Using Rasch Modeling to Analyze Standards-Based Assessment Items Aligned to Middle School Chemistry Ideas

Cari F. Herrmann-Abell, George E. DeBoer, and Jo Ellen Roseman

AAAS Project 2061

Poster presented at the DR-K12 PI Meeting
November 8, 2009
Washington, DC

Background

Project 2061 is designing a process for creating clusters of multiple choice assessment items to test middle school students' understanding of science ideas from AAAS's *Benchmarks for Science Literacy* and the National Research Council's *National Science Education Standards*. Items are being developed in 16 topic areas from the life, physical, and earth sciences; mathematics; and the nature of science. Distractor-driven multiple-choice items make use of common misconceptions as distractors, which allows users to simultaneously test for correct science ideas and documented misconceptions. Items expect students to make use of a variety of cognitive skills including: recognizing the truth of scientific facts and principles, and using the targeted ideas to explain, predict, and analyze phenomena. In this poster, we present the steps in our item development process and a summary of how we use Rasch Modeling to analyze the items.

Steps in the Item Development Procedure

The work presented here is part of a larger project funded by the National Science Foundation to develop items that are precisely aligned with national content standards. Each item is developed using a procedure designed to evaluate an item's match to important science ideas and its overall effectiveness as an accurate measure of what students do and do not know about those ideas. The steps in the development procedure are outlined below. The poster focuses on the highlighted steps**.**

- Select target content standards
- **Tease apart content standards into finer-grained key ideas**
- Create assessment maps that visually portray the relationships among target ideas
- Write clarification statements for each key idea that define the boundaries of knowledge that will be tested
- Review the literature on student learning to identify misconceptions that students may have and that can be used as distractors in test questions
- Interview students to fill in gaps in the research literature regarding ideas students have
- Develop assessment tasks using the criteria of *necessity* and *sufficiency* to align items to content standards and improve construct validity of items by paying attention to comprehensibility, test-wiseness, and appropriateness of task context

- Review and revise items internally
- Pilot test items with a sample of approximately 150 students per item
- Formally review items using teams of external and interval reviewers
- **Use Rasch modeling to compare the difficulty of items and determine a discrimination measure for each answer choice.**
- Revise items based on results of reviewer comments and results of pilot testing
- **Field test with a national sample of 2000 students per item**
- Analyze field test results and upload items and item characteristics to our Web site

Progress So Far

As of the Fall of 2009, field testing has been completed for the following topics.

- Atoms, Molecules, and States of Matter
- Control of Variables
- Properties of Substances, Chemical Reactions, and Conservation of Matter
- Force and Motion
- Plate Tectonics
- Matter and Energy Transformations in Living Systems
- Interdependence in Living Systems
- Forms of Energy
- Weathering and Erosion
- Basic Functions in Humans
- Sexual Reproduction, Genes, and Heredity

The following topics have been pilot tested and will be field tested in the Spring of 2010.

- Weather and Climate
- Cross-Cutting Themes: Models
- Energy Transformations, Energy Transfer, and Conservation of Energy
- Cell Structure and Function
- Evolution and Natural Selection

Identify Key Ideas Based On Content Standards

The ideas presented here are based on Chapter 4, Section D, of *Benchmarks for Science Literacy* (BSL) (AAAS, 1993), and Physical Science Content Standard B of the *National Science Education Standards* (NSES) (NRC, 1996).

AM-A: All matter is made up of atoms.
AM-B: All atoms are extremely small.
AM-E: All atoms and molecules are in constant motion.
AM-I: For any single state of matter, the average speed of the atoms or molecules increases as the temperature of a substance increases and decreases as the temperature of a substance decreases.

AM-G: For any single state of matter, changes in temperature typically changes the average distance between atoms or molecules. Most substances or mixtures of substances expand when heated and contract when cooled.

AM-F: There are differences in the spacing, motion, and interaction of atoms and molecules that make up solids, liquids, and gases.

AM-H: Changes of state can be explained in terms of changes in the arrangement, motion, and interaction of atoms and molecules.

SC-A: A pure substance has characteristic properties, such as density, a boiling point, and solubility, all of which are independent of the amount of the substance and can be used to identify it.

SC-C: Many substances react chemically in predictable ways with other substances to form new substances with different characteristic properties.

SC-D: When substances interact to form new substances, the atoms that make up the molecules of the original substances rearrange into new molecules.

SC-G: Whenever substances interact with one another, regardless of how they combine or break apart, the total mass remains the same.

SC-H: Whenever atoms interact with each other, regardless of how they are arranged or rearranged, the number of each kind of atom stays the same and, therefore, the total mass stays the same.

Field Test Items

Revised items are field tested on a national sample of approximately 2000 students to:

- determine psychometric properties of the items and clusters of items;
- construct difficulty scores, item discrimination indices, and differential item functioning scores; and
- describe student understanding of the targeted science ideas.

The results presented here are from two field tests involving over 7000 middle school students in grades 6 through 8 that were administered in the Spring of 2007 and the Spring of 2008. Each item was taken by approximately 2000 students. Additional testing was done during the Fall of 2009 at a highly selective private boarding school located in the Northeast with over 500 students in grades 9 through 12. Students at two universities (a public university in a southern state and a public university in a northeastern state) were also tested in 2008. These students included 474 students who had taken high school chemistry and were enrolled in a college level introductory chemistry course but who had not yet had any instruction at the college level, and over 1200 students who had had at least one semester of college level chemistry instruction.

Rasch Modeling

Rasch modeling is used to estimate and compare item difficulty and the popularity of each answer choice for students of differing ability (Linacre, 2006). The item - person map below shows the distribution of student abilities on the left and the item difficulties on the right. When item difficulty and person ability match, the person has a 50% chance of answering the item

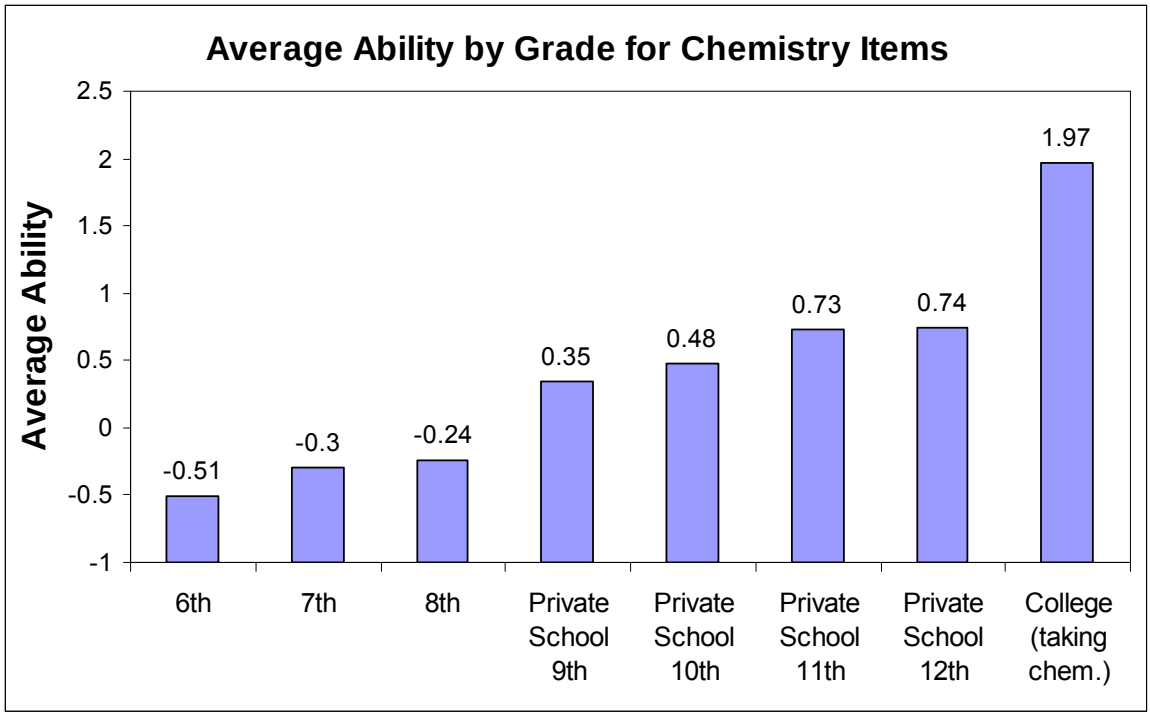
correctly. Item difficulties are shown for all 91 chemistry items in our field tests. The items are color coded by idea to illustrate the range of difficulties for items in each idea cluster.

```
        PERSONS - MAP - ITEMS
             <more>|<rare>
 4            .##  +
                   |
                .  |
               .#  |
                .  |
               .#  |
                .  |
 3                 +
             .##   |
                .  |
              ##   |
              .# T|
              .#   |
              .#   |
 2           .###  +  AM-H
             .##   |
             .##   |
             .##   |  SC-A SC-C
           .####   |  SC-G
           .### S|T
            ####   |  AM-G
 1         .####   +
          .#####   |  AM-E AM-G SC-A SC-A SC-G
         .######   |S AM-E AM-G SC-G SC-H
          .#####   |  AM-E AM-E AM-E SC-G
          .#####   |  AM-F AM-H SC-A SC-C SC-D SC-H SC-H
         .######   |  AM-B AM-F AM-F AM-H SC-A SC-C SC-G SC-H
        .#######   |  AM-A AM-G AM-G AM-G AM-H AM-I AM-I SC-A SC-A SC-D SC-D SC-G SC-G SC-H
 0    .######### M+M AM-H AM-H SC-C SC-D
      .#########   |  AM-F AM-G SC-H
       .########   |  AM-F AM-G AM-G AM-G AM-H SC-A SC-A SC-A SC-H
    .###########   |  AM-I SC-A SC-A SC-C SC-D SC-D
     .##########   |  AM-A SC-A SC-A SC-C SC-H
      .#########   |S AM-I AM-I SC-A SC-C
   .############   |  AM-A AM-F AM-F AM-I SC-C
-1    .#########   +  AM-A AM-B AM-B SC-C
        ######## S|   AM-B AM-F
         .######   |T SC-C
           .####   |  SC-G
           .####   |
             .##   |
              .#   |
-2            .#   +
               .   |
               .   |
               . T|
               .   |
               .   |
               .   |
-3             .   +
                   |
                   |
               .   |
               .   |
               .   |                        M = mean ability, mean difficulty
                   |                        S = 1 standard deviation away from mean
-4             .   +                        T = 2 standard deviations away from mean
            <less>|<frequ>                   Each '#' is 49 students.
```

The table below lists the chemistry ideas to which the items are aligned in order of difficulty based on the average of the item measures. Students were least successful on questions about atomic motion, changes of state, conservation of mass, and thermal expansion. They were most successful on questions about the idea that all matter is made up of atoms and the idea that atoms are extremely small.

Idea		Average Measure
AM-B	All atoms are extremely small.	-0.69
AM-A	All matter is made up of atoms.	-0.58
AM-I	Average speed of atoms increases as temperature increases	-0.40
AM-F	States of Matter	-0.28
SC-C	Chemical Reactions (Substance level)	-0.25
SC-D	Chemical Reactions (Atomic level)	-0.03
SC-A	Characteristic Properties of Substances	0.04
SC-H	Conservation of Mass (Atomic level)	0.12
AM-G	Thermal Expansion	0.25
SC-G	Conservation of Mass (Substance level)	0.34
AM-H	Changes of State	0.35
AM-E	All atoms are in constant motion.	0.64

The bar graph below shows the average ability in logits of the students in each of the different grade levels as measured by the entire set of 91 items. There is an increase in ability with increasing grade-level. Knowledge of the topic increases slowly throughout middle school in the national sample, and it increases slowly throughout high school in the single high school that was tested. The larger increases between middle school and ninth grade and between high school and college can be attributed to the greater selectivity of the sample at those points.

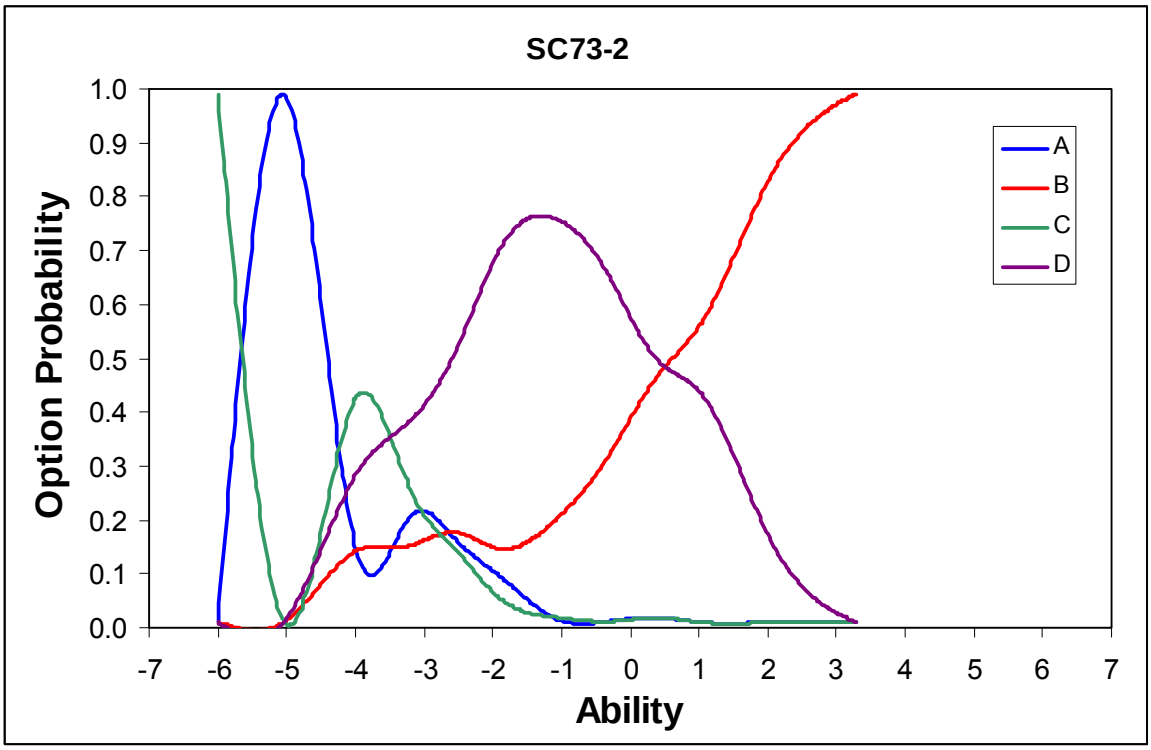


Distractor Analysis

Winsteps was used to obtain option probability curves for each item (Linacre, 2006). The option probability curves show the probability of selecting each answer choice as a function of student ability. For traditional multiple choice items, the curve for the correct answer typically increases monotonically with increasing ability, and the curve for the set of distractors typically decreases monotonically with increasing ability (Haladyna, 1994). In dichotomous scoring (correct-incorrect) the incorrect answer choices are lumped together and no attempt is made to determine the difference in discriminating ability of those incorrect answers. However, it has also been shown that, for distractor-driven items like the ones we developed, the curves do not match the monotonic behavior of traditional items (Sadler, 1998). The unique shape of the curves for distractor-driven items allows for a more precise analysis.

We use the option probability curves because they present a visual image of the distribution of correct answers and misconceptions across a wide spectrum of student ability (ranging in this case from 6th grade to college students who are taking chemistry). This enables us to see if the shape of the curves matches our expectations or if there is something unusual that could indicate a structural problem with an item. The shape of the curves also may be suggestive of a hierarchy of misconceptions which disappear in sequence as students become more knowledgeable about a topic either through out-of-school experience or through formal instruction. Below are examples of option probability curves for four chemistry items in our study.

Example 1: The nature of chemical reactions



Item SC73-2

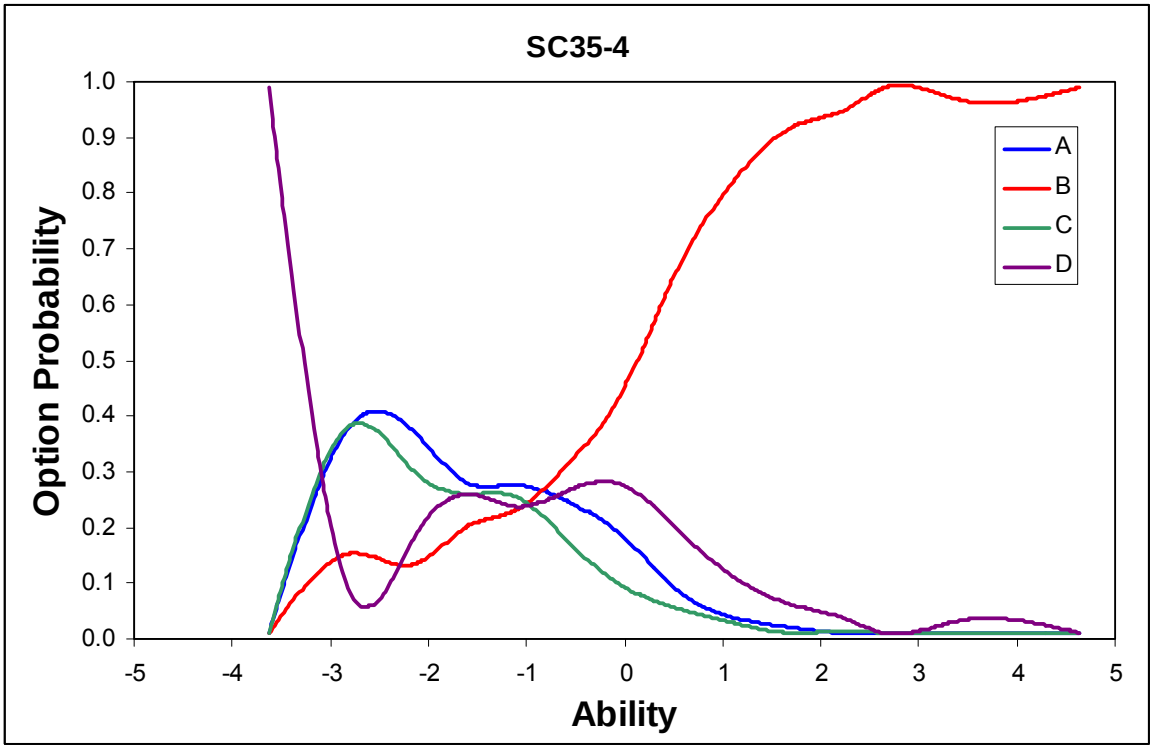
Which of the following statements about chemical reactions is TRUE?

A. All chemical reactions are dangerous. (8%)
B. Some chemical reactions can be reversed. (24%)*
C. Chemical reactions occur only in a laboratory. (7%)
D. Chemical reactions require starting with at least two substances. (61%)

Students of low ability are more likely to choose either answer choice A (all reactions are dangerous) or C (reactions only occur in a lab). The probability of choosing these answer choices drops to almost 0 around a student ability of -1. Students with moderate ability are most likely to choose answer choice D (reactions require starting with 2 substances), which is the most popular distractor. The students with abilities greater than 1 are more likely to choose the correct answer (some reactions can be reversed).

This pattern of response by ability is consistent with our expectations. It is reasonable to assume that prior to instruction students' ideas about chemical reactions are based largely on impressions they gain from the popular media. These include the idea that chemical reactions are dangerous and the idea that chemical reactions occur only in a laboratory. As students gain more familiarity with actual chemical reactions during instruction, many of the reactions they encounter involve two starting materials. As instruction progresses into the upper grades, there are increasingly more opportunities for students to learn that at least some chemical reactions are reversible. It is unlikely that students would draw that conclusion from out-of-school experience alone.

Example 2: Chemical reactions (Atomic level)



Item SC35-4

Two substances interacted and a chemical reaction occurred. What happened to the atoms and molecules of the substances during the chemical reaction?

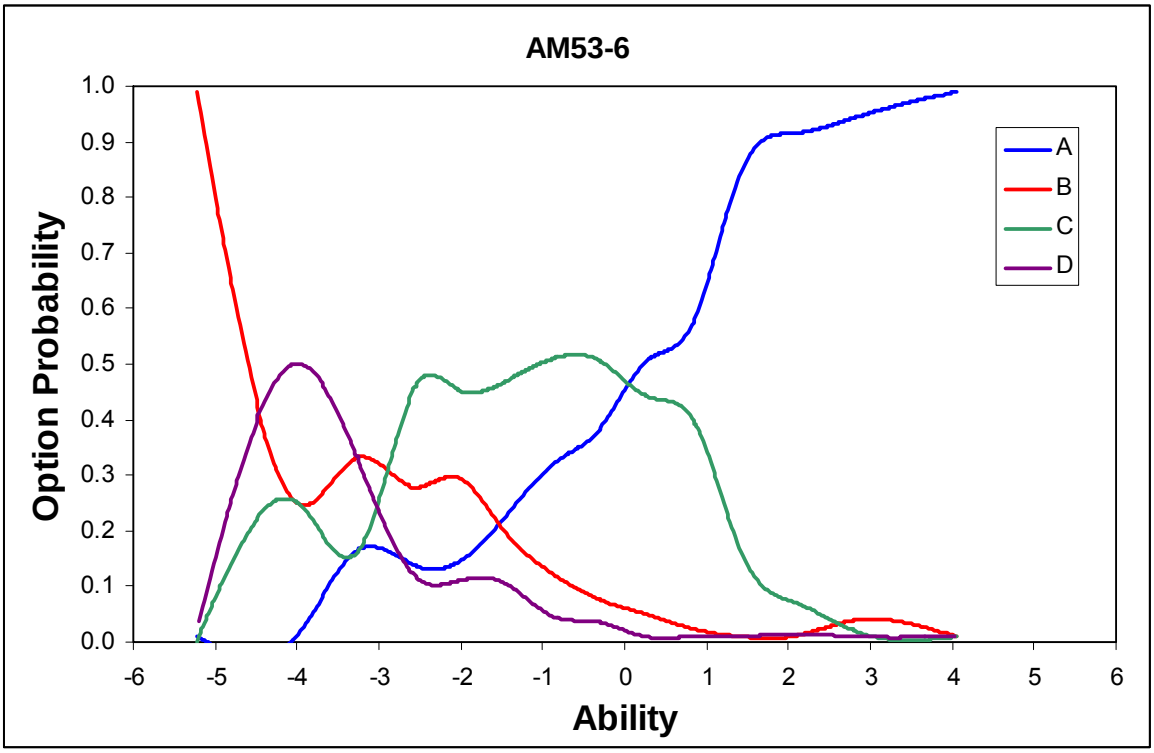
A. All of the atoms changed into new atoms and formed new molecules. (18%)
B. All of the atoms stayed the same but rearranged to form new molecules. (48%)*
C. The molecules and the atoms they are made of are the same after the chemical reaction. (14%)
D. Some atoms rearranged to form new molecules, and other atoms were changed into new atoms. (20%)

The probability of selecting the answer choice involving some atoms rearranging and some changing into new atoms is highest for the lowest ability students (D). Between ability -3 and 0, students are about equally likely to choose any of the four answer choices. For students above ability level 0, the probability of selecting the correct answer increases rapidly.

Unlike the item in example 1, this item requires a formal mental model of chemical reactions, and neither the correct answer nor the misconceptions are derived from everyday experience.

The item depends on the knowledge that that during a chemical reaction atoms rearrange to form new molecules. Students develop this idea through formal instruction. Because the item deals with a single idea, there is no logical progression involved. This is evidenced by the approximately equal selection of incorrect answers prior to a steady increase in the correct answer as instruction increases.

Example 3: All atoms are in constant motion.



Item AM53-6

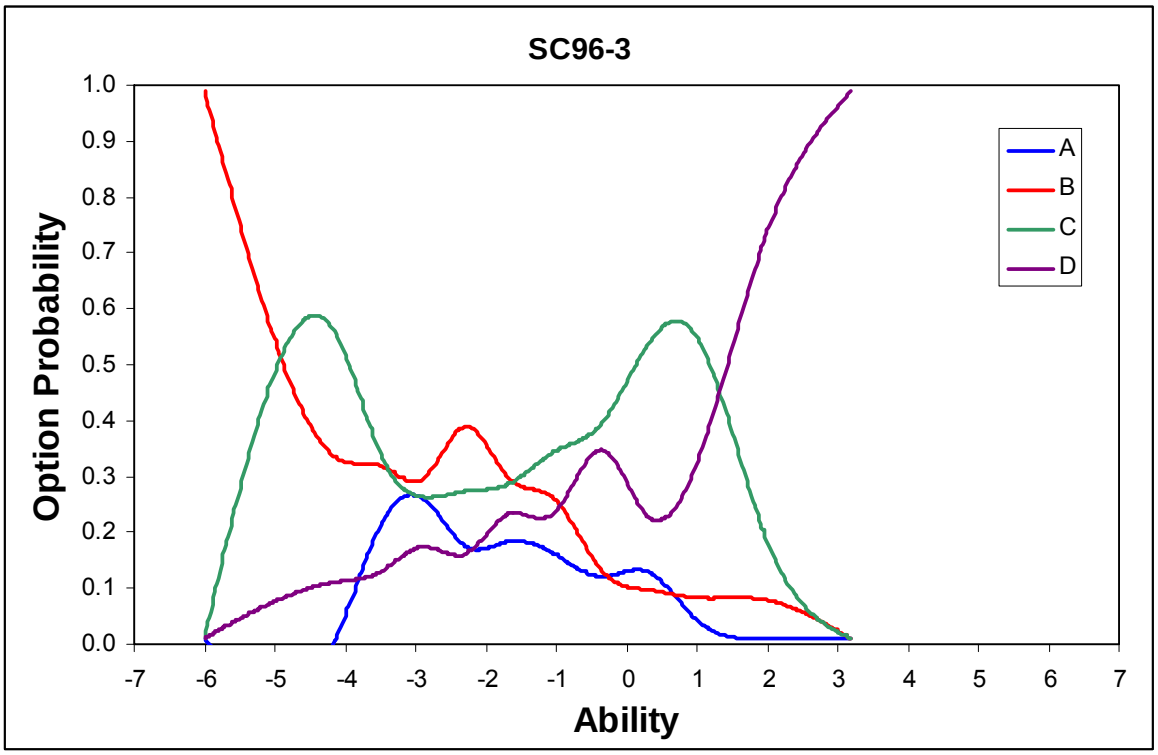
A balloon full of air is placed on a chair. Which of the following statements about the atoms and molecules of the chair and the atoms and molecules of the air in the balloon is TRUE?

A. The atoms and molecules of both the chair and the air in the balloon are moving. (39%)*
B. The atoms and molecules of both the chair and the air in the balloon are not moving. (13%)
C. The atoms and molecules of the chair are not moving, and the atoms and molecules of the air in the balloon are moving. (43%)
D. The atoms and molecules of the chair are moving, and the atoms and molecules of the air in the balloon are not moving. (6%)

Lowest-ability students tend to select the answer choice that says neither the atoms of a gas nor the atoms of a solid move (B). The answer choice that says the atoms of the solid move but the atoms of the air do not (D) peaks in probability around -4 ability and then decreases. The answer choice that says the atoms of the air move but the atoms of the solid do not (C) peaks in probability around 0 ability and then decreases. The probability of selecting the correct answer (A) (atoms of both are moving) increases as expected with increasing ability.

The pattern of responses revealed in the option probability curves is interesting for two reasons. First, the curve shows that the misconception that atoms of solids do not move is held by students over a large range of ability (from ability -3 to 0). This is a very common misconception that persists even at the higher grades. Second, the results match the expected progression of understanding. Without instruction or with limited instruction, students have no reason to think that atoms are in motion. As they learn that atoms are in motion, it is not surprising that they think atoms of some kinds of materials are in motion but not others. Finally, with increased instruction, they develop the correct idea that all atoms are in motion. The progression is from none to some to all.

Example 4: Characteristic properties of substances



Item SC96-3

You have two balloons. Balloon 1 is filled with a gas, and Balloon 2 is filled with a gas. The gases are pure substances. What could you do to help decide if the gases in the balloons are the same substance?

A. Compare the temperature of the gas in Balloon 1 to the temperature of the gas in Balloon 2. (17%)

B. Compare the volume of the gas in Balloon 1 to the volume of the gas in Balloon 2. (27%)

C. Compare the weight of the gas in Balloon 1 to the weight of the gas in Balloon 2. (32%)

D. Compare the odor of the gas in Balloon 1 to the odor of the gas in Balloon 2. (23%)*

The idea that you can use volume to identify the gas (B) drops steadily from a high point at ability -6. The idea that you can use temperature to identify the gas (A) peaks at ability -3 and then drops off steadily as ability increases. Those results are not surprising. But the curve for the answer choice in which students say that you can tell whether the gases in the two balloons are the same substance by comparing the weights of the gases (C) is unusual because it shows two distinct peaks, one at very low ability levels and the other at relatively high ability levels. The curve for the correct answer (D) increases steadily along the ability spectrum except for where it shows a dip corresponding to the second peak of answer choice C for the more able students.

We speculated that some of the more able students may have assumed that the balloons were the same size and, therefore, had the same volume. If this were the case, then comparing the weights would be a very effective, and probably the best, way to decide if the gasses in the balloons were the same substance because equal volumes of different substances usually have different weights. This is a case where the shape of the curve quickly revealed a structural problem with the item that could be corrected by saying that the balloons are different sizes.

Conclusions

Rasch Modeling was used to analyze the results from a set of 91 multiple choice assessment items aligned to middle school ideas about chemistry. For this set of items, we found that students had the most difficulty with ideas about atomic motion, changes of state, conservation of mass, and thermal expansion. The students were most successful with items testing the ideas that all matter is made up of atoms and atoms are extremely small. We also found a steady increase in understanding of chemistry from 6th grade students to college students. We examined the option probability curves for each item, which revealed hierarchies of misconceptions. Additionally, the unusual shape of one of the distractor probability curves for an item indicated a structural problem with that item.

Acknowledgements

The authors would like to acknowledge our research team: Abigail Burrows, Jean Flanagan, Francis Molina, David Pollock, Brian Sweeney, Jill Wertheim, and Ted Willard. This work is funded by the National Science Foundation (Grant No. ESI-0352473).

References

American Association for the Advancement of Science. (1993). *Benchmarks for science literacy*. New York: Oxford University Press.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, NJ: Erlbaum.

Linacre, J. M. (2006). WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com.

National Research Council. (1996). *National science education standards*. Washington, DC: National Academy Press.

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, *35*, 265-296.