of agreement, students may end up choosing SA, or they may end up choosing A. Therefore, the researchers decided to regard SA and A as the same response, and SD and D the same response. Moreover, in researchers'

interviews, they found students have varied interpretations toward N, which makes it hard to define the ordinal meaning of this category. They, therefore, decided to neglect students' responses in the "N" category. Finally, the researchers decided the standard way of analyzing CLASS is to count the number of students' responses, which have the same tendency as experts' responses, and use the percentage of these responses to be a student's score[1]. This percentage is called favorable percentage.

An example of calculating this percentage is that if the experts' response for a statement is SA/ A and a student selected SA/ A for it, this statement will be counted into this student's favorable percentage. If a student did not select SA/ A, no matter which category he or she selected, the measurement of this student's attitudes would not be affected. In other words, this student's responses in any category other than SA/ A are neglected in the measurement of his or her attitudes.

If the Likert scale in CLASS can be treated as an interval scale, taking responses in every categories into the consideration of measurements can be possible (this is in-depth discussed in Chapter II, 2.3.2 and Chapter III, 3.2). Fortunately, analyzing the Likert scale in CLASS as an interval is possible. The researchers from Texas State University quickly found that the Rasch model is a ready model to analyze the Likert scale as an interval scale.

The Rasch Analysis has very different features from the standard analysis. Not only does it treat the Likert scale as an interval scale, the Rasch Analysis also considers the different weight of each statement. In addition, the Rasch Analysis displays the distribution of students' ability of holding expert-like attitudes along a vertical continuum which is calibrated by the difficulty of every statement. The change of the distribution of

students' ability has never been investigated. Moreover, the Rasch Analysis is able to discover the drawbacks of a survey, which can provide researchers with clues to improve the quality of this survey.

To date, the Rasch Analysis has been never applied to any of the epistemological surveys for Physics. If Rasch Analysis is applicable to CLASSs and shows advantages over the standard analysis, it may also be applied to other epistemological survey and improve the measurements.

II. LITERATURE REVIEW

2.1 The influence of students' epistemological beliefs

David Hammer[8] (1994) studied six students with different beliefs of physics, and discussed the correlation between their beliefs and behaviors in learning physics. He found these six students could be characterized by their beliefs about the structure of physics. The students who believed physics to be comprised of fragmented facts, formulas and procedures were defined as students characterized by Apparent Concepts. These students casually make or break the conceptual connections. The students who believed physics to be consist of concepts governed by general principles were defined as students characterized by Concepts. These students, in contrast, carefully build and modify their understanding of concepts. Hammer asked three questions of each student to probe for common misconceptions. Students with Apparent Concepts beliefs each showed fundamental misconceptions in his or her responses. However, students with Concepts beliefs did not display any misconceptions on any of the three questions. Hammer noticed students with Apparent Concepts beliefs have disconnections between their intuitive knowledge and what they considered as "physics knowledge". They dismissed the disconnections, instead of trying to account for them. These students decided they understand physics when they find apparent conceptual connections or they are able to follow the details of the formal manipulations. However, students with Concepts beliefs would like to question and modify their understandings, although sometimes they could not resolve the conflicts of their knowledge. Hammer concluded that some students' knowledge remains fragmented, partly because they do not expect it to be coherent. They retain their misconceptions, partly because they do not think

conceptual knowledge is important, and it is not necessary to modify their own understandings.

Laura Lising and Andrew Elby [9] (2005) carefully discussed the causality of a student's learning difficulty and her epistemological beliefs. In the classes about "electrical field" and "light and shadow", this student, "Jan", showed a series of troubles in understanding and using the relation $E = F/q$, as well as the behavior of light in group discussions. The researchers noticed that this student showed uncommon (comparing with her group mates) inclinations toward using technical and mathematical terms in explaining phenomena and debating with her group mates. She refused to listen to her group mates when they used intuitive/ everyday terms, even if they were right. The researchers refuted the possibilities that Jan's mathematical skills, expectations of this class (whether formal explanations would be rewarded), learning habits (whether she checks her answers), or confidence in using intuitive/ common sense reasoning are the reasons for her words and behaviors in her discussion, and they concluded that Jan's epistemological belief that there is a "wall" between formal reasoning and intuitive/ everyday reasoning, and the intuitive/ everyday reasoning cannot be used to explain physics phenomena was the reason.

House J. Daniel[10] claimed students' expectations of a course are a significant predictor of their grades. He investigated students' initial (at the beginning of a course) expectations toward an introductory college chemistry course (whether they wanted to make a B from this course, or whether they wanted to be honor students), their ACT grades, the number of years of mathematical courses taken, and their final grades of the introductory college chemistry course at the end of that semester. He found that a

student's expectations of their achievement are better predictor than this student's former academic achievement, and mathematical ability, for his or her future academic achievement.

2.2 Methods of measuring students' attitudes

The existing methods of measuring students' attitudes include interviews, observations and surveys. Hammer interviewed students after he read their homework (mentioned in Chapter II, 2.1); Lising and Elby observed a student's behavior in group discussions (mentioned in Chapter II, 2.1); and more research has been done using surveys.

Widely-used epistemological surveys include the Views about Science Survey (VASS), the Maryland Physics Expectation Survey. (MPEX), the Epistemological Beliefs Assessment for Physical Science (EBAPS), and the Colorado Learning Attitudes about Science Survey (CLASS).

2.2.1 Views about Science Survey

VASS was developed by Ibrahim Halloun and David Hestenes. It was used to assess students' views about learning science for the purpose of assessing the influence of their views on learning. This survey has a Biology version, a Chemistry version and a Physics version. Each one has 30 items. Item 13 is from VASS, the Physics version. This item is presented below for the purpose of showing the format of VASS.

13. The first thing I do when solving a physics problem is:

(a) represent the situation with sketches and drawings.

(b) search for formulas that relate givens to unknowns.

Answer Options

① Only (a), Never (b); ② Mostly (a), Rarely (b); ③More (a) Than (b); ④ Equally (a) & (b);

⑤ More (b) Than (a); ⑥Mostly (b), Rarely (a); ⑦ Only (b), Never (a); ⑧ Neither (a) Nor (b)

VASS was designed in Contrasting Alternatives Design (CAD) format, which "provides students with two contrasting statements and asks students to express their position relative to both"[11]. These researchers chose this format because they found when the VASS was administered in essay format (the initial format of VASS), a student replied that the first thing he did when solving a physics problem is searching for formulas that relate givens to unknowns. However, in the follow-up interview, when the interviewer asked him whether he ever considered drawing some kind of a diagram, he replied actually the first thing he usually did when solving a physics problem was representing the situation with sketches and drawings. He thought this procedure was too trivial, and not worth mentioning in his essay response. To avoid this situation from happening again, the researchers decided to present students with two contrasting alternatives (a folk alternative and an expert alternative) at the same time, and prompt them to choose their positions from an eight-point scale. Each point on the scale represents a different inclination to these two alternatives. Options 1 to 3 imply that a student agrees with (a) more than (b), and options 5 to 7 imply that a student agrees with (b) more than (a). Each option implies an inclination to a different extent. Option 4 implies that a student agree with these two alternatives to the same degree, and option 8

implies that a student does not agree with either of these alternatives. The expert view is either (a) or (b). It is determined by both the VASS creators and a number of science experts. When (a) is the expert view, if a student's choice is one of the options 1 to 3, this student is considered having an expert tendency for this item.

VASS has two parts, scientific and cognitive parts. The scientific part contains three dimensions, the structure, methodology and validity of science. The cognitive part contains four dimensions, the learnability of science, students' critical thinking in learning science and their personal relevance toward science. The analysis results are usually presented as Figure 1, which shows the distributions of students who expressed a tendency towards expert views on a given number of items for one of the dimensions in VASS. Figure 2 is the common method of presenting the results of pretest-posttest comparison.

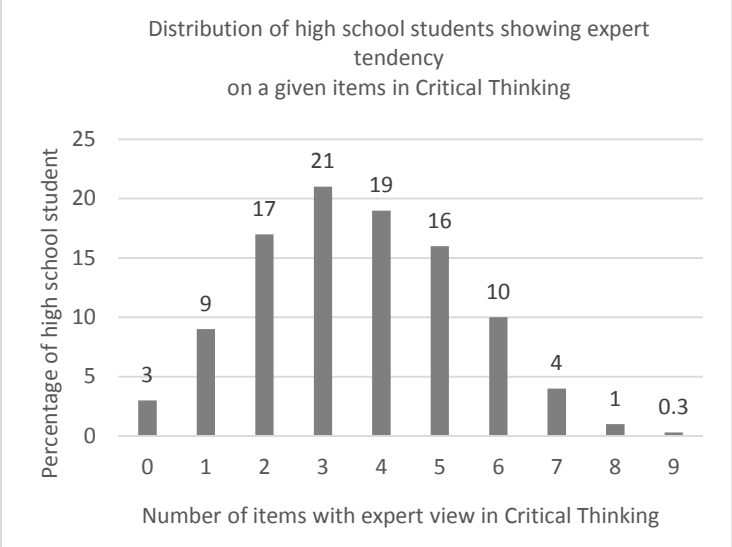

Figure 1 The common method of presenting the VASS results

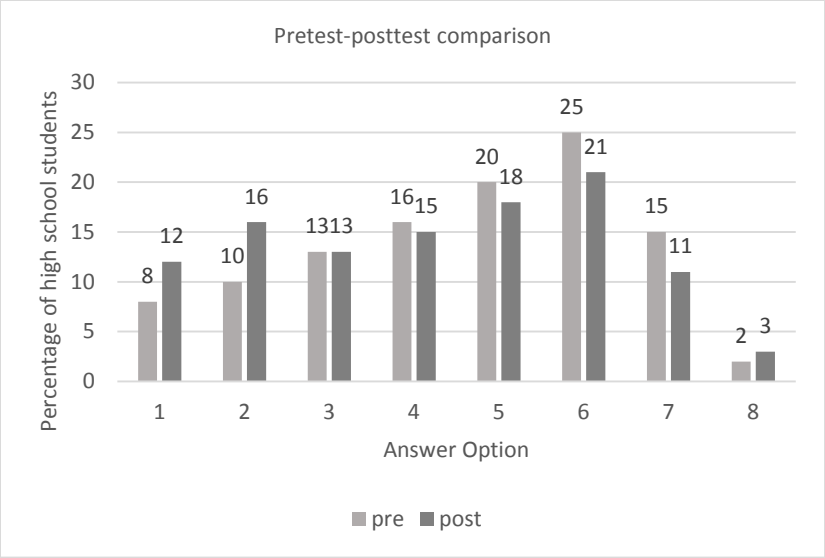

Figure 2 The common method of presenting pretest-posttest comparison in VASS

By using VASS, researchers got the following conclusions[11]:

1. High school and college students usually have novice tendency about science and learning science. However, scientists and teachers usually have expert tendency.
2. Every high school and college student has a mixture of novice and expert tendencies in any of the dimensions.
3. High school teachers and university professors do not have expert tendency on every item. University professors show higher percentage of agreement with expert views than high school teachers.
4. Traditional science courses do not have significant influence on students' views. Sometimes, these courses even make students shift further away from expert views.
5. Students' novice tendency are deep-rooted and do not change after years of high school and college schooling.

6. Learning science does not affect students' views. However, students' views seem to affect the outcome of learning. High academic achievers are usually students who have expert tendency; High risk students are usually novice tendency; and average students are usually mixed tendency.
7. Students show different views toward different disciplines (Biology, Chemistry and Physics).
8. In both high school level and college level, female students show higher tendency toward expert views than male students.

2.2.2 Maryland Physics Expectation Survey

MPEX was designed by the Physics Research Group from the University of Maryland. This survey measures students' expectations about what kind of things they will learn, what skills will be required, what they will be expected to do in physics courses, and also their views of the nature of science information. The objectives of this survey are to investigate (1) students' expectations of physics and a physics course at the beginning of a semester; (2) the effect of students' expectations on their learning during this semester; and (3) the effect of the physics course on students' expectations.[12]

The MPEX has 34 statements. Each statement is followed by a five-point Likert scale (includes SD, D, N, A, SA). Every student needs to select one stance from the scale to represent his or her attitude toward this statement. Item 8 (Table 1) from MPEX is presented below for the purpose of showing the format of MPEX.

Table 1 Item 8 from MPEX

8	In this course, I do not expect to understand equations in an intuitive sense; they must be taken as givens.	1 2 3 4 5[a]

Explanatory note of item 8:

a. The meaning of each number: 1: Strongly Disagree; 2: Disagree; 3: Neutral; 4: Agree; 5: Strongly Agree.

The 34 statements in MPEX belong to 6 dimensions: (1) learning independence (Independence); (2) coherence of physics knowledge (Coherence); (3) notion of constructing and understanding concepts (Concepts); (4) knowledge-reality link (Reality link); (5) physics-math link (Math link); and (6) personal effort (Effort). The Independence dimension investigates whether a student learns independently and tries to construct his or her own understanding, or simply takes what is given by authorities without evaluation; The Coherence dimension investigates whether a student considers physics is a connected, consistent framework, or it is made up of separated facts; The Concepts dimension investigates whether a student thinks understanding concepts or memorizing formulas is important; The Reality link dimension investigates whether a student thinks physics is relevant in real contexts; The Math link dimension investigates whether a student considers mathematics as a way of representing physics phenomena; The last dimension, the Effort dimension investigates whether a student thinks it is necessary to make sense out of and make use of what he or she has learnt from physics[12].

The standard method of analyzing the MPEX data is to calculate the mean of every student's favorable and unfavorable percentages for the overall statements and each dimension. In pretest-posttest comparison, the standard MPEX analysis calculates the

gain of students' favorable percentage, and uses Paired Sample t Test to examine the significance of the favorable gain. The results of MPEX is usually presented as Table 2.

Table 2 The common method of presenting the MPEX results

	Overall	Independ.	Coherence	Concepts	Reality link	Math	Effort
Pre	55[a]/21[b]	49/34	53/21	41/28	57/13	68/16	74/11
Post	66/17	60/24	78/11	68/12	78/6	82/8	55/28
Main gain score	11[c]	11[c]	25[c]	27[c]	21[c]	14[c]	-18[c]
s.d. of gain	20	28	32	32	36	33	29

Explanatory note of Table 2:

a. The mean of every student's favorable percentages;

b. The mean of every student's unfavorable percentages;

c. $p<0.01$, significant gain.

Some main conclusions derived by using the MPEX includes:

1. Students usually have expectations that are different from professors' expectations of their roles in physics courses, what they can learn from a physics course, and the nature of physics and physics courses.
2. Professors usually aren't aware of or don't deal with students' different expectations. Consequently, course goals are always not achieved[4].
3. Traditional physics courses do not help with the discrepancy between the professors' and students' expectations. They can even exacerbate the differences between them[13].
4. Best reformed curricula like Problem-Based Learning (PBL), which have been very successful at improving students' efficiency in learning concepts, cannot make students' expectations shift to experts' expertations[14].

5. Epistemological – focused courses can lead to significant favorable shifts in students' expectations. Epistemological-focused courses take epistemology into consideration in students' homework- and test-question selection, homework-grading policy, classroom discussion and even labs[13].

2.2.3 Epistemological Beliefs Assessment for Physical Science

EBAPS investigates students' understanding of the nature of science and science learning. It is also designed by the Physics Education Research Group from the University of Maryland. This survey is aimed at high school students and college students in introductory physics, chemistry or physical science courses[15].

EBAPS has 30 items in total. 17 items are statements with five-point Likert scales; 6 items are multiple choice questions; and 7 items are written mini debates for students to choose a side on a continuum. To make the format of EBAPS easy to understand, one item in each form is presented in Appendix.

The scoring scheme of EBAPS is complicated. Each item is scored on a scale of 0 (least sophisticated) to 4 (most sophisticated). The scoring scheme is non-linear, which also can be seen from the Appendix. EBAPS has five dimensions, including (1) the structure of scientific knowledge; (2) the nature of knowing and learning; (3) the real-life applicability; (4) evolving knowledge; and (5) the source of ability to learn. The standard way of presenting EBAPS results is calculating students' overall score of EBAPS and each dimension. The score is the average of the student's scores on every item in EBAPS or the dimension.

Table 3 The common method of presenting the EBAPS results

	Overall	Structure of knowledge Concepts, Coh.	Nature of learning Independence	Real-life Applicability Reality link	Evolving knowledge	Source of ability…
Pre	66.5	62.5	68.4	73	63.9	72.6
Post	71.8	70.9	75.0	73.5	67.9	77.4
Main gain score	5.3[a]	8.4[a]	6.6[a]	0.5[a]	4.0[a]	4.8[a]
s.d. of gain	8.7	17.7	11.9	15.6	21.5	17.3

a. $p<0.02$

2.2.4 Colorado Learning Attitudes about Science Survey

CLASS was developed by W. K. Adams et al. from University of Colorado, Boulder. This survey is established on the aforementioned surveys (VASS, MPEX, and EBAPS). CLASS was designed for students of all educational levels and it examines a broader range of issues that educators consider to be important for learning physics when comparing with other surveys[1]. It is the most up-to-date and most widely used attitudes survey at present.

CLASS has 42 statements. One of them is a filter statement, which asks students to select category 4 as their answer to ensure that the students are paying attention to the survey. The format of CLASS is the same as MPEX, which can be seen from the Appendix.

To determine the experts' attitudes to each statement, the researchers from University of Colorado, Boulder, interviewed several experts (physics education researchers and practicing physicists who are interested in teaching physics) many times, and asked 16 experts to give their responses to CLASS. For five statements, physics experts were not able to reach a consensus on a response. Therefore, only 36 statements are counted in the scoring process. The standard CLASS analysis is also the same as the

analysis method adopted by MPEX, which calculates students' favorable or unfavorable percentages.

By using CLASS, researchers found:

1. In traditional physics classes, students' attitudes usually do not become more expert-like after a semester's study, and sometimes even become more novice-like. That is to say, traditional physics courses may not benefit students' attitude development and sometimes even do harm to it[1].
2. Physics major students have more expert-like attitudes than non-physics majors. Students' expert-likes attitudes are largely a pre-existing trait of students who choose to be a physics major rather than something developed through courses[16].
3. To help students make attitudinal gains, it is necessary to explicitly instruct students about the nature of science and learning science, and this kind of instruction must be embedded within content. "Physics and Everyday Thinking" (PET) and "Physical Science and Everyday Thinking" (PSET) curricula are this kind of curricula. Several researchers reported attitudinal gains by using these curricula[17].
4. In a study which investigated the effect of Modeling Instruction on students' attitudes, the researchers found significant favorable shifts on students' attitudes (students' attitudes became more expert-like) during semesters. They also found little favorable shift during winter breaks[18].
5. "Physics by Inquiry" (PbI) curricula also helps students' with positive attitudinal gains[19].

The following paragraphs will briefly describe the features of the aforementioned curricula:

In a general PET or PSET class, students develop their own model through experiments and consensus discussion, and after they established their own models, they read about the historical development of that concept or watch videos about these experiments. In this way, they find the generally accepted models. They usually have time to discuss about the consistency or inconsistency between their models and the accepted ones[17]. The Modeling Instruction curriculum mimics the expert physicists' process of constructing a knowledge structure. Students have to generate their own model for some new knowledge based on what they already know and reexamine their model cyclically. This keeps the coherence of physics knowledge. In addition, the Modeling Instruction curriculum aims to create an authentic science discovery process, the problems students are facing are more similar to the problems that scientists encounter[18].

PbI is not an attitude-centered curriculum. It only addresses students' attitudes of science and the nature of science in an implicit way. In this curriculum, students mirror the process that practicing scientists do to create new knowledge. With some guided questions, students use simple equipment to conduct an in-depth study of a few fundamental concepts in physics. They collect their data, record their observation, generate and modify their models of how things work. The exams in this curriculum are open book and open notes. The exams require students to apply the ideas they have learned to deal with problems they have never seen, so learning by rote is not helpful to get good grades. Even though in these classes, knowledge about experts' attitudes and beliefs were not explicitly explained, students' behavior were similar to experts'

behavior, so there was a potential that this curriculum would lead students to think about physics in an expert-like way[19].

2.3 Measurement of attitudes

2.3.1 Types of scale

In Psychometrics, there are four types of scale: (1) nominal, (2) ordinal, (3) interval, and (4) ratio[7].

On a nominal scale, there is no ordering among the categories. For example, a survey may ask: "What's your favorite color? 1. Blue; 2. Yellow" The numbers that used to label the categories do not imply one category is superior to the other category[20]. On an ordinal scale, the numbers of the categories indicate the ordering among them. For example, in a car-race, numbers are assigned to cars by the orders they finish the race[7]. This is an ordinal scale. However, the differences between adjacent categories on an ordinal scale are not comparable. (It's not reasonable to say the distance from the first car to the second car, is the same as the distance from the second car to the third car.) On an interval scale, however, the numbers not only have ordinal meaning, the difference between the numbers are also meaningful and comparable. The Fahrenheit scale is an interval scale. First, the numbers have ordinal meanings: 80 degree > 60 degree > 40 degree. Second, the differences are comparable: the temperature differences from 80 to 60 is the same as 60 to 40. However, it is apparent that the ratio of 80 to 60 is not the same as the ratio of 60 to 40. In a ratio scale, the ratio of numbers is also meaningful and comparable. For the ratio scale, the zero position means the absence of the quantity being

measured. (On the Fahrenheit scale, the zero position does not imply the absence of temperature.) The Kelvin scale is a ratio scale[20].

These four types of scale from nominal to ratio increases its power in the meaningfulness for the numbers used for measurement. In general, a person's latent trait (e.g. attitudes) does not have an absolute zero. The ratio scale is therefore not usually applicable in this kind of measurement, which makes the interval scale become the most informative scale in measuring persons' latent trait[7].

The standard CLASS analysis regards its Likert scale as ordinal scale. When regarding a Likert scale as an ordinal scale, just as mentioned in the previous paragraph, each category has an ordinal meaning. However, the CLASS developers found that in students' responses, "Strongly Agree" and "Agree", and "Strongly Disagree" and "disagree" do not always have ordinal difference. Because every student has different interpretation toward the word "Strongly". Students who agree with a statement to the same extent may end up choosing "Strongly Agree" or "Agree". Regarding these two responses differently is therefore inapplicable.

2.3.2 The Rasch theory

The Rasch theory is based on the following idea: each student has a probability of answering a question correctly, and this probability is related with this student's ability and a question's difficulty. The Rasch theory ignores every other factor that may affect the probability, and expresses the probability in the following way:

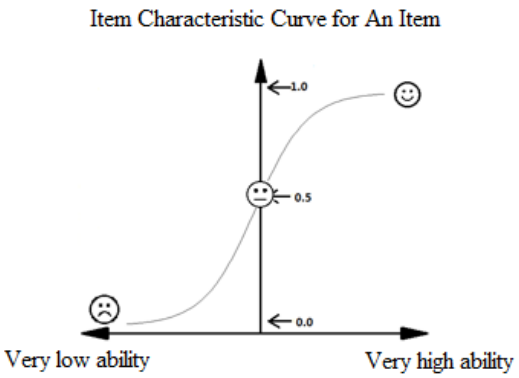

Figure 3 The assumption of the Rasch theory[7]

For a dichotomous question (only asks a student to select his or her answer from an incorrect option and a correct option) with a certain difficulty, the more able a student is, the higher the probability the student is able to answer is. The expression of the probability curve is:

$$p = \mathrm{P}(\mathrm{X} = 1) = \frac{\exp(B_n - D_i)}{1+\exp(B_n - D_i)} \quad (1)$$

The B_n represents a student n's ability and the D_i represents a question i's difficulty. X=1 means this student answers this question correctly, and p represents the probability. By changing the form of equation (1), we have equation (2) and (3):

$$\log\left(\frac{p_1}{1-p_1}\right) = B_1 - D_r \quad (2)$$

$$\log\left(\frac{p_2}{1-p_2}\right) = B_2 - D_r \quad (3)$$

The equation (2) is for a student with abilityB_1, and (3) is for a student with abilityB_2. The left sides of both equations are called log odds. A log odd is the logarithm of the ratio of the probability that a student answers a question correctly over the probability that this student answers this question incorrectly.

If we take the difference of these two equations, we get equation (4):

$$\log\left(\frac{P_1}{1-P_1}\right) - \log\left(\frac{P_2}{1-P_2}\right) = (B_1 - D_r) - (B_2 - D_r) = B_1 - B_2 \text{ (4)}$$

The difference of these two students' abilities is therefore expressed by the difference of their log odds. Therefore, it is reasonable to use a student log odds to represent this student's ability. The unit of students' ability is logit.

$$\log\left(\frac{P_{1\prime}}{1-P_{1\prime}}\right) - \log\left(\frac{P_{2\prime}}{1-P_{2\prime}}\right) = (B_n - D_1) - (B_n - D_2) = D_2 - D_1 \text{ (5)}$$

Similarly, statements' difficulty can also be represented by the log odds.

For a Likert scale, the Rasch theory only looks at two categories a time, and assuming one of the categories is a "correct" option, and the other one is an "incorrect" option. Equations (6) to (11) shows the process to obtain the relationship among a student's probability of choosing a category from the three-point Likert scale and a student's ability and the categories difficulty.

$$p_{0/0,1} = \Pr[(X=0)/(X=0 \; or \; X=1)] = \frac{\Pr(X=0)}{\Pr(X=0)+\Pr(X=1)} = \frac{1}{1+\exp(B-D_1)} \text{ (6)}$$

$$p_{1/0,1} = \Pr[(X=1)/(X=0 \; or \; X=1)] = \frac{\Pr(X=1)}{\Pr(X=0)+\Pr(X=1)} = \frac{\exp(B-D_1)}{1+\exp(B-D_1)} \text{ (7)}$$

$$p_{1/1,2} = \Pr[(X=1)/(X=1 \; or \; X=2)] = \frac{\Pr(X=1)}{\Pr(X=1)+\Pr(X=2)} = \frac{1}{1+\exp(B-D_2)} \text{ (8)}$$

$$p_o = \Pr(X=0) = \frac{1}{1+\exp(B-D_1)+\exp(2B-(D_1+D_2))} \text{ (9)}$$

$$p_1 = \Pr(X=1) = \frac{\exp(B-D_1)}{1+\exp(B-D_1)+\exp(2B-(D_1+D_2))} \text{ (10)}$$

$$p_2 = \Pr(X=2) = \frac{\exp(2B-(D_1+D_2))}{1+\exp(B-D_1)+\exp(2B-(D_1+D_2))} \text{ (11)}$$

The p_o represents a student's probability of choosing the first category on the Likert scale. The difficulty of the first category is assumed to be 0. The $D_1$is the difficulty of the second category.

$$P_r(X_{ni} = x) = \frac{\exp \sum_{k=0}^{x}(B_n - D_{ik})}{\sum_{h=0}^{m_i} exp \sum_{k=0}^{h}(B_n - D_{ik})} \text{ (12)}$$

$$\text{Where } exp \sum_{k=0}^{K}(B_n - D_{ik}) = 1 \text{ (13)}$$

Equations (12) and (13) are the general expressions for a Likert scale. $P_r(X_{ni} = x)$ is the probability for student n to choose the x category for question i. The k represents the number of categories.

The Rasch theory thinks that students' responses are sufficient information to estimate their ability and the statements' difficulty. The Rasch model adopts the Joint Maximum Likelihood Estimation (JMLE) to do the estimations. The expressions of JMLE are as follows[21]:

$$s_{ik} = \sum_n^N p_{nik}, \text{ i=1,…, I (14)}$$

$$r_n = \sum_{ik}^{IK} p_{nik}, \text{ n=1,…, N (15)}$$

$$\text{where } p_{nik} = \exp(B_n - D_{ik}) / (1 + \exp(B_n - D_{ik})). \text{ (16)}$$

The s_{ik} is the score that category k of question i received, and the r_n is the raw score that the student n received. Raw score is obtained by assuming category 1 counts 1 point, category 2 counts 2 points; etc. The Rasch model needs to assume every student's ability and every statement's difficulty, and based on the information of the raw scores, the Rasch model will find the best fit of students' ability and statements' difficulty.

In Rasch analysis, the results are always presented in Wright Maps. In a Wright Map, the students' abilities are located on the left side of the scale, and the items' difficulties are located on the same scale but the right side. When a student's ability equals to an item's difficulty, then the probability of "success" is 0.5. For instances, when the difference between the student's ability and the item's difficulty is 1 logit, the probability of success is 0.73; when the difference is 2 logits, the probability is 0.88; and

when the difference is 3 logits, the probability is 0.95. On the contrary, when the difference is -1 logit, the probability is 0.27; when the difference is -2 logit, the probability is 0.12; when the difference is -3 logit, the probability is 0.05.

Even though the Rasch model has significant advantages in analyzing Likert-scale data, the application of the Rasch model in science education is very rare. One of the reason is the complexity of the mathematic process, which limited the application of the model. After late the 1960s', due to the development of computer technology, the implementation of Rasch analysis was finally realized.

One of the primary application of the Rasch model is to examine the construct and item properties of assessments. Lamb, Annetta, Meldrum and Vallett used the Rasch model to assess the psychometric properties of the Science Interest Survey. This survey is a five-point Likert survey, and the items are empirically categorized into five subscales. In addition, the confirmatory factor analysis confirmed the empirical categorizations. However, Winsteps exhibited interval consistency over the items, which suggested the fitness of the items and also the unidimensionality of the structure. The authors chose to accept the conclusion from Rasch Analysis[22].

K. Neumann, I. Neumann and Nehm examined the (1) model fit and dimensionality, (2) reliability, and (3) item quality in the Nature of Science survey (NOS) and the Scientific Inquiry (SI) instrument. For the dimensionality, the researchers examined the derivations when applying the one-dimensional model and two dimensional model to the data, and found that the deviations was smaller for the latter, which indicates that the NOS and SI instrument are two dimensional assessments. They assessed the item fit with two indices: (1) mean square fit statistics (including infit and outfit statistics) and

(2) standardized z values (ZSTD). Even though an item should be regarded as over-fitting if the infit or outfit is less than 1.0. In this case, too few items fit this criteria, so the fit range had to be widened from 0.8 to 1.2. The acceptable range of ZSTD is from -2 to 2. There were different items poorly fitted when concerning the mean square fit statistics[23].

Glynn evaluated TIMMS (Trend in International Mathematics and Science Study) with Rasch Analysis. 16,005 students participated in the test. By examining results with Winsteps, they found relatively high reliability for TIMMS and they found most items fit the Rasch model. When exploring the item fit, they didn't consider ZSTD because this index is too sensitive (biased toward misfit), which is not suitable for a large sample. Only infit and outfit were considered and the acceptable range of the data is from 0.8-1.2. The Wright map of persons and items suggested that the items in TIMMS is a high stakes test. Items trend to be difficult, and there is a lack of easy items. Also, there are redundant items for same difficulty levels[24].

The Rasch program Sondergeld and Johnson applied in their research study toward the STEM Awareness Community Survey (SACS) is also Winsteps. Their choice was based on the polychotomous responses used in this survey and the underlying unidimensionality of SACS. They tested the item/person mean square fit statistics. The range they adopted is 0.6-1.4. When the mean square of an item/person is greater than 1.4, more misinformation will be brought into the study results than valuable information. When the mean square is less than 0.6, the response patterns are considered too predictable, which don't add new information to the construct. These items/persons do not harm the measure. Besides the infit/outfit assessment, the direction of point-biserial correlation for each item with the overall construct is also assessed for item fit. Items

with negative values were deleted, since they worked in opposition to the measure's construct. The separation and reliability of the items and persons were also examined[25].

2.3.3 The assumptions of the Rasch Analysis

Rasch Analysis assumes that a high-quality survey must have three properties: unidimentionality, equal item discrimination and no error to guessing. The unidimentionality implies that a measurement tool must measuring one latent trait at a time. If a survey has multiple dimensions, this survey should be analyzed by separate dimensions. The Rasch Analysis has two essential measurements, the measurement on a person's ability and an item's difficulty. The Rasch Analysis requires every item in a survey to have the same discrimination, because different item discrimination will make the item difficulty meaningless. Figure 4 and Figure 5 will illustrate this problem[26].

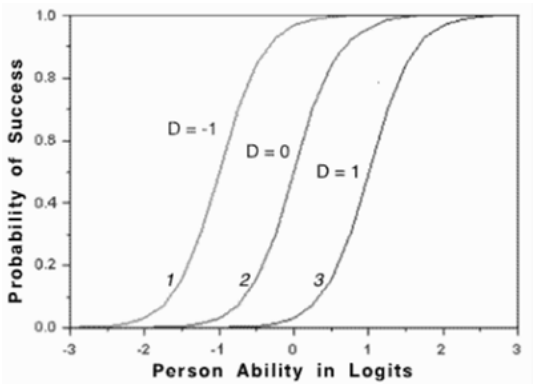

Figure 4 Item Characteristic Curves of items with the same discrimination[27]

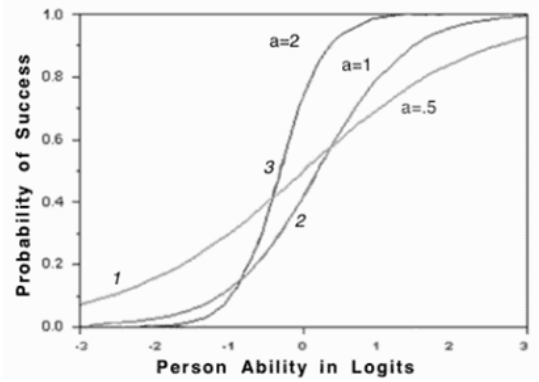

Figure 5 Item Characteristic Curves of items with different discriminations[27]

Figure 4 and Figure 5 are both Item Characteristic Curves (ICCs) for three different items. The horizontal axis of the ICCs graph represent person's ability, and the vertical axis represent a person's ability of being successful on an item. The item numbers are labeled beside the curves (1, 2, and 3), and the discrimination is the slope of the curve. In Figure 4, which the item discriminations are the same, no matter what a person's ability is, the order of the item difficulty (from easiest to the most difficult) is always item 1, item 2 and item 3. In Figure 5, the order of the item difficulty (from easiest to the most difficult) is item 1, item 2 and item 3. However, the three items are different in discrimination. For a person with -1 logits ability, the order of the item difficulty (from easiest to the most difficult) is item 1, item 2 and item 3. However, for a person with 2 logits ability, the order of the item difficulty is item 3, item 2 and item 1. In other words, the order of item difficult is meaningless in this survey when items' discrimination is not the same. Equal item discrimination is embodied in the mathematical expression in Rasch Model. Each item corresponds to only one item difficulty.

The third feature a test or a questionnaire must have is no success due to guessing. Gershon and Waller's study shows that success due to guessing doesn't show up

randomly. This type of success is based on some certain trait, such as risk-taking, culture background, test-wiseness and so on. A high quality test or questionnaire should be able to avoid success due to guessing. Rasch measurement will diagnose if the test or the questionnaire meets these three requirements.

2.3.4 Optimal usage of response categories

Linacre raised a set of criteria of optimal usage of response categories[27].

(1) Regular observation;

(2) A minimum of 10 observation for each category is required;

(3) Category measures must increase monotonically with categories;

(4) Outfit mean square statistics should be less than 2.00;

(5) Step calibrations (Rasch-Andrich thresholds) should increase monotonically with categories;

(6) For 3-point Likert scale, step calibrations should be at least 1.4 to 5 logits apart. For 5-point Likert scale, step calibrations should be at least 1.0 to 5 logits apart.

III. RESEARCH DESIGN

3.1 Objectives

This study attempts to check if the results and conclusions obtained by the standard analysis are consistent with those obtained by Rasch Analysis. In addition, this study also intends to seek extra information that Rasch Analysis can reveal about students' responses toward CLASS.

3.2 Rationale

The standard CLASS analysis uses a student's favorable percentage as the representative of this student's ability of holding expert-like attitudes toward Physics. The favorable percentage is the percentage of the statements toward which a student has shown a tendency to experts' attitudes. For example, if the experts' attitude for a statement is SA or A (these two responses are considered as the same response in the standard CLASS analysis), and a student selected SA or A for it, this statement will therefore be counted into this student's favorable percentage. However, if a student did not select SA or A, no matter which category he or she selected, the measurement on this student's ability would not be affected.

However, when considering the rating scale in CLASS as an ordinal scale (Chapter II in this thesis), the approach adopted by the standard CLASS analysis is possibly the most reasonable way of dealing with students' responses (Chapter II). Another way of analyzing a Likert scale is considering it as an interval scale (Chapter II). The developers of CLASS have discussed the possibility of analyzing the Likert scale in this survey as an interval scale. The approach they thought of was to assign the five

categories with the values from 1 to 5, which is a common approach to manipulate a Likert scale as an interval scale. However, these researchers immediately realized that it is not reasonable to assume the interval between different adjacent categories to be the same, such as the intervals between “Strongly Agree” and “agree”, and “agree” and “neutral”. Therefore, they did not find another appropriate way of analyzing students’ responses on the five-category Likert scale other than the standard CLASS analysis.

Rasch Analysis is an analysis which manipulates Likert scales as interval scales. Instead of simply assigning values 1 to 5 to each category, this analysis uses the joint maximum likelihood estimation (JMLE) (Chapter II) to estimate the value that should be assigned to each category of each statement, for example, the value that should be assigned to “Strongly Agree” of statement 1. (This value will be called statement’s difficulty in the rest of this thesis). Meanwhile, the estimation determines the level of students’ expert thinking. (This value will be called student’s ability in the rest of this thesis). Due to the mechanism of Rasch Analysis, a student’s responses in every category on the rating scale are figured in the calculation of this student’s ability. Therefore, the student’s ability measured in this way may be more accurate than that obtained by the standard CLASS analysis.

This study will use the Rasch Analysis to measure the students’ ability, and compare the change in the means of students’ ability measured in this way in the pre-course survey and the post-course survey. This change is considered as the students’ change in their attitude during this course in this study. In addition, to determine the significance of this change, this study will use Paired Sample t Test.

3.3 Description of courses

This study uses the same data from the former CLASS study, which was done using the standard CLASS analysis. The data were gathered from a series of elementary physics courses at Texas State University in the fall of 2010, and the spring and fall of 2011. This series of courses includes lecture courses Physics 1310 and Physics 1320, which employed the *Conceptual Physics* text by Hewitt, with chapters 1-18 covered in Physics 1310 and the remaining content covered in Physics 1320. Both courses were taught in sections of approximately 100 students, and students were allowed to take these two courses in arbitrary sequence. Physics 1310 and Physics 1320 are compulsory courses for students who are pre-service teachers or students who are Health Profession majors.

25 sections of classes in total had participated in the former study (15 sections of Physics 1310 and 10 sections in Physics 1320). These sections were taught by 7 different instructors, who followed the textbook to different degrees. In addition, their teaching style varied from traditional lecture through incorporation of interactive activities (concept questions and student voting), to the use of the Learning Physical Science (LEPS) curriculum, and team-based learning methods. No systematic observation of the classes was conducted. Table 5 demonstrates the distribution of the instructors in different semesters.

Table 4 Instructor distribution by semester

Term / Inst. #[a]	Fall 2010 1310	Fall 2010 1320	Spring 2011 1310	Spring 2011 1320	Fall 2011 1310	Fall 2011 1320
1	2[b]	/[c]	/	/	/	/
2	2	/	2	1	2	/
3	1	2	/	3	2	2
4	/	/	2	/	/	/
5	/	/	/	/	1	/
6	/	/	/	/	/	/
7	/	/	/	/	/	2

Explanatory Note for Table 5:

a. Instructor. # is the label of each instructor;

b. The number in each cell is the number of sections that the corresponding instructor taught for that term;

c. "/" means the corresponding instructor has not taught any sections for that term.

3.4 Participants

The enrollment of undergraduate students for the courses varies from 60% to 80% female over different semesters. The percentage of the students who were pre-service (elementary and middle school) teachers varies from 50% to 80%, and those who were Health Profession majors varies from 4% to 12 % over the course of different semesters. Specific demographic data is listed in Table 6.

Table 5 Demographic composition of the courses in study

Course and Semester	Enrollment	% Female	% Male	% Pre Serv. Teach	% Health Prof
1310 Fall 2010	469	84.0	16.0	65.2	9.4
1320 Fall 2010	289	85.5	14.5	80.3	7.6
1310 Spring 2011	386	80.8	19.2	62.2	12.4
1320 Spring 2011	379	84.4	15.6	74.7	4.7
1310 Fall 2011	551	59.7	40.3	51.4	11.6
1320Fall 2011	384	79.2	20.8	72.6	5.5

The sample used in this study is selected from that used in the former study. Statement 31 in CLASS is a filter statement (Chapter II). The students who had responded to this statement incorrectly were eliminated from the analysis in this study. In addition, this study determines the significance of students' change in ability by using the Paired Sample t Test, which requires matched students' measurements in the pre-course survey and the post-course survey. Therefore, the sample used in this study must be matched in both surveys. Table 7 shows the number of students remained after each step of sample selection.

Table 6 Sample selection of the courses in study

Term	Survey	Original[a]	Statement 31[b]	Match[c]
Fall 2010 1310	Pre	202	187	180
	Post	202	190	
Fall 2010 1320	Pre	53	52	48
	Post	53	49	
Fall 2011 1310	Pre	292	270	232
	Post	292	259	
Fall 2011 1320	Pre	338	311	155
	Post	207	197	
Spring 2011 1310	Pre	142	139	129
	Post	142	135	
Spring 2011 1320	Pre	149	142	138
	Post	149	145	
Winter break	Before	57	57	55
	After	57	57	

Explanatory Note for Table 7:

a. Number in original sample;

b. Number remaining after Statement 31 filter;

c. Number of pre/post matched pairs.

3.5 Administration

CLASS was administered during the regular class times in the first week and the final week of the each course. The exact timing of each administration was determined by the instructor of that course. Additionally, students were assured that their performance in the survey would not affect their grades.

3.6 Data analysis instruments

To estimate students' ability and statements' difficulty, Rasch Analysis has to perform a relatively complicated computation (Chapter II). When the sample size or the amount of statements which need to be analyzed is large, it will be challenging for

researchers to calculate the students' ability and statements' difficulty without using Rasch software.

There are several different Rasch software that are available for doing Rasch Analysis, (Chapter II). Among those software package, Winsteps stands out for its remarkable merits. First, Winsteps was created by two authorities in the field of Rasch Analysis, who are Dr. Michael Linacre and Prof. Benjamin D. Wright. Dr. Linacre has abundant practical experience in using the Rasch model in analyzing practical data, which can be traced back to the 1980s. Prof. Wright was a student and colleague of the creator of Rasch Analysis, Georg Rasch, and Prof. Wright is the leading advocate of Rasch Measurement from 1967 until 2001. Dr. Linacre is currently the Research Director of Winsteps. com, which contains a forum for Winsteps users to discuss questions with him. Users are always able to solve their problems efficiently. Second, Winsteps is a pragmatic Rasch software. When compared with other existing software, it has more input and output files which can realize various functions that are usually used in data analysis. In addition, Winsteps can report more than thirty major tables, files, plots and graphs about a survey. It is therefore suitable software to use if abundant information about the survey is wanted.

Winsteps can perform the Rasch measurement on students' ability and statements' difficulty. In addition, to determine the significance of the change in the means of students' ability for different surveys, the Pair Sample t Test in the SPSS statistics is also needed. The most recent versions (when the study began) of these two software packages were used in this study, which are Winsteps 3.81.0 and SPSS statistics 22.

3.7 Analysis design

As mentioned in 3.2, to determine students' change in ability, the measurements of students' ability in both pre-course survey and post-course survey are needed. The estimation method (JMLE) used in the Rasch Analysis that determines the measurement of a student's ability is related to the responses of every student in the same sample (Chapter II). Students' responses to the pre-course survey and the post-course survey are usually different. The measurements of ability in the pre-course survey and the post course survey are therefore incomparable. However, Winsteps provides Item-Structure File (IFILE) and Structure-Threshold File (SFILE) to address this problem. IFILE and SFILE are output files of some designated sample. IFILE contains the values of the statements' difficulty derived from the computation for that sample. SFILE contains the values of the step calibrations of the rating scale, which is are Rasch parameters for categories. (The combination of the statements' difficulty and the step calibrations will be called measurement system in the rest of this thesis.) IFILE and SFILE can be inserted into another sample. The measuring system in this sample is therefore forced to be the same as the previous sample.

Because this research seeks students' change in attitude during a semester or a break, it is therefore reasonable to use the measurement system built by the pre-course survey as the standard of comparison. The students' ability will be measured by the same measurement system. The change measured by this system is the students' change in ability for this course.

To determine the significance of students' attitudinal change, the students' ability in the pre-survey and the post-survey will be analyzed by the Paired Sample t Test in

SPSS. The t-value reported by this test will decide the significance of the attitudinal change. A similar method will be used to find out if the students' attitude truly changes during the winter break. In addition, this study will examine additional information provided by the Rasch model that may be useful providing more information about CLASS.

IV. RESULTS

4.1 Students' change in ability during semesters and the winter break

To investigate the students' change in attitude during each long semester (the fall and spring semesters) or the winter break, the first task was to establish an appropriate measurement system (Chapter III) for each term. To establish such a system, two requirements need to be satisfied: (1) the mean square fit statistics (including infit mean square value and outfit mean square value) of the students who are used for establishing this system should be less than 2; (2) the number of students who satisfied the requirement (1) should be no less than 36. Students whose mean square fit statistics are greater than 2 (the outliers) will distort or degrade the accuracy of the measurement systems. These students therefore should be eliminated from constructing the systems. However, sometimes the elimination is unnecessary if it does not make significant difference in measurement. For example, the measurement of students' ability does not significantly change after removing the outliers. The method to determine the necessity of an elimination is by doing a cross-plot of the students' ability measured by the system constructed before, and after removing the outliers. The horizontal axis of the cross-plot displays the measurements of students' ability before removing the outliers, and the vertical axis displays the measurements after removing them. If the cross-plot is approximately a straight line, which means the ratio between the measurements of two students stays the same and the elimination changes every measurement of students' ability to the same degree, then the elimination is unnecessary. In addition, if the measurement system is expected to generate stable results, the number of the students who are used to establish the system should be no less than the number of statements.

The number of statements that can be scored in CLASS is 36; therefore, at least 36 students are thus required for establishing the system.

Table 7 Sample size by term

Term	Winter break	Fall 2010		Spring 2011		Fall 2011	
		1310	1320	1310	1320	1310	1320
N	53	158	37	96	113	192	126

In the process of constructing the measurement systems for each term, the aforementioned requirements were strictly obeyed. The outliers were eliminated from measurement system construction. The cross-plots were also checked to determine the necessity of the elimination, and the plots showed that every elimination was necessary.

Sometimes after removing old outliers, new outliers appeared. The process of checking students' mean square fit statistics, deleting the outliers and doing the cross-plots therefore were thus reiterated several times, until there were no more outliers. The number of students within the acceptable range of fitness (mean square fit statistics <2) was greater than 36 for each term, which satisfied the requirement of the minimum number of students for performing stable measurements. In the measurement systems thus established, the mean square fit statistics of every statement were less than 2. Therefore, no statement was eliminated from analysis.

These systems were then used to assess the students' ability in the pre-course survey and the post-course survey. The first thing that is worth looking at is the Wright Map, which displays the relationship between the students' ability and the statements' difficulty. The left side of the Wright Map shows the distribution of students from the

least able at the bottom to the most able at the top, and the right side shows the distribution of the statements from the least difficult at the bottom to the most difficult at the top. Figure 6 to Figure 12 are the Wright Maps of pre-course surveys and post-courses survey in each term. The Wright Maps on the left side are for pre-course surveys (which will be called the “pre-maps”), and the Wright Maps on the right side are for the post-course surveys (which will be called the “post-maps”). For most courses during the semesters and the winter break, the distributions of the students’ ability are similar to normal distributions. Most students have moderate ability. Only a few students have expert-like ability or novice-like ability. The distributions of the students’ ability had slightly expanded in the post-course survey during the semesters, except for the PHYS 1320 in the fall semester in 2011. In other words, some students became more able after a semester’s study, and some others became less able, while the mean ability remained approximately the same.

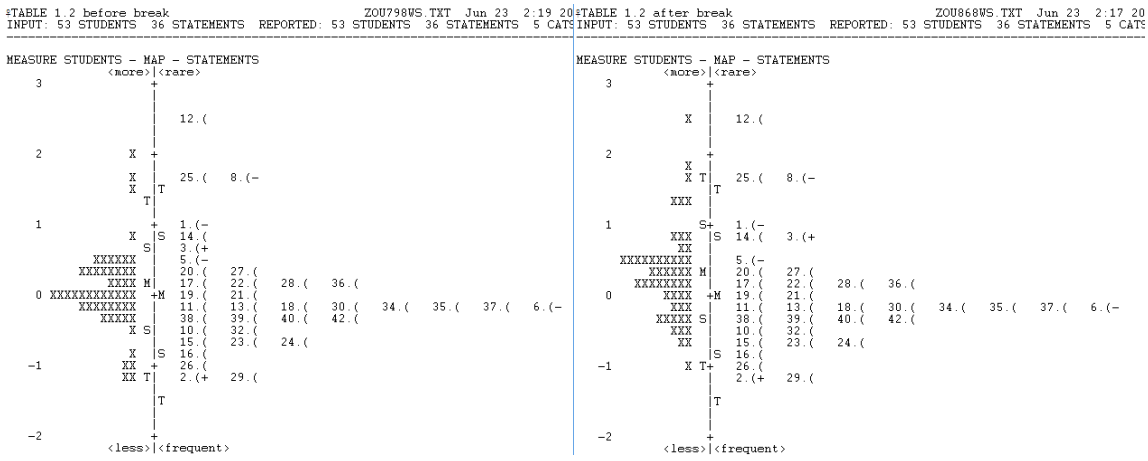

Figure 6 Wright Maps before and after break

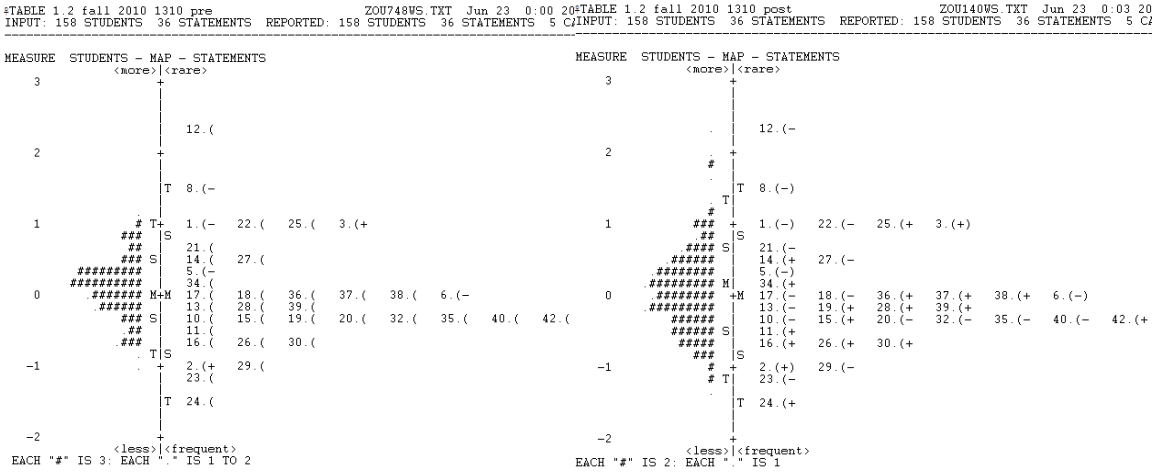

Figure 7 Wright Map Fall 2010 1310 pre/post

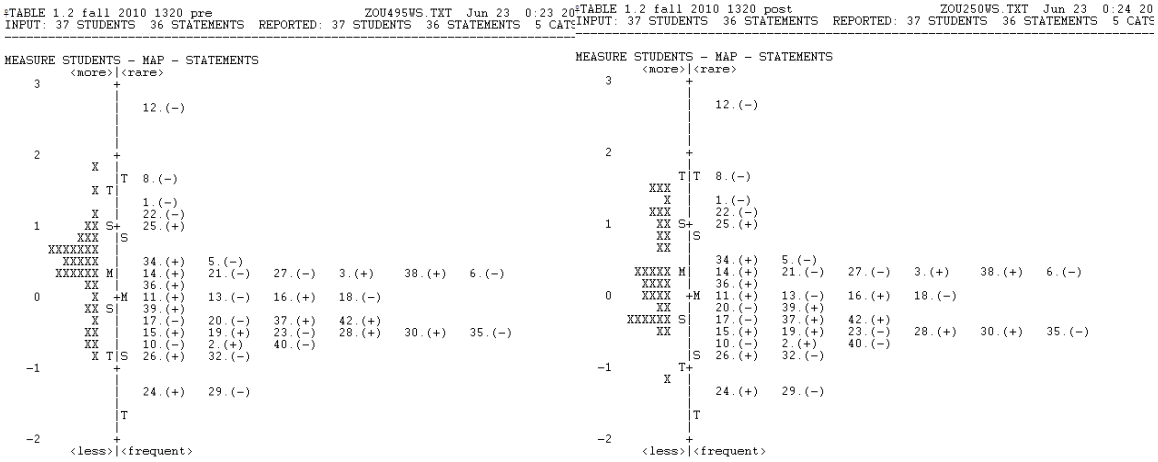

Figure 8 Wright Map Fall 2010 1320 pre/post

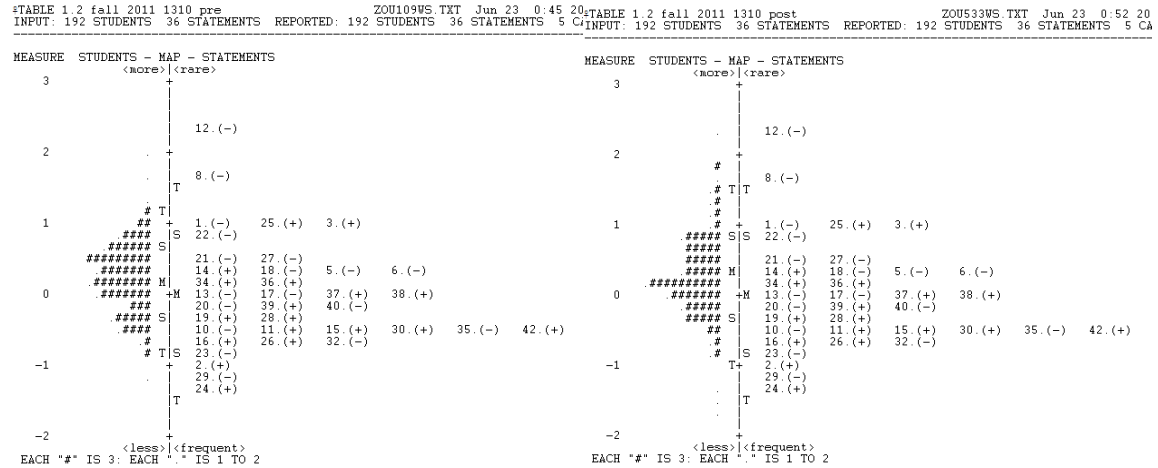

Figure 9 Wright Map Fall 2011 1310 pre/post

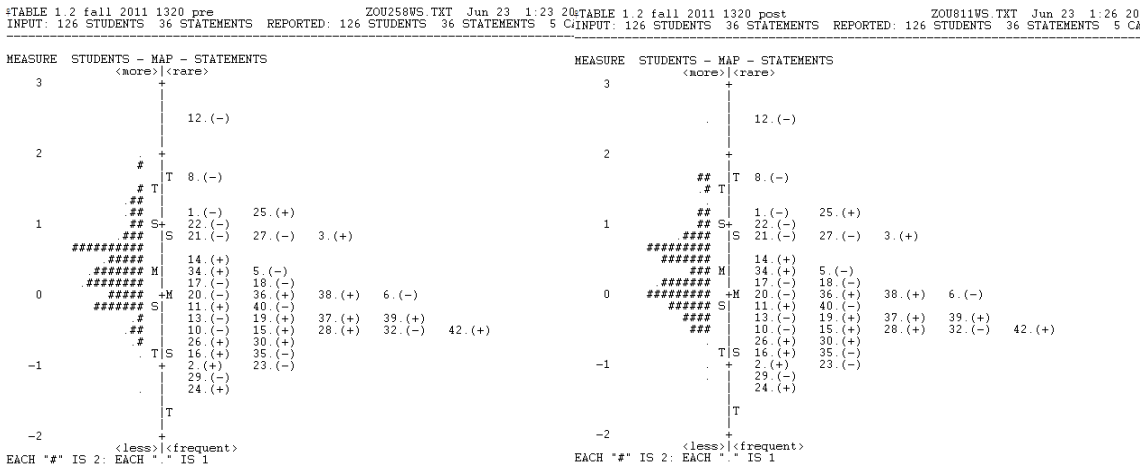

Figure 10 Wright Map Fall 2011 1320 pre/post

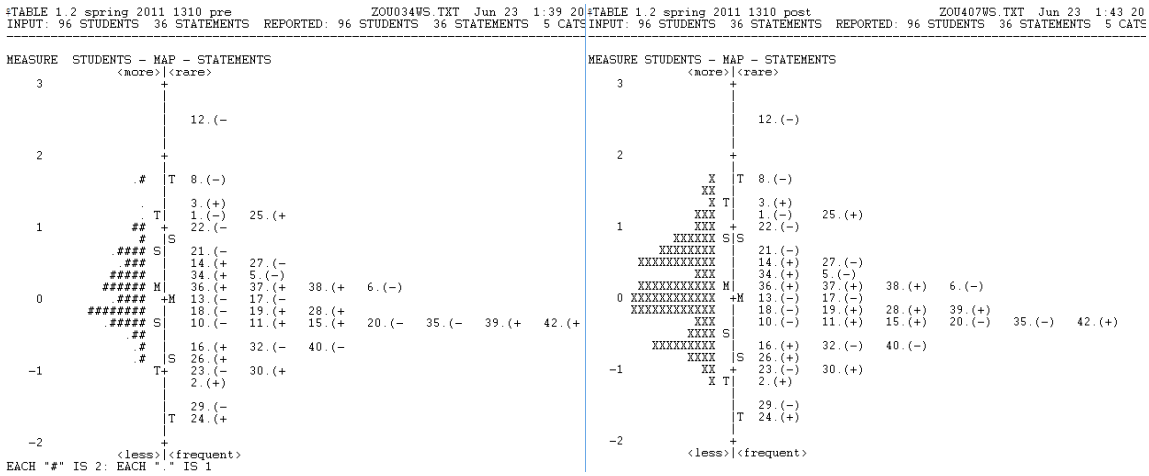

Figure 11 Wright Map Spring 2011 1310 pre/post

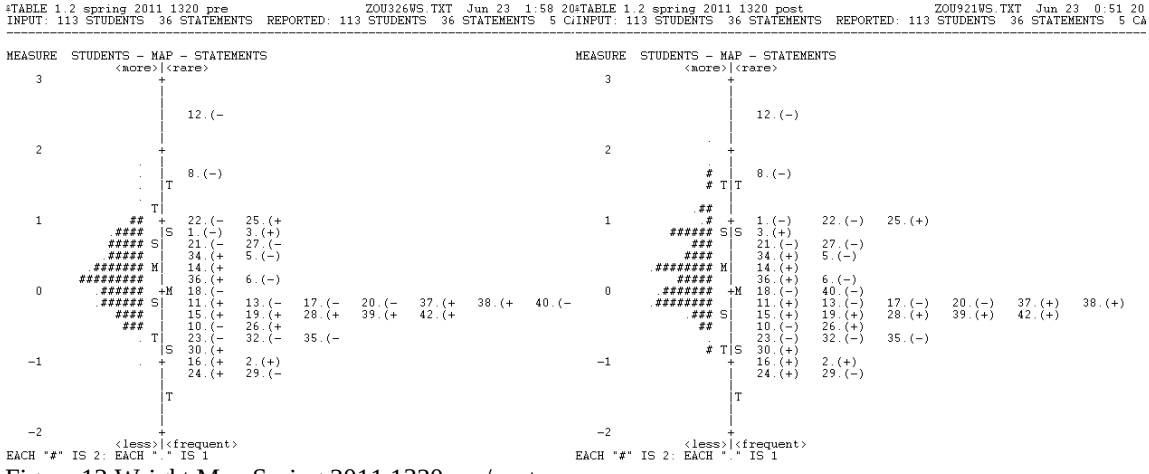

Figure 12 Wright Map Spring 2011 1320 pre/post

Explanatory note for Figure 6-12: “X” in the Wright Map only represents one student.

Table 8 Students' change in ability by term

Term	Survey	Student Mean	p-value
Winter break	Pre	.0836	.021*
	Post	.2906	
Fall 2010 1310	Pre	.0811	.690
	Post	.0960	
Fall 2010 1320	Pre	.3873	.531
	Post	.3341	
fall 2011 1310	Pre	.2362	.321
	Post	.2709	
fall 2011 1320	Pre	.3972	.505
	Post	.3602	
spring 2011 1310	Pre	.1580	.852
	Post	.1426	
spring 2011 1320	Pre	.2549	.168
	Post	.3104	

Explanatory note for Table 9:

*$p<0.05$ means the change is significant, if the significance level is 5%.

To check whether there is significant change in students' attitude, this study conducted the Paired Sample t Tests for each term, and found no significant change had taken place during any long semester. The winter break, however, significantly influenced in students' attitude (see Table 9). These results are consistent with those obtained by the standard CLASS analysis.

Students' interpretation of the statements in CLASS may change due to a semester's learning of Physics. This invalidates pre/post comparison for the statement in question. To assess whether this happened, change of students' interpretations toward each statement for each term is measured by using the Differential Item Functioning (DIF) Test in Winsteps. Rasch-Welch (logistic regression) t-test is a DIF Test for rating scale surveys. This test estimates the statements' difficulty for each person group by using a logistic regression model. The difference in the DIF Measures (which are the statements' difficulty) indicates that one group of students is scoring better than another

group of students on this statement. This could mean that different groups are having different interpretation toward the same statement.

Table 9 DIF Test by term

Term	Statement	PRE DIF MEASURE	POST DIF MEASURE	DIF CONTRAST	Rasch-Welch t
Fall 2011 1310	3.(+)	1.02	0.31	0.71	6.21
	24.(+)	-1.24	-0.74	-0.5	-3.56
	29.(-)	-1.2	-0.65	-0.54	-3.93
Spring 2011 1310	24.(+)	-1.72	-0.84	-0.88	-4.21
	37.(+)	0.22	-0.5	0.73	4.16

Explanatory notes for Table 10:

1. When a DIF contrast for a statement is > |0.5|, and the Rasch-Welch t is > |2|. Students' interpretation of this statement is considered significantly changed.

According to Table 10, students' interpretations of statements have not changed for courses during most terms. For course PHYS 1310 in the fall and spring semesters of 2011, students' interpretations for some statements have changed significantly. However, according to Linacre, if there are not a large amount of statements that have changed, or if the direction of the changes for different statements are not mostly the same, the measurements of students' ability in the pre-course survey and the post-course survey are still comparable. The changes in the fall and spring semester have met these criteria, the results obtained about those terms are therefore still reliable.

In the former study, which used the standard CLASS analysis, the researchers discussed the connections between CLASS and a survey about the transformative experience, which was develop by K. J. Pugh for assessing the engagement of classroom knowledge in students' life[28]. Transformative experience is the experience when a student

actively applies the knowledge he or she has obtained from school to everyday life, and views the world in a different and meaningful way. The levels of the transformative experience vary from merely engaging the related knowledge in classroom activities to engaging this knowledge in everyday life, which is a continuum of less transformative behavior to more transformative behavior. This survey contains three facets of transformative experience, including motivated use, expansion of perception, and experiential value. The transformative survey contains 33 statements, and some of its statements have parallel meaning with some statement in CLASS. Statement 3 on the CLASS ("I think about the physics I experience in everyday life") is similar to statement 11 on the transformative experience survey ("I find myself thinking about adaptation and/or natural selection in all kinds of everyday situations"). Statement 37 on the CLASS ("To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed") is similar to statement 12 on the transformative experience survey ("I seek out opportunities to apply my knowledge of adaptation and/or natural selection in my everyday life"). Connections are found between statement 28 on the CLASS (Learning physics changes my ideas about how the world works") and statement 19 on the transformative experience survey ("I can't help but see animals and/or plants in terms of adaptation and/or natural selection now"), and item 14 on the CLASS ("I study physics to learn knowledge that will be useful in my life outside of school") and item 26 on the transformative experience survey ("Knowledge of adaptation and/or natural selection is useful in my current, everyday life") are also very similar. In addition, statement 11 on the CLASS ("I am not satisfied until I understand why something works the way it does") may address motivated use, and statement 30

("Reasoning skills used to understand physics can be helpful to me in my everyday life" and 35 ("The subject of physics has little relation to what I experience in the real world") address experiential value. After removing the statements in CLASS which are related to the transformative experience, the students' change in attitude becomes comparably insignificant (0.075). One possible implication of this phenomenon is, during the winter break, students engage classroom physics knowledge to their everyday life to a greater extent, and it may be the main respect of students' change in attitude during the break. However, because the sample used to investigate the Winter Break Effect is small (55 students), the conclusion should be carefully drawn, and further study on it is needed.

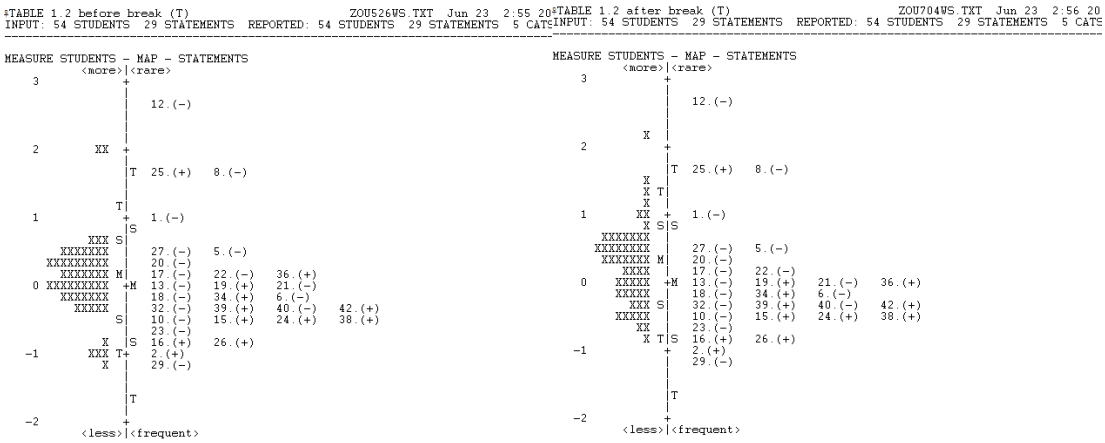

Figure 13 Wright Map before and after break (transformative related statements deleted)

Table 10 Students' change in ability (transformative related statements deleted)

Term	Survey	Student Mean	p-value
Winter break	Pre	0.1106	.075
	Post	0.2607	

4.2 Students' change in attitude due to the teaching methods

One thing that is worth discussing is that while none of the instructors were found to have a significant influence on students' attitude in the study analyzed by the CLASS analysis, the students of instructor 5, however, were found to have a significant positive change in their attitudes by using the Rasch Analysis.

Table 11 Students' change in ability by instructor

Instructor	Stage	Using percentage	Student Mean	Sig
1	Pre	(74/82)90%	.0512	.138
	Post		-.0306	
2	Pre	(205/228)90%	.1819	.173
	Post		.1167	
3	Pre	(337/377)89%	.3888	.196
	Post		.4261	
4	Pre	(47/54)87%	.1269	.286
	Post		.2806	
5	Pre	(64/70)91%	.1913	.011*
	Post		.3566	
7	Pre	(205/228)90%	.2660	.309
	Post		.1907	

Explanatory notes for Table 12

1. Instructor 6 has only 18 students' response s of CLASS. Because the sample size is too small, it is not appropriate to analyze this sample.

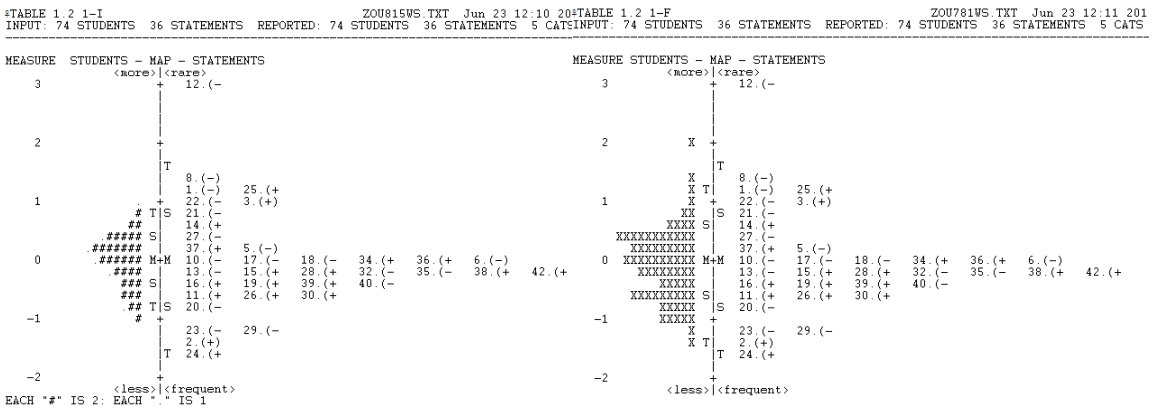

Figure 14 Wright Map for Inst. #1 pre/post

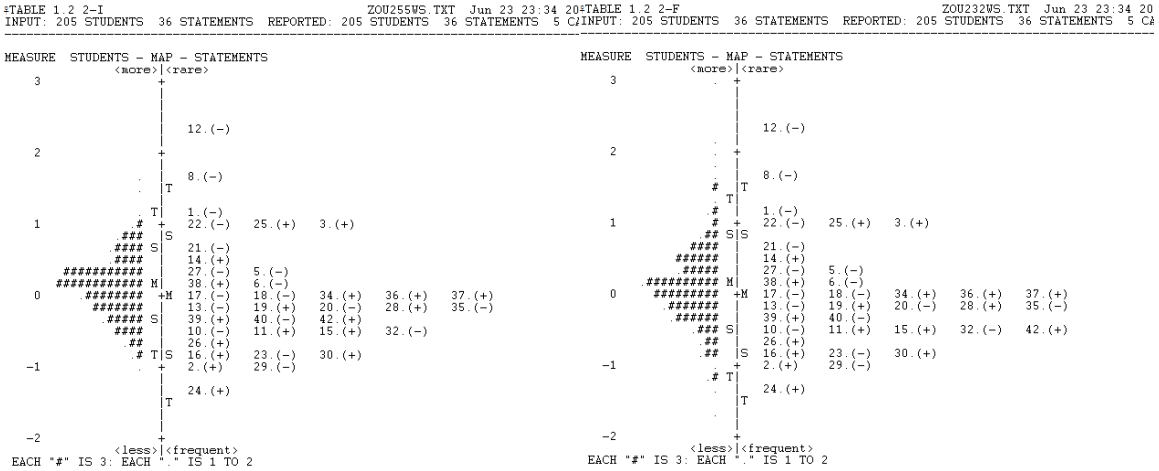

Figure 15 Wright Map for Inst. #2 pre/post

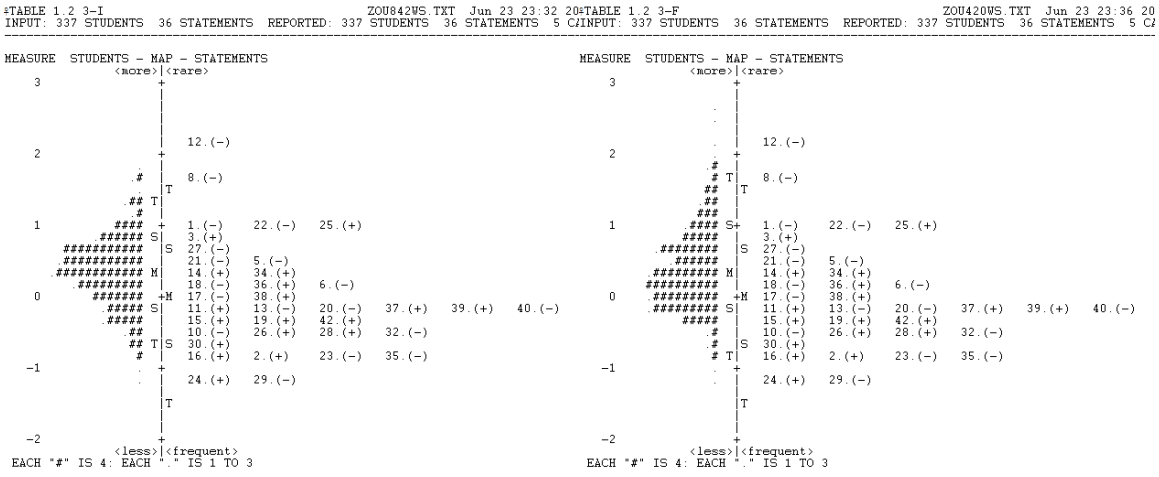

Figure 16 Wright Map for Inst. #3 pre/post

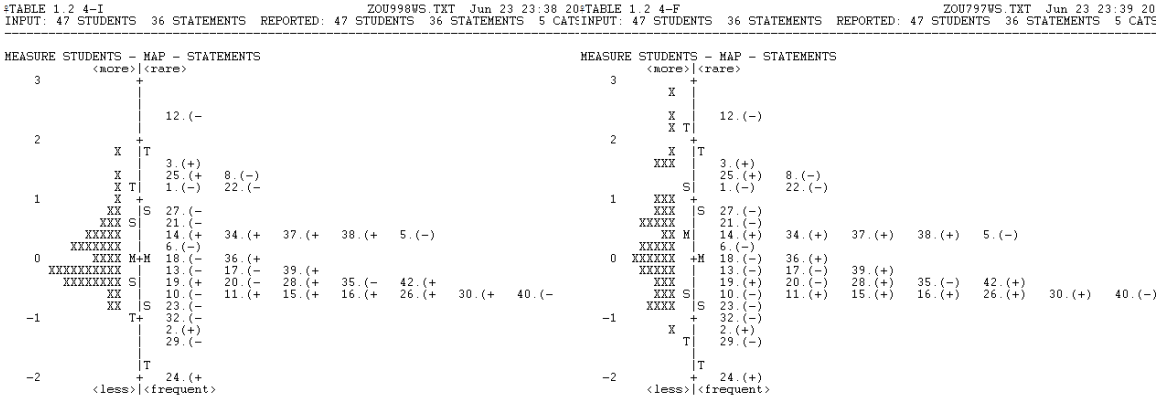

Figure 17 Wright Map for Inst. #4 pre/post

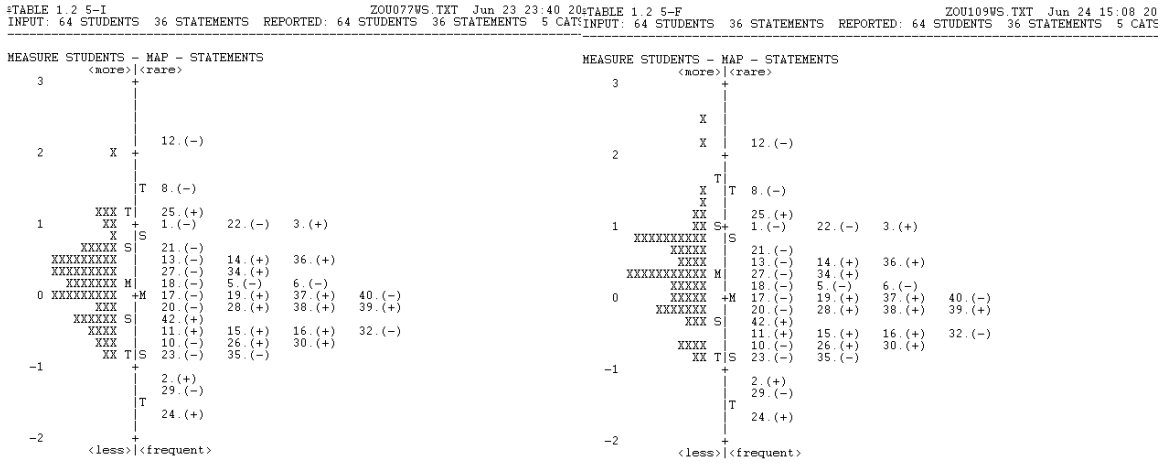

Figure 18 Wright Map for Inst. #5 pre/post

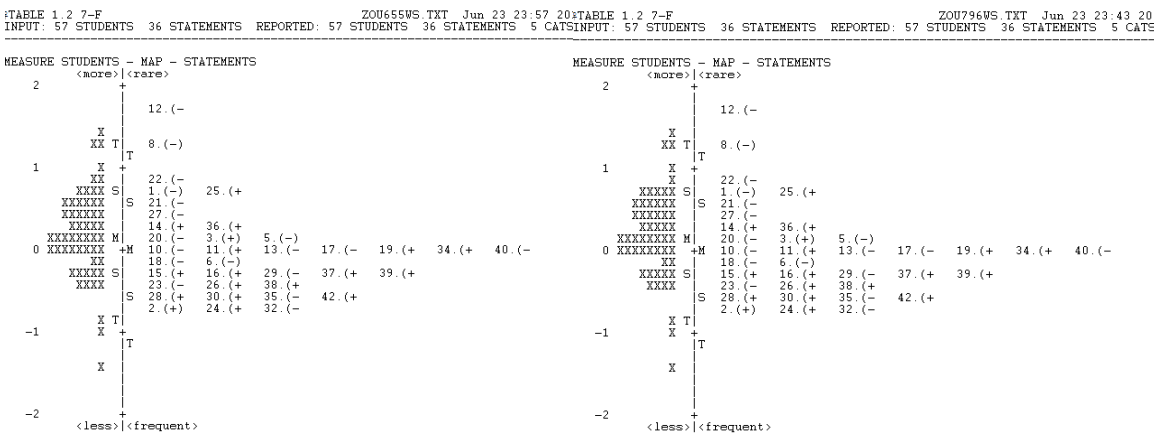

Figure 19 Wright Map for Inst. #7 pre/post

In addition, this study checked the change of students' interpretation to every statement for each instructor. Even though Instructor 4 and 5 each have six statements that had changed their meaning, the direction of the changes are not all the same. The pre/post comparison is therefore, still reliable.

Table 12 The DIF Test by instructor

Inst#	Name	PRE DIF MEASURE	POST DIF MEASURE	DIF CONTRAST	Rasch-Welch t
1	10.(-)	-0.04	-0.56	0.52	2.68
	12.(-)	2.73	1.98	0.75	2.86
2	3.(+)	0.95	0.44	0.51	4.76
	38.(+)	0.09	-0.45	0.53	4.75
3	3.(+)	0.75	0.23	0.52	6.04
4	3.(+)	1.28	0.26	1.02	4.29
	11.(+)	-0.49	0.09	-0.58	-2.45
	19.(+)	-0.35	0.2	-0.55	-2.33
	24.(+)	-1.59	-0.72	-0.87	-3.05
	37.(+)	0.28	-0.52	0.8	3.36
	38.(+)	0.31	-0.4	0.7	2.98
5	19.(+)	-0.03	-0.55	0.52	2.51
	23.(-)	-0.83	-0.24	-0.59	-2.73
	24.(+)	-1.44	-0.75	-0.69	-2.81
	29.(-)	-1.21	-0.67	-0.54	-2.29
	34.(+)	0.31	-0.16	0.46	2.37
	35.(-)	-0.7	-0.26	-0.44	-2.08
7	12.(-)	2.61	1.78	0.83	3.07
	24.(+)	-1.42	-0.76	-0.66	-2.39

4.3 Evaluation of CLASS by using Rasch Analysis

4.3.1 The design of the statements

Figure 20 is also a Wright Map. To evaluate the statement difficulty of CLASS, both pre and post survey responses from students (except the outliers) were used in make the Wright Map in Figure 20. This figure shows that CLASS has statements with different difficulties, and the range of the difficulty is broad enough to measure the varying ability of students. The measurements of the students' ability and the statements' difficulty support this opinion. The range of the students' ability is from -1.64 logits to +2.29 logits, and the range of the statements' difficulty is from -1.15 logits to +2.25 logits. Except one student (the one whose ability is +2.29 logits), other students' ability are all lower than +2.25 logits. However, for an ideal survey, the easiest statement on the Wright Map should have lower position than the least able student, and the most difficult statement should have higher position than the most able student in a standardized sample. According to the Wright Map for CLASS, a most able student in a standardized sample exhibited 50% possibility of responding the most difficult statement (statement 12) in CLASS, which indicates that the most difficult statement in CLASS is difficult enough. However, there was a noticeable gap between the most difficult statement and the second hardest statement (statement 8), which indicates that some statements, which are easier than statement 12 but more difficult than statement 8, should be designed into CLASS. In addition, even the easiest statement in CLASS is too hard to some students to

answer correctly. Some easier statements therefore need to be added into CLASS.

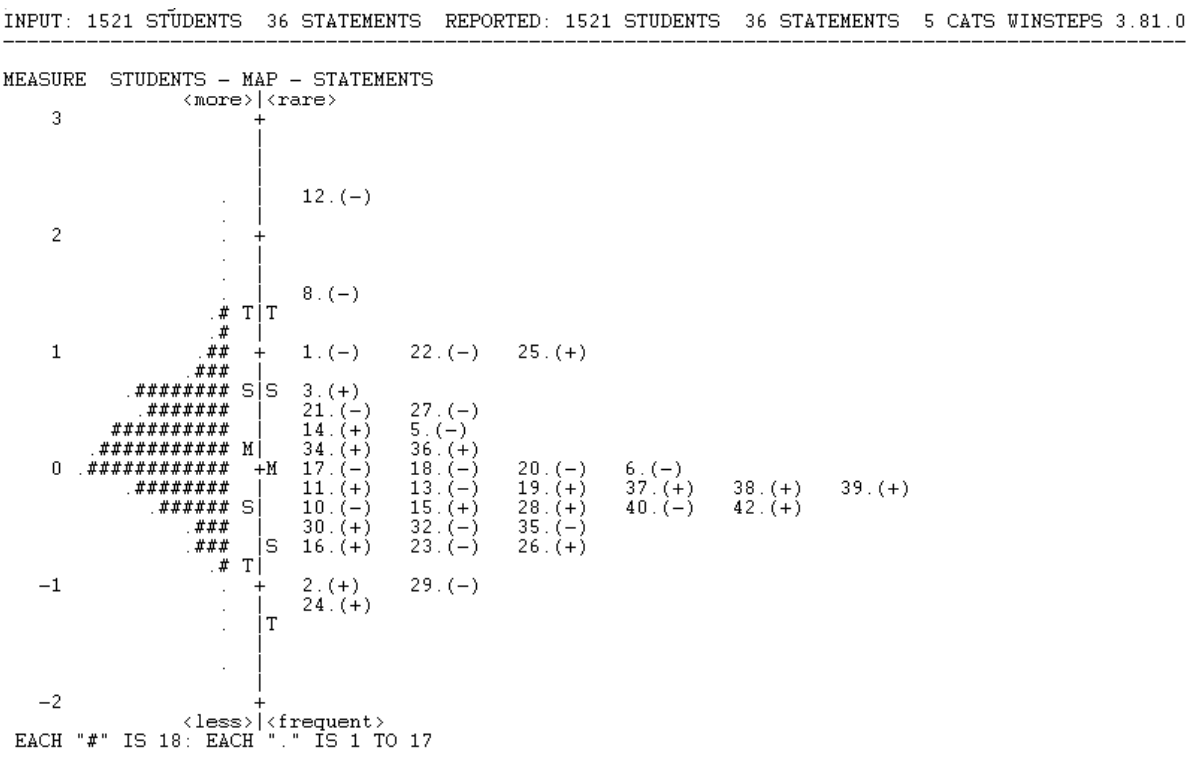

Figure 20 Wright Map of all students

4.3.2 The design of the rating scale

Linacre once proposed eight guidelines to examine whether a rating scale is functioning well. Those guidelines are demonstrated with the indices and graphs generated by Winsteps, and the guidelines have been carefully discussed in Chapter II. The rating scale of CLASS meets all the guidelines except the one about the distance between adjacent step calibrations (see Table 14). The step calibration of category 3 is only 0.4 logits apart from that of category 4, which is much less than what is required for the smallest distance between them. In Figure 21 (in page 54), the intersections of the adjacent categories are the step calibrations. The second intersection is the step calibration of category 3, and the third intersection is the step calibration of category 4. Figure 21 emphasizes the peak of the

"neutral" category is too low, which is why the two calibrations are too close to each other. Even at the peak where a student is most likely to choose "neutral", he or she shows almost the same possibility of choosing neighboring categories. Its influence on the measurement is not clear, which requires for further study. However, one possible explanation of this occurrence is that the "neutral" category is not significantly favored by students who have moderate relative ability. Some students who choose "agree" or "disagree" may actually have ability correspondending to the "neutral" category. Those "agree" and "disagree" responses are yet still counted into the students' favorable or unfavorable percentages in the standard CLASS analysis, and therefore distort the result of measurement to some extent. The problem also exists when considering "Strongly Agree" and "agree" the same responses, and "Strongly Disagree" and "disagree" the same responses (see Table 15 and Figure 22). The accuracy of the Rasch measurement is degraded as well, because the problem stems from the imperfection of the rating scale.

Table 13 The summaryummary of rating scale (five-point rating scale)

Category	Observed		Observed Average	Infit	Outfit	Andrich Threshold	Category Measure
	Count	%					
SD	2966	5	-1.14	.99	1.01	NONE	-3.29
D	10940	20	-.39	1.02	1.04	-2.06	-1.45
N	14819	27	.1	.93	.92	-.43	-.16
A	22036	40	.61	.96	.97	-.03	1.39
SA	3995	7	1.02	1.07	1.05	2.52	3.67

Table 14 The summary of rating scale (three-point scale)

Category	Observed Count	Observed %	Observed Average	Infit	Outfit	Andrich Threshold	Category Measure
SD/D	13906	25	-.46	1.02	1.08	NONE	-1.57
N	14819	27	.31	.92	.89	-.10	.00
S	26031	48	.99	1.00	1.05	.10	1.57

Explanatory Notes on Table 14 and 15

1. The sample is the students from every course over different semesters. Those students whose mean square fit statistic is >2 have been eliminated (from analysis).
2. The Andrich thresholds are the step calibrations, which are the points where the possibility of choosing either of the adjacent responses is equivalent.
3. The observed average is the average ability of the people who have responded in that category.

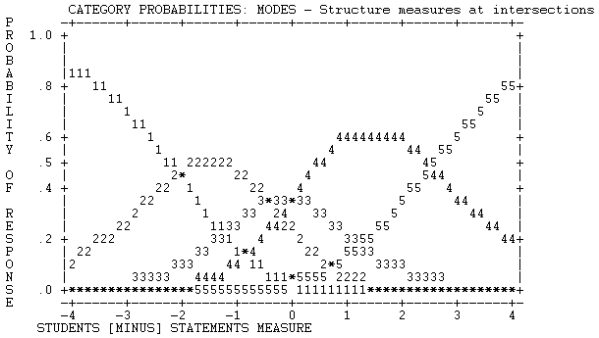

Figure 21 Category Probability Curves (five-point rating scale)

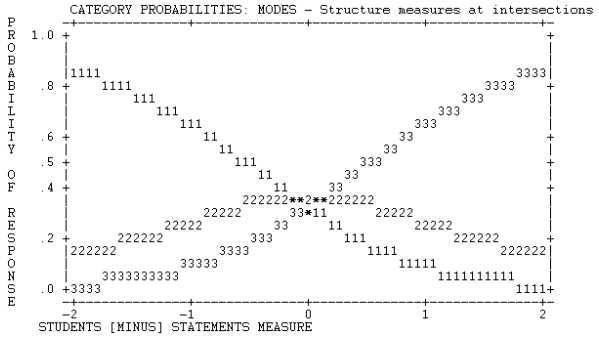

Figure 22 Category Probability Curves (three-point rating scale)

4.4 Evaluation of the standard CLASS method by using Rasch Analysis

The CLASS developers from University of Colorado, Boulder, decided the reasonable way to score students' responses is to look at their average percentage of favorable or unfavorable responses. These researchers dismissed the data collected by the category of "neutral", and they also neglected the difference between the categories of "Strongly Agree" and "agree", and the difference between the categories of "Strongly Disagree" and "disagree". This scoring process has its rationality, but also has its limitations and drawbacks.

The Rasch Analysis shows the rationality of neglecting the difference between "Strongly Agree" and "agree" in the standard analysis. In the standard analysis, the Likert scale in CLASS is considered as an ordinal scale. Usually, when a scale is considered ordinal, every response on the scale is therefore automatically assigned with ordinal meaning, which means "Strongly Agree" must be a better or a worse response than "agree", depending on what the correct response on the scale is, but they are never regarded as the same responses. The graph of Category Probability Curves (Figure 23) for the rating scale in CLASS, which is obtained by the Winsteps, however, illustrates the problem of regarding these two responses differently. For a student whose ability is 3.5 logits beyond the difficulty of the statement (which will be referred to "relative ability"), he or she has approximately a 27% probability of choosing "agree" and a 73% probability of choosing "Strongly Agree"; or, for a group of students whose ability are 3.5 logits beyond the difficulty of the statement, 27% of the them will choose "agree", and 73% of them will choose "Strongly Agree". This phenomenon implies that for a student who chooses "Strongly Agree", his or her ability on this statement may be the same as a student who

chooses “agree”. “Strongly Agree” is therefore not the same as “agree”. This conclusion obtained by the Rasch model is consistent with what the CLASS developers learnt from their interviews, which shows that students who agreed with the statement to the same extent might end up choosing different categories. As a result, when considering the Likert scale in CLASS as an ordinal scale, it is necessary to regard “Strongly Agree” and “agree” as the same responses, just as the CLASS developers decided to score students’ responses in those two categories.

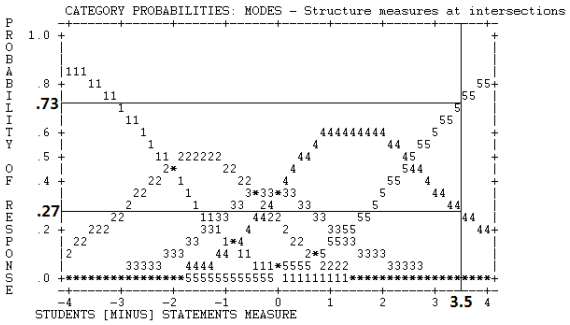

Figure 23 Category Probability Curves (ability 3.5)

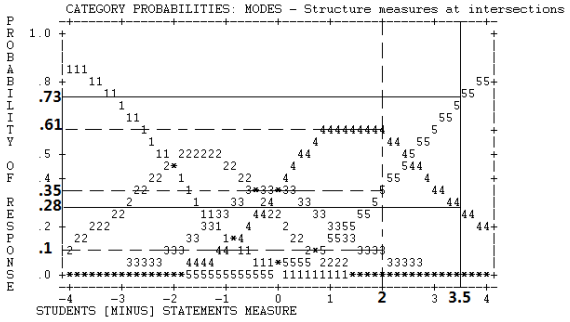

Figure 24 Category Probability Curves (ability 2 and 3.5)

However, Figure 24 also demonstrates that a student who has higher ability, the probability for him or her to choose response with a higher order is greater than a student who has lower ability. For example, a student with a relative ability of 2 logits and a student with a relative ability of 3.5 logits, both have possibilities of choosing “Strongly Agree” and “agree”. However, the student with the lower ability (2 logits) has higher chance of

choosing “agree” than “Strongly Agree”, and the student with the higher ability (3.5 logits) has higher chance of choosing “Strongly Agree” than “agree”. That is to say, the five responses are indeed different indicators of students’ ability, and the difference is inherited in the possibility of choosing each response. Regarding any two responses as the same therefore introduces a loss of information in the measurement. However, the standard CLASS analysis is not able to estimate the possibility of choosing different responses. The loss of information is therefore inevitable.

In addition, the Rasch Analysis shows additional information the standard CLASS analysis has not taken into account. That is the difference between the difficulties of statements. In the standard CLASS analysis, every statement’s contribution to students’ favorable or unfavorable percentage is identical. A favorable response to a statement which is difficult for most students to agree with is regarded the same as a favorable response to a statement which is easy for most students to agree with. Figure 20 has already shown the statements do not have the same difficulty, and in fact the difficulties vary with a considerable range. The measurement of students’ ability in standard analysis is therefore less accurate than the Rasch Analysis, which can estimate the difficulty of each statement.

V. CONCLUSIONS AND DISCUSSION

As discussed in the previous chapter, the Rasch Analysis showed an advantage over the standard analysis. The Rasch Analysis more fully used the information derived from different responses. The results thus obtain by Rasch Analysis are more accurate.

The results obtained by Rasch Analysis about students' change in attitude in each courses during different semesters and the winter break is consistent with the results obtained by the standard analysis. However, the results about the instructors' impact on students' attitude are not completely the same as the standard analysis. In the standard analysis, no instructor demonstrated significant influence on students' attitude. The Rasch Analysis, however, shows that instructor 5, who used LEPS curriculum had significantly changed students' attitude toward more expert-like attitude.

Different statements in CLASS demonstrated different difficulties in the Rasch Analysis. To improve the quality of CLASS, which can contribute to more accurate measurement, some difficult statements and some easy statements need to be added into CLASS.

In addition, further research needs to be done for seeking more effective rating scale structure for CLASS. The following content will discuss about possible methods of improving the quality of the rating scale for CLASS survey.

One possible method is to omit the neutral category and make the survey a 4-point Likert scale (Strongly Disagree, disagree, agree and Strongly Agree, or word them as disagree, slightly agree, slightly disagree and disagree). This action may force students to put more thought in their selection so that the measure of their latent trait can be more precise.

Ron has done a 4-point Likert scale and 5-point Likert scale comparison in the marketing area. He found that the unfavorable percentage significantly increased and the favorable percentage significantly dropped by 10% after deleting the neutral category[29]. Worcester and Burns once suggested that survey-participants are inclined to choose more positive categories to please survey administers, or to make themselves feel more socially acceptable. It is possible that, when students think they disagree with the "right answer", they might choose neutral stance instead.

Another direction of improving the effectiveness of the rating scale structure for CLASS is extending the response categories to 7 or more categories (for example, strong disagree, disagree, slightly disagree, neutral, slightly agree, agree and Strongly Agree). Omitting neutral stance and forcing students to make a choice may result in students skipping statements [30], or students with no preference reporting unreal attitudes. However, by increasing the number of categories from 5 to 7, the usage of neutral stance will possibly decrease. Students with slightly dissentient opinions can choose slightly disagree on the 7-point Likert scale instead of forcing themselves selecting disagree, or more likely, skip the statement or select neutral stance conservatively. Therefore, extending the rating scale is also a worth-studying direction.

Linacre has claimed that "Unless the rating scales which form the basis of data collection are functioning effectively, any conclusions based on those data will be insecure[27]." Due to the fact that the data were collected by the 5-point scale CLASS, the only action that can be taken to change the rating scale is collapsing it into 3-point scale. However, no matter which scale is being used, the problematic neutral data will always be employed in the present Rasch Analysis on CLASS.

The standard CLASS analysis has two important merits. Even though the survey was designed using a 5-point Likert scale, in the scoring process, the Strongly Disagree and disagree are combined (similar for Strongly Agree and agree). The necessity of taking this step is not only proved by the CLASS developers' interviews toward students, it's also supported by the category probability curves in Rasch Analysis. The standard analysis doesn't take the neutral data into account. The influence from students who are not able to explicitly state their preferences are naturally dismissed in this process. This procedure avoided using problematic data, which is not achievable in Rasch Analysis before the appropriate rating scale being found for CLASS.

However, from the category probability curves, it is clear that the five categories are behaving differently. Combining categories will result in loss of information, making the measurement less precise. If an appropriate rating scale for CLASS can be found, more accurate measurement on interval level will bring the analysis on CLASS to a more accurate and informative level.

APPENDIX SECTION

EBAPS

Sample statement and answer options

Part 1

A: Strongly Disagree	B: Somewhat Disagree	C: Neutral	D: Somewhat Agree	E: Strongly Agree

1. Tamara just read something in her science textbook that seems to disagree with her own experience. But to learn science well, Tamara shouldn't think about her own experiences; she should just focus on what the book says.

Part 2

19. Scientist are having trouble predicting and explaining the behavior of thunder storms. This could be because thunder storms behave according to a very complicated or hard-to-apply set of rules. Or, that could be because some thunder storm don't behave consistently according to *any* set of rules, no matter how complicated and complete that set of rules is.

In general, why do scientists sometimes have trouble explaining things? Please read all options before choosing one.

(a) Although things behave in accordance with rules, those rules, those rules are often complicated, hard to apply, or not fully known.
(b) Some things just don't behave according to a consistent set of rules.
(c) Usually it's because the rules are complicated, hard to apply, or unknown; but sometimes it's because the thing doesn't follow rules.
(d) About half the time, it's because the rules are complicated, hard to apply, or unknown; and half the time, it's because the thing doesn't follow rules.
(e) Usually it's because the thing doesn't follow rules; but sometimes it's because the rules are complicated, hard to apply, or unknown.

Part 3

24.

Brandon: A good science textbook should show how the material in one chapter relates to the material in other chapters. It shouldn't treat each topic as a separate "unit," because they're not really separate.

Jamal: But most of the time, each chapter is about a different topic, and those different topics don't always have much to do with each other. The textbook should keep everything separate, instead of blending it all together.

With whom do you agree? Read all the choices before circling one.

(a) I agree almost entirely with Brandon.
(b) Although I agree more with Brandon, I think Jamal makes some good points.
(c) I agree (or disagree) equally with Jamal and Brandon.

(d) Although I agree more with Jamal, I think Brandon makes some good points.
(e) I agree almost entirely with Jamal.

Scoring scheme:

1. A=4, B=3, C=1, D=0.5, E=0

19. A=4, B=0, C=3, D=2, E=1

24. A=4, B=4, C=2, D=1, E=0

CLASS

This survey will be used to assess the influence of physics courses on students' attitudes toward learning science. Responses will be kept confidential.

Please bubble in your name and Texas State ID number in the appropriate spaces on the Scantron form.

Here are a number of statements that may or may not describe your beliefs about learning physics. You are asked to rate each statement by circling a letter between a and e where the letters mean the following:

a. Strongly Disagree

b. Disagree

c. Neutral

d. Agree

e. Strongly Agree

Choose one of the above five choices that best expresses your feeling about the statement. If you don't understand a statement, leave it blank. If you understand, but have no strong opinion, choose c.

Survey

1. A significant problem in learning physics is being able to memorize all the information I need to know. (-)

2. When I am solving a physics problem, I try to decide what would be a reasonable value for the answer. (+)

3. I think about the physics I experience in everyday life. (+)

4. It is useful for me to do lots and lots of problems when learning physics. (+)

5. After I study a topic in physics and feel that I understand it, I have difficulty solving problems on the same topic. (-)

6. Knowledge in physics consists of many disconnected topics. (-)

7. As physicists learn more, most physics ideas we use today are likely to be proven wrong. (-)

8. When I solve a physics problem, I locate an equation that uses the variables given in the problem and plug in the values. (-)

9. I find that reading the text in detail is a good way for me to learn physics. (+)

10. There is usually only one correct approach to solving a physics problem. (-)

11. I am not satisfied until I understand why something works the way it does. (+)

12. I cannot learn physics if the teacher does not explain things well in class. (-)

13. I do not expect physics equations to help my understanding of the ideas; they are just for doing calculations. (-)

14. I study physics to learn knowledge that will be useful in my life outside of school. (+)

15. If I get stuck on a physics problem my first try, I usually try to figure out a different way that works. (+)

16. Nearly everyone is capable of understanding physics if they work at it. (+)

17. Understanding physics basically means being able to recall something you've read or been shown. (-)

18. There could be two different correct values to a physics problem if I use two different approaches. (-)

19. To understand physics I discuss it with friends and other students. (+)

20. I do not spend more than five minutes stuck on a physics problem before giving up or seeking help from someone else. (-)

21. If I don't remember a particular equation needed to solve a problem on an exam, there's nothing much I can do (legally!) to come up with it. (-)

22. If I want to apply a method used for solving one physics problem to another problem, the problems must involve very similar situations. (−)

23. In doing a physics problem, if my calculation gives a result very different from what I'd expect, I'd trust the calculation rather than going back through the problem. (-)

24. In physics, it is important for me to make sense out of formulas before I can use them correctly. (+)

25. I enjoy solving physics problems. (+)

26. In physics, mathematical formulas express meaningful relationships among measurable quantities. (+)

27. It is important for the government to approve new scientific ideas before they can be widely accepted. (-)

28. Learning physics changes my ideas about how the world works. (+)

29. To learn physics, I only need to memorize solutions to sample problems. (-)

30. Reasoning skills used to understand physics can be helpful to me in my everyday life. (+)

31. We use this statement to discard the survey of people who are not reading the questions. Please select agree-option d (not Strongly Agree) for this question to preserve your answers.

32. Spending a lot of time understanding where formulas come from is a waste of time. (-)

33. I find carefully analyzing only a few problems in detail is a good way for me to learn physics. (+)

34. I can usually figure out a way to solve physics problems. (+)

35. The subject of physics has little relation to what I experience in the real world. (-)

36. There are times I solve a physics problem more than one way to help my understanding. (+)

37. To understand physics, I sometimes think about my personal experiences and relate them to the topic being analyzed. (+)

38. It is possible to explain physics ideas without mathematical formulas. (+)

39. When I solve a physics problem, I explicitly think about which physics ideas apply to the problem. (+)

40. If I get stuck on a physics problem, there is no chance I'll figure it out on my own. (-)

41. It is possible for physicists to carefully perform the same experiment and get two very different results that are both correct. (-)

42. When studying physics, I relate the important information to what I already know rather than just memorizing it the way it is presented. (+)

Finally, we would like some additional information:

43. What physics course are you currently enrolled in?

    a. Physics 1310
    b. Physics 1320

44. Would you be willing to be interviewed about your responses to this survey?

    a. Yes
    b. No

The “+” means experts Strongly Agree or agree with this statement, and the “-” means experts Strongly Disagree or disagree with this statement.

REFERENCES

1 W. Adams, K. Perkins, N. Podolefsky, M. Dubson, N. Finkelstein and C. Wieman, Physical Review Special Topics - Physics Education Research 2, (2006).

2 P. Zhang and L. Ding, Physical Review Special Topics - Physics Education Research 9, (2013).

3 N. Songer and M. Linn, J. Res. Sci. Teach. 28, (1991).

4 D. Hammer, Cognition And Instruction 12, (1994).

5 L. Lising and A. Elby, Am. J. Phys. 73, (2005).

6 M. Henerson, L. Morris and C. Fitz-Gibbon, *How To Measure Attitudes* (Sage Publications, Beverly Hills, Calif., 1978).

7 M. Wu and R. Adams, *Applying The Rasch Model To Psycho-Social Measurement* (Educational Measurement Solutions, Melbourne, 2007).

8 D. Hammer, Cognition And Instruction 12, (1994).

9 L. Lising and A. Elby, Am. J. Phys. 73, (2005).

10 J. House, The Journal Of Social Psychology 135, (1995).

11 I. Halloun and D Hestenes, Sci. Educ. Netherlands. 6, 553 (1998).

12 Physics.Umd.Edu (2016).

13 E. Redish, Am. J. Phys. 66, (1998).

14 M. Sahin and N. Yorek, Sci. Res. Essays. 4, 753 (2009).

15 Physics.Umd.Edu (2016).

16 E. Gire, B. Jones and E. Price, Phys. Rev. ST Phys. Educ. Res, 5, (2009).

17 V. Otero and K. Gray, Phys. Rev. ST Phys. Educ. Res, 4, (2008).

18 E. Brewe, L. Kramer and G. O'Brien, Phys. Rev. ST Phys. Educ. Res, 5, (2009).

19 B. Lindsey, L. Hsu, H. Sadaghiani, J. Taylor and K. Cummings, Phys. Rev. ST Phys. Educ. Res, 8, (2012).

20 Onlinestatbook.Com (2016).

21 Wikipedia (2016).

22 R. Lamb, L. Annetta, J. Meldrum and D. Vallett, Int J Of Sci And Math Educ 10, (2011).

23 I. Neumann, K. Neumann and R. Nehm, International Journal Of Science Education 33, (2011).

24 S. Glynn, J. Res. Sci. Teach. 49, (2012).

25 T. Sondergerd and C. Johnson, Science Education 98, (2014).

26 J. Sick, Jalt.Org (2016).

27 M. Linacre, J. Appl. Meas. 1, 85 (2002).

28 K. Pugh, Teachers College Record 104, (2002).

29 R. Garland, Marketing bulletin, 66 (1991).

30 M. Matell and J. Jacoby, Journal Of Applied Psychology 56, (1972).