Using distractor-driven standards-based multiple-choice assessments and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items

Cari F. Herrmann-Abell* and George E. DeBoer

Received 27th October 2010, Accepted 4th March 2011
DOI: 10.1039/C1RP90023D

Distractor-driven multiple-choice assessment items and Rasch modeling were used as diagnostic tools to investigate students' understanding of middle school chemistry ideas. Ninety-one items were developed according to a procedure that ensured content alignment to the targeted standards and construct validity. The items were administered to 13360 middle school, high school, and college students from across the USA, and the student data were analyzed using Rasch modeling. A cross-sectional analysis was performed to examine the progression of understanding of chemistry from middle school to college and revealed an overall increase in understanding with increasing grade level. Option probability curves for several of the items were used to provide insight into how students' thinking changes with increasing knowledge of the ideas being tested. In some cases, hierarchies of misconceptions were detected in which the misconceptions decrease in a series as the overall student performance level increases. Additionally, for one item, the peculiar shape of the option probability curve for a distractor indicated a flaw in the item.

Keywords: assessment, misconceptions, Rasch modeling

Introduction

In response to growing concerns in the United States of America about the quality of science assessment items and their alignment to standards (American Federation of Teachers, 2006), Project 2061 has been involved in a multi-year, NSF-funded project to develop multiple-choice assessment items for middle school science topics that are precisely aligned with established content standards, including those in *Benchmarks for Science Literacy* (American Association for the Advancement of Science, 1993) and the *National Science Education Standards* (National Research Council, 1996). Assessment items often test facts and definitions and provide little information about students' conceptual understanding. In addition, the alignment between test items and learning goals is often poor so that it is difficult to use the results of assessment to diagnose problems in students' thinking so that instruction can be modified to address those problems (Black and Wiliam, 1998; Rothman, 2003; DeBoer, Herrmann-Abell *et al*., 2008; DeBoer, Lee *et al*., 2008).

The multiple-choice format is often criticized because it is thought that multiple-choice items require only the recall of facts and cannot be used to diagnose students' thinking (Klassen, 2006). However, previous studies have disputed this line of thinking by showing that the use of distractors based on common student misconceptions can increase the diagnostic power of multiple-choice items (Hamilton *et al*., 1997; Sadler, 1998). The items developed for this study have been designed in this way so that they can be used diagnostically to reveal what students do and do not know, and also, what misconceptions and alternative ideas they have. The expectation is that the results of these types of assessments will then be used to inform and improve science instruction.

Our item development process begins with a precise definition of the knowledge being targeted by elaborating and drawing boundaries around the knowledge stated in each content standard. Items are then written that are carefully aligned to the target knowledge and that precisely measure student understanding of that knowledge along with common misconceptions that students have. During item development, pilot testing is used to obtain feedback from students about the items. Then scientists and science education experts review the items to ensure content alignment and construct validity. After revisions are made based on the reviews and student feedback, the items are field tested on a large national sample to determine the psychometric properties of the items. Details of the item development procedure have been described elsewhere (DeBoer *et al*., 2007; DeBoer, Herrmann-Abell *et al*., 2008; DeBoer, Lee *et al*., 2008).

Not only do we pay close attention to the qualitative validity of the items by aligning them to clearly defined learning goals, but we also use various quantitative techniques to ensure the validity of the items. For example, when analyzing data from the assessment items, we use Rasch modeling to examine the pattern of responses for each answer choice rather than treating all incorrect answers the same. In dichotomous scoring of multiple choice tests (correct-incorrect), the incorrect answer choices are lumped

AAAS Project 2061, 1200 New York Ave. NW, Washington, DC, 20005USA. E-mail: cabell@aaas.org

Table 1 Demographic information for field test participants

	Total	Female	Male	Primary language is English	Primary language is not English
Grade	% (N)	%	%	%	%
6th Grade	14% (1806)	50%	49%	91%	8%
7th Grade	21% (2815)	51%	48%	87%	12%
8th Grade	19% (2486)	49%	50%	87%	12%
9th Grade	10% (1269)	51%	46%	89%	8%
10th Grade	10% (1298)	52%	46%	87%	10%
11th Grade	10% (1322)	48%	49%	88%	9%
12th Grade	4% (593)	47%	50%	84%	12%
Entering Freshman on 1st day of College Chemistry	4% (474)	58%	42%		
College chemistry students	9% (1218)	52%	48%		
Chemistry graduate students	0.2% (31)	32%	65%		
Total	100% (13360)	50%	48%	76%	9%

together and no attempt is made to determine if there is a difference in how those incorrect answers discriminate among students. With dichotomous scoring, the curve corresponding to the probability of selecting the correct answer typically increases monotonically with increasing student understanding, and the curve for the set of distractors typically decreases monotonically with increasing student understanding (Haladyna, 1994). However, it has been shown that for distractor-driven items, such as the ones we have developed, the curves do not match the monotonic behavior of traditional items (Sadler, 1998). Therefore, to represent the data more accurately and to provide additional information about the incorrect ideas students have, we analyze each answer choice separately. We use option probability curves generated during Rasch modeling to show the probability of selecting each answer choice as a function of the overall performance level of the students on the topic being tested (as measured by the entire set of items). The four curves, one for each answer choice, show the probability that students who have a particular level of understanding of this topic will choose that answer choice. What typically results is that students whose understanding of the topic is low will be drawn to a particular misconception, and students whose understanding of the topic is higher will be drawn to other misconceptions. Detecting hierarchies of misconceptions allows educators to diagnose more accurately students' thinking, which enables them to target instruction more effectively.

In this paper, we summarize the results of the field testing of multiple-choice assessment items that are precisely aligned to U.S. national content standards about middle school chemistry ideas and that incorporate commonly held student misconceptions as distractors. We also describe several examples of how Rasch modeling and the resulting option probability curves can be used to reveal hierarchies of misconceptions in students' understanding of chemistry and structural problems with individual items.

Method

Description of the sample tested

The student data reported on here resulted from the field testing of assessment items aligned to middle school ideas about chemistry. Table 1 shows a summary of the demographic information for the student sample. Students from 122 schools in 30 states participated in the middle school field tests, and students from 188 schools in 41 states and one school in Puerto Rico participated in the high school field tests. We also administered the items to populations of college students who were likely to have the knowledge being targeted by the items as a way of further validating the items. The college students were from two universities (a public university in the southern region of the U.S. and a public university in the northeastern region of the U.S.). These students included 474 students who had taken high school chemistry and were enrolled in, but had not yet received, instruction in a college level introductory chemistry course, 1218 students who had received at least one semester of college level chemistry instruction, and 31 chemistry graduate students.

Description of the target learning goals

The ideas on which students were tested are based on Chapter 4, Section D of *Benchmarks for Science Literacy* (AAAS, 1993) and Physical Science Content Standard B of the *National Science Education Standards* (NRC, 1996). The key ideas are:

- All matter is made up of atoms.
- All atoms are extremely small.
- All atoms and molecules are in constant motion.
- 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.

- For any single state of matter, changes in temperature typically change the average distance between atoms or molecules. Most substances or mixtures of substances expand when heated and contract when cooled.
- There are differences in the spacing, motion, and interaction of atoms and molecules that make up solids, liquids, and gases.
- Changes of state can be explained in terms of changes in the arrangement, motion, and interaction of atoms and molecules.
- 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.
- Many substances react chemically in predictable ways with other substances to form new substances with different characteristic properties.
- When substances interact to form new substances, the atoms that make up the molecules of the original substances rearrange into new molecules.
- Whenever substances interact with one another, regardless of how they combine or break apart, the total mass remains the same.
- 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.

Misconceptions as distractors

Often students who do not have the targeted knowledge or who hold misconceptions about it are, nevertheless, able to respond correctly to traditional assessment items by guessing or using test-wiseness strategies. Incorporating misconceptions in the distractors provides these students with plausible answer choices to select from, so they are less likely to guess the correct answer choice. The following is a list of the misconceptions that were tested in the example items discussed below in the Results and discussion section.

- Atoms or molecules are embedded in matter (Renstrom *et al*., 1990; Griffiths and Preston, 1992; Lee *et al*., 1993; Johnson, 1998).
- Matter exists only when there is perceptual evidence of its existence (Stavy, 1990).
- Biological materials are not matter (Stavy, 1991).
- Atoms or molecules of a solid are not moving when the solid itself is not moving (Novak and Musonda, 1991; Lee *et al*., 1993).
- Atoms of molecules of a gas are not in motion (Novick and Nussbaum, 1981).
- The atoms of the reactants of a chemical reaction are transformed into other atoms (Andersson, 1986).
- Students may have difficulty with the distinction between properties of objects (e.g., weight and volume) and characteristic properties of the substance(s) of which they are made (e.g., density) (Smith *et al*., 2006).
- Temperature is a characteristic property of the substance from which an object is made (Thomaz *et al*., 1995).

Table 2 Summary of fit statistics

	Min	Max	Median
Standard error	0.02	0.05	0.04
Infit mean-square	0.82	1.47	0.96
Outfit mean-square	0.68	2.18	0.94
Point-measure correlation coefficients	0.17	0.59	0.47
Item separation index (reliability)		17.38 (1.00)	
Person separation index (reliability)		2.30 (0.84)	

Description of the field test

A total of 91 items aligned to the learning goals listed above were included in the field tests. Because we were testing more items than students could finish in a typical class period, we created multiple test forms that contained subsets of the available items. Each student received 26 to 30 items, and each item was answered by an average of 4000 students. Linking items allowed us to use Rasch modeling to compare item characteristics across forms. For each item, the students were asked to choose the one correct answer; students who chose more than one answer were marked incorrect.

Description of Rasch modeling

We used Rasch modeling to analyze the field test data. In the dichotomous Rasch model, the probability that a student will respond to an item correctly is determined by the difference in the student's overall performance level and the difficulty of the item, according to the following equation:

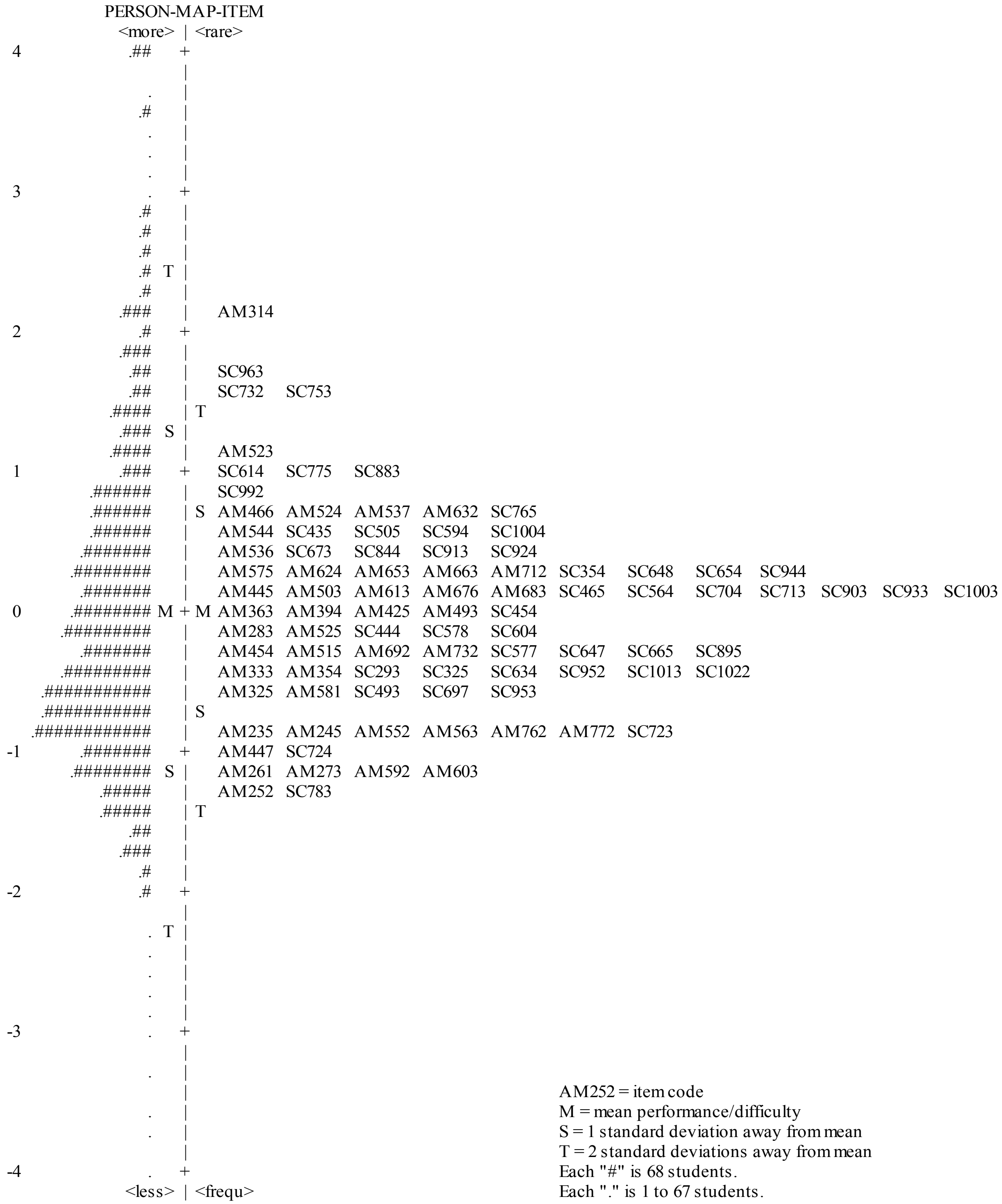
$$\ln\left(\frac{P_{ni}}{1-P_{ni}}\right)=B_n-D_i$$

where P_{ni} is the probability that student *n* of overall performance level B_n will respond correctly to item *i* with a difficulty of D_i (Liu and Boone, 2006; Bond and Fox, 2007). It is important to note that the student and item measures, B_n and D_i, are expressed on the same interval scale and are mutually independent, which is not the case for percent correct statistics. Student performance level and item difficulty are measured in the unit of logarithm called log odds or logits, which can vary from -∞ to +∞.

WINSTEPS (Linacre, 2009) was used to estimate the students' understanding of the topic and the item difficulties. From these parameters, we were able to determine if the range of item difficulty was appropriate for the students who were sampled and the extent to which scores on each of the items correlated with the entire set of items (point-measure correlation). We also looked to see if the pattern of student responses followed expectations such that:

> *…a person having a greater ability than another should have the greater probability of solving any item of the type in question, and similarly, one item being more difficult than another one means that for any person the probability of solving the second item correctly is the greater one* (Rasch, 1960 in Wright and Stone, 1999).

Any discrepancies prompted us to examine the items more closely to determine the cause.



```
              PERSON-MAP-ITEM
              <more> | <rare>
4              .##   +
                     |
                 .   |
                .#   |
                 .   |
                 .   |
                 .   |
3                .   +
                .#   |
                .#   |
                .#   |
                .# T |
                .#   |
              .###   |   AM314
2               .#   +
              .###   |
               .##   |   SC963
               .##   |   SC732  SC753
             .####   | T
              .### S |
             .####   |   AM523
1             .###   +   SC614  SC775  SC883
           .######   |   SC992
           .######   | S AM466  AM524  AM537  AM632  SC765
           .######   |   AM544  SC435  SC505  SC594  SC1004
          .#######   |   AM536  SC673  SC844  SC913  SC924
         .########   |   AM575  AM624  AM653  AM663  AM712  SC354  SC648  SC654  SC944
          .#######   |   AM445  AM503  AM613  AM676  AM683  SC465  SC564  SC704  SC713  SC903  SC933  SC1003
0        .######## M + M AM363  AM394  AM425  AM493  SC454
        .#########   |   AM283  AM525  SC444  SC578  SC604
          .#######   |   AM454  AM515  AM692  AM732  SC577  SC647  SC665  SC895
        .#########   |   AM333  AM354  SC293  SC325  SC634  SC952  SC1013 SC1022
      .###########   |   AM325  AM581  SC493  SC697  SC953
      .###########   | S
     .############   |   AM235  AM245  AM552  AM563  AM762  AM772  SC723
-1        .#######   +   AM447  SC724
         .######## S |   AM261  AM273  AM592  AM603
            .#####   |   AM252  SC783
            .#####   | T
               .##   |
              .###   |
                .#   |
-2              .#   +
                     |
                 . T |
                 .   |
                 .   |
                 .   |
                 .   |
-3               .   +
                     |
                 .   |
                     |                            AM252 = item code
                 .   |                            M = mean performance/difficulty
                 .   |                            S = 1 standard deviation away from mean
                     |                            T = 2 standard deviations away from mean
-4               .   +                            Each "#" is 68 students.
              <less> | <frequ>                    Each "." is 1 to 67 students.
```

Fig. 1 Item–person map for the 91 items included in the field tests.

Results and discussion

Model fit

A summary of the fit statistics for the field test results is presented in Table 2. The data had a good fit to the dichotomous Rasch model, indicating a unidimensional set of items targeting a single construct. The separation indices and corresponding reliabilities were high and the standard errors for the items were low. The infit and outfit mean-square values for the majority of the items were within the acceptable range of 0.7 to 1.3 for multiple-choice tests (Bond *et al.*, 2007).

Figure 1 shows the item-person map for the entire set of 91 items. The map shows the range of student performance on the left side of a vertical line and item difficulties on the right side of the line. Low performance/difficulty is

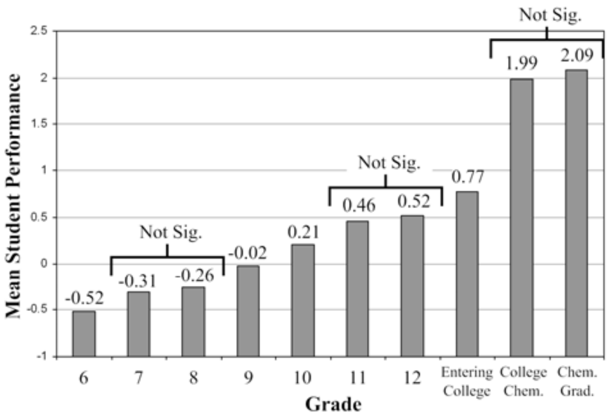


Fig. 2 Mean student performance level by grade

represented at the bottom of the map and high performance/difficulty is represented at the top of the map. The mean of the item difficulties is set at zero. When item difficulty and student performance level match, the student has a 50% chance of answering the item correctly. The map reveals that for these items and these students the mean item difficulty is equal to the mean student performance level and that along most of the range of student performance, there are test items that provide a reliable measure of what they know about this topic. However, for the lowest and highest parts of the student performance distribution there are no items. The lack of items to assess student understanding of the topic is especially evident at the high end of the scale. This is not surprising, given that it is mostly college level chemistry students who are at the upper end of the scale and the items were meant to test middle school ideas about chemistry.

Grade-to-grade improvement

The bar graph in Figure 2 shows that there was a general increase in the performance of students from sixth grade through high school and college. The graph shows the mean performance level measured in logits of the students in each of the different grades as measured by the entire set of items. One-way ANOVA revealed overall statistically significant differences in the means (F(9, 13293) = 595.09, $p < 0.001$), and a Bonferroni post hoc test showed that the differences in mean performance level are statistically significant at the 0.01 level for all grades, except between grades seven and eight, grades eleven and twelve, and between college chemistry students and chemistry graduate students, as noted in Fig. 2. Overall, there is an increase in understanding of the topic with increasing grade level. This gradual increase in students' understanding of chemistry from sixth to twelfth grade was also found in a previous study (Liu, 2007). The larger increase between high school and college can be attributed to the greater selectivity of the sample of college students.

Distractor analysis

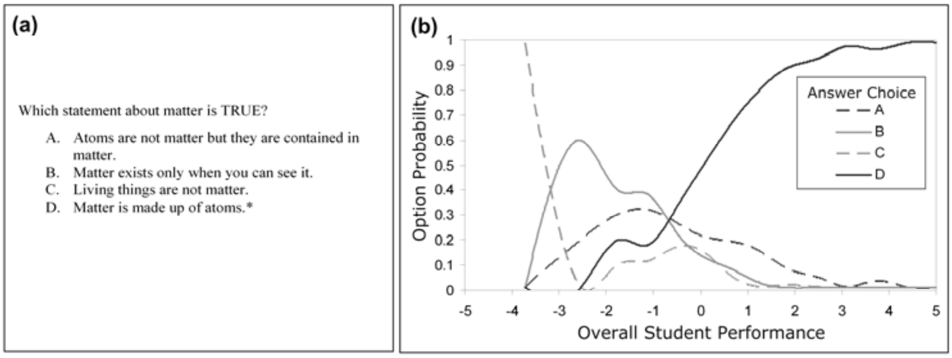
WINSTEPS was used to obtain option probability curves for each item (Linacre, 2009). We use the option probability curves because they present a visual image of the distribution of correct answers and misconceptions across the spectrum of student knowledge (ranging from sixth grade to college chemistry students). 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 may also suggest hierarchies of misconceptions that disappear in sequence as students become more knowledgeable about a topic either through out-of-school experience or through formal instruction. In this paper, we present examples of option probability curves for four chemistry items in our study.

Example 1. In the first example, shown in Figure 3(a), an item tested students' understanding of the nature of matter, namely, that all matter is made up of atoms. The corresponding option probability curves are shown in Figure 3(b).

Students who have a very low understanding of the topic (less than -3 on the overall performance scale) were more likely to chose answer choice C (living things are not matter). Students with an overall performance level between -3 and -1 were more likely to chose answer choice B (matter exists only when you can see it) and students with an overall performance level greater than -1 were more likely to choose the correct answer choice D (all matter is made up of atoms). This pattern of response by performance level is understandable. At the lowest level are students whose definition of matter includes only inanimate objects, which are often the first examples that are used when teaching about matter. This also explains the misconception that only things that can be seen (i.e. liquids and solids but not gases) are matter. As instruction progresses into the upper grades, students gain an understanding of gases and that all matter is made up of atoms. The documented misconception that atoms are embedded in matter (answer choice A) is significant for a wide range of student performance levels (-2 to 1), but it is never the most likely to be chosen answer choice at any level for this item.

Example 2. In the next example, shown in Figure 4(a), an item tested students' understanding of the idea that all atoms are in constant motion, including the atoms that make up solids and gases. The option probability curves for this item are shown in Figure 4(b).

The option probability curves show that students with the lowest understanding of the topic are more likely to select the answer choice that states the misconception that neither the atoms of the air nor the atoms of the chair move (B). The selection of answer choice D (the atoms of the chair move but the atoms of the air do not) peaks in probability around a performance level of -4 and then decreases. Students with performance levels between -3.5 and 0 are more likely to choose the answer choice stating that the atoms of the air move but the atoms of the chair do not (C). The probability of selecting the correct answer choice A

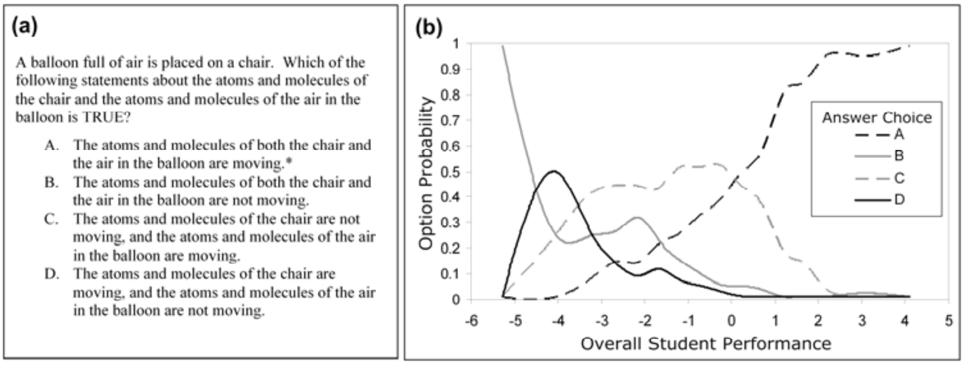


(a)

Which statement about matter is TRUE?

A. Atoms are not matter but they are contained in matter.
B. Matter exists only when you can see it.
C. Living things are not matter.
D. Matter is made up of atoms.*

Fig. 3 (a) Sample item (AM59-2) testing the idea that all matter is made up of atoms. (b) Corresponding option probability curves.

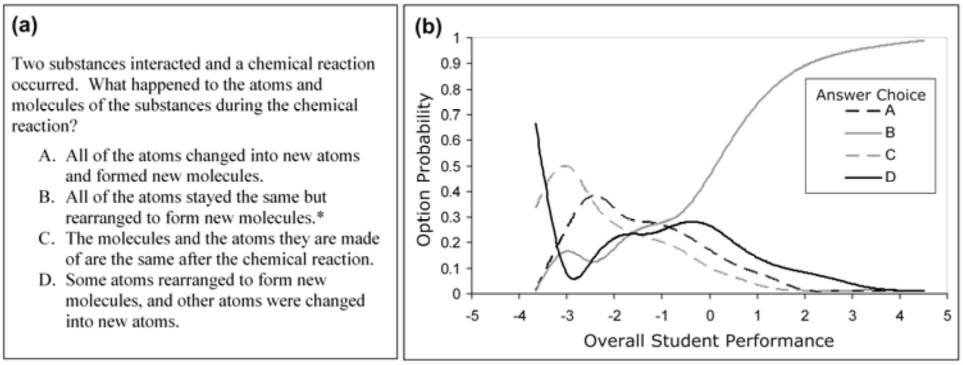


(a)

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.*
B. The atoms and molecules of both the chair and the air in the balloon are not moving.
C. The atoms and molecules of the chair are not moving, and the atoms and molecules of the air in the balloon are moving.
D. The atoms and molecules of the chair are moving, and the atoms and molecules of the air in the balloon are not moving.

Fig. 4 (a) Sample item (AM53-6) testing the idea that the atoms and molecules of solids and gases are constantly moving. (b) Corresponding option probability curves.

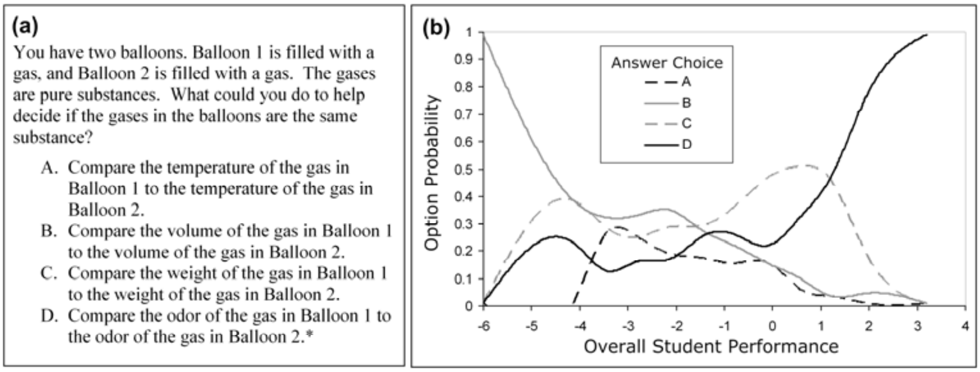


(a)

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.
B. All of the atoms stayed the same but rearranged to form new molecules.*
C. The molecules and the atoms they are made of are the same after the chemical reaction.
D. Some atoms rearranged to form new molecules, and other atoms were changed into new atoms.

Fig. 5 (a) Sample item (SC35-4) testing the idea that the atoms rearrange to form new molecules during a chemical reaction. (b) Corresponding option probability curves.



(a)

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.
B. Compare the volume of the gas in Balloon 1 to the volume of the gas in Balloon 2.
C. Compare the weight of the gas in Balloon 1 to the weight of the gas in Balloon 2.
D. Compare the odor of the gas in Balloon 1 to the odor of the gas in Balloon 2.*

Fig. 6 (a) Sample item (SC96-3) testing the idea that the odor is a characteristic property that can be used to identify a substance. (b) Corresponding option probability curves.

(atoms of both the chair and the gas move) increases, as expected, with increasing understanding of the topic. The pattern of responses revealed in the option probability curves is interesting for two reasons. First, the curve for answer choice C shows that the misconception that atoms of solids do not move is held by students over a large range of overall understanding of the topic. For all the items in our study, answer choices that included this misconception were selected by 48% of middle school students, 40% of high school students, and 29% of college students. Second, the results match the expected progression of understanding. Without instruction or with limited instruction, students have no reason to think that any atoms are in motion. As they learn that atoms are perpetually in motion, it is not surprising that they think atoms of some kinds of materials are in motion, but atoms of other kinds of materials, especially solids, are not. Finally, with further instruction, they develop the correct idea that all atoms are in motion. The progression is from none to some to all.

Example 3. The item shown in Figure 5(a) tested students' understanding of what happens to atoms and molecules during a chemical reaction. The corresponding option probability curves are shown in Figure 5(b).

The probability of selecting answer choice D, which says that some atoms rearrange and some change into new atoms during a chemical reaction is highest for the students with the lowest overall performance. The probability of selecting answer choice C, which says that molecules and the atoms they are made of are the same after a chemical reaction, peaks at a performance level of -3 and then decreases with increasing performance. Between performance levels -2 and -1, students are about equally likely to choose any of the four answer choices. For students above a performance level of -0.5, the probability of selecting the correct answer B, which says that in chemical reactions the atoms stay the same but rearrange to form new molecules, increases rapidly. The option probability curves for this item differ from those for the previous examples in that there is a significant performance range where the probability of selecting any of the incorrect answers is equal prior to a steady increase in probability of selecting the correct answer. This could be due to the fact that the item requires a formal mental model of chemical reactions, and neither the correct answer nor the misconceptions can be derived from everyday experience. Students are most likely to develop the idea that during a chemical reaction, atoms rearrange to form new molecules through formal instruction.

Example 4. Figure 6(a) shows an item that was intended to test students' knowledge that odor is a characteristic property that can be used to identify a substance. The option probability curves for this item are displayed in Figure 6(b).

The probability of selecting the answer choice corresponding to the misconception that volume is a characteristic property (B) is highest for the students with the lowest understanding of the topic and then decreases with increasing understanding. The probability of selecting the answer choice corresponding to the misconception that temperature is a characteristic property (A) is the highest around a performance level of -3 , but at no point on the knowledge continuum is it the most likely answer choice to be selected by students. Those results are not surprising. But the curve for the answer choice involving the comparison of the weights of the gases (C) is unusual because it shows two peaks in probability, one at very low performance levels and the other at relatively high performance levels. The curve for the correct answer choice (D) increases along the knowledge spectrum except where it shows dips corresponding to the peak of answer choice A and the second peak of answer choice C.

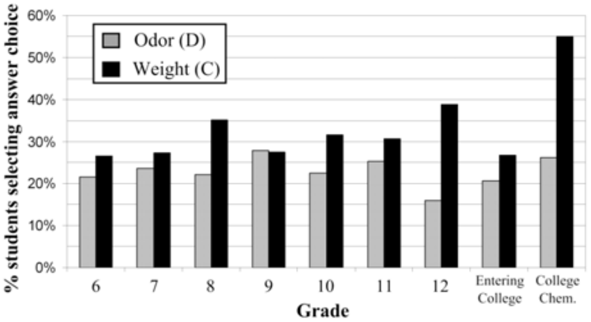


Fig. 7 Percentage of students selecting answer choice C and the correct answer choice D in sample item 4 (SC96-3) shown in Figure 6(a).

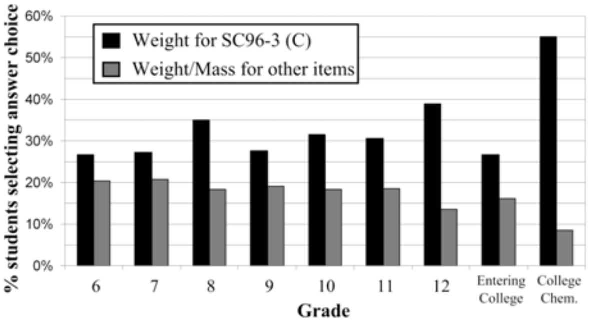


Fig. 8 Percentage of students selecting answer choice C in sample item 4 (SC96-3) shown in Figure 6(a) compared to the percentage of students selecting the answer choices in five other items that correspond to the misconception that weight/mass is a characteristic property.

This unusual shape of the curve for answer choice C prompted us to take a closer look at the item and at the answer choice selections of similar items testing the misconception that weight or mass is a characteristic property that can be used to identify a substance. As shown in Figure 7, the percentage of students selecting the correct answer choice (D) that odor is a characteristic property does not increase from grade to grade. However, the percentage of students choosing the answer choice that involves using weight to identify the gas in the balloons (C) does increase from grade to grade ending with about 55% of the college students selecting this option. It did not make sense to us that over half of the college students would actually think that weight is a characteristic property of a substance, so we looked at the results of five other items that tested for this misconception. Figure 8 compares the percentage of students choosing the answer that involves the comparing weights of the balloons (C) to the percentage of students choosing answer choices that involve weight/mass as a characteristic property in the five other items. The percentage of students who indicated that weight/mass is a characteristic property on those five items is significantly lower than the percentage of students who chose weight as a characteristic property on the balloon item, especially for the case of the college students where less than 10% selected the weight/mass answer choices in those other five items. This discrepancy suggested a problem with the structure of the balloon item.

We speculated that some of the higher performing students may have assumed that the balloons referred to in the item 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 gases 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 option probability curve and an analysis of the answer choice selections quickly revealed a structural problem with the item that could be corrected by stating in the item stem 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 U.S. national content standards about chemistry. Combining qualitative item development procedures and quantitative data analysis allowed us to probe more deeply into the progression of understanding of chemistry by obtaining information about the ideas students know and do not know, and also the misconceptions that they hold. The option probability curves generated during Rasch modeling revealed that the prevalence of some misconceptions is greatest for low-performing students and decreases steadily as performance increases, but for other misconceptions the prevalence increases steadily before decreasing in favor of the correct answer. The availability of this type of information about which misconceptions are most prevalent for which groups of students can be valuable in informing and improving instruction. The option probability curves were also used to investigate the validity of the items and, for one item, the unusual shape of one of the distractor probability curves indicated a structural problem with that item.

The science assessement items developed for this study, along with Web-based resources that support the development and use of items aligned to science content standards, are available on the Project 2061 Web site (http://assessment.aaas.org/). Because these items are carefully aligned with the key ideas about chemistry but not to any single curriculum or instructional approach, our hope is that researchers and developers of curriculum materials will be able to use the items and quantitative methods similar to the ones discussed in this paper to compare the effectiveness of various materials and approaches with more precision and objectivity. Additionally, although this study did not address the question of individual student performance or growth, it is expected that the items will be useful in helping teachers diagnose individual students' thinking, so that they can target instruction more effectively.

Acknowledgements

This work was funded by the National Science Foundation under Grants No. ESI 0227557 and ESI 0352473.

References

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

American Federation of Teachers, (2006), *Smart testing: let's get it right (policy brief no. 19)*, American Federation of Teachers, Washington, DC,.

Andersson B. R., (1986), Pupils' explanations of some aspects of chemical reactions, *Sci. Educ.*, **70**, 549-563.

Black P and Wiliam, D., (1998), Inside the black box: raising standards through classroom assessment, *Phi Delta Kappan*, **80**, 139-148.

Bond T. G. and Fox C. M., (2007), *Applying the Rasch model: fundamental measurement in the human sciences*, Mahwah, NJ: Lawrence Erlbaum Associates.

DeBoer G. E., Herrmann-Abell C. F. and Gogos A., (2007), Assessment linked to science learning goals: probing student thinking during item development, in *Proceedings of the National Association for Research in Science Teaching Annual Conference*, New Orleans, LA.

DeBoer G. E., Herrmann-Abell C. F., Gogos A., Michiels A., Regan T. and Wilson P., (2008), Assessment linked to science learning goals: probing student thinking through assessment, in Coffey J., Douglas R. and Stearns C. (eds.), *Assessing student learning: Perspectives from research and practice*, Arlington, VA, NSTA Press, pp. 231-252.

DeBoer G. E., Lee H. S. and Husic F., (2008), Assessing integrated understanding of science, in Kali Y., Linn M. C. and Roseman J. E. (eds.), *Coherent science education: implications for curriculum, instruction, and policy*, New York, NY, Columbia University Teachers College Press, pp. 153-182.

Griffiths A. K. and Preston K. R., (1992), Grade-12 students' misconceptions relating to fundamental characteristics of atoms and molecules, *J. Res. Sci. Teach.*, **29**, 611-628.

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

Hamilton L. S., Nussbaum E. M. and Snow R. E., (1997), Interview procedures for validating science assessments, *Appl. Meas. Educ.*, **10**, 169-207.

Johnson P., (1998), Progression in children's understanding of a 'basic' particle theory: a longitudinal study, *Int. J. Sci. Educ.*, **20**, 393-412.

Klassen S., (2006), Contextual assessment in science education: background, issues, and policy, *Sci. Educ.*, **90**, 820-851.

Lee O., Eichinger D. C., Anderson C. W., Berkheimer G. D. and Blaskeslee T. D., (1993), Changing middle school students' conceptions of matter and molecules, *J. Res. Sci. Teach.*, **30**, 249-270.

Linacre J. M., (2009), *Winsteps Rasch measurement computer program*, Chicago, Winsteps.com.

Liu X., (2007), Elementary to high school students' growth over an academic year in understanding the concept of matter, *J. Chem. Educ.*, **84**, 1853-1856.

Liu X. and Boone W. J., (2006), Introduction to Rasch measurement in science education, in Liu, X. and Boone, W. J. (eds.), *Applications of Rasch measurement in science education*, Maple Grove, Minnesota, JAM Press, pp. 1-22.

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

Novak J. D. and Musonda D., (1991), A twelve-year longitudinal study of science concept learning, *Am. Educ. Res. J.*, **28**, 117-153.

Novick S. and Nussbaum J., (1981), Pupils' understanding of the particulate nature of matter: a cross-age study, *Sci. Educ.*, **65**, 187-196.

Rasch G., (1960), *Probabilistic models for some intelligence and attainment tests*, Chicago, University of Chicago Press.

Renstrom L., Andersson B. and Marton F., (1990), Students' conceptions of matter, *J. Educ. Psychol.*, **82**, 555-569.

Rothman R., (2003), *Imperfect matches: the alignment of standards and tests,* Washington, DC, National Research Council.

Sadler P. M., (1998), Psychometric models of student conceptions in science: reconciling qualitative studies and distractor-driven assessment instruments, *J. Res. Sci. Teach.*, **35**, 265-296.

Smith C. L., Wiser M., Anderson C. W. and Krajcik J., (2006), Implications of research on children's learning for standards and assessment: a proposed learning progression for matter and the atomic-molecular theory, *Measurement: Interdisciplinary Research and Perspectives*, **4**, 1-98.

Stavy R., (1990), Children's conceptions of changes in the state of matter: from liquid (or solid) to gas, *J. Res. Sci. Teach.*, **27**, 247-266.

Stavy R., (1991), Children's ideas about matter. *Sch. Sci. Math.*, **91**, 240-244.

Thomaz M. F., Malaquis I. M., Valente M. C. and Antunes M. J., (1995), An attempt to overcome alternative conceptions related to heat and temperature, *Phys. Educ.*, **30**, 19-26.

Wright B. D. and Stone M. H., (1999), *Measurement essentials*. Wilmington, Delaware, Wide Range, Inc.