Science
Education

Diagnostic Opportunities Using Rasch Measurement in the Context of a Misconceptions-Based Physical Science Assessment

STEFANIE A. WIND,[1] JESSICA D. GALE[2]
[1]*The University of Alabama, Tuscaloosa, AL 35487;* [2]*Center for Education Integrating Science Mathematics and Computing, Georgia Institute of Technology, Atlanta, GA 30308, USA*

Received 21 August 2014; accepted 25 February 2015
DOI 10.1002/sce.21172
Published online 11 May 2015 in Wiley Online Library (wileyonlinelibrary.com).

ABSTRACT: Multiple-choice (MC) items that are constructed such that distractors target known misconceptions for a particular domain provide useful diagnostic information about student misconceptions (Herrmann-Abell & DeBoer, 2011, 2014; Sadler, 1998). Item response theory models can be used to examine misconceptions distractor-driven multiple-choice (MDDMC) items as a method for examining patterns of student answer choices given their achievement level. Furthermore, changes in response patterns over time can yield useful information about changes in student understanding following instruction. Because it meets the requirements for invariant measurement, the Rasch model is a promising tool for examining the diagnostic properties of MDDMC items within a coherent measurement framework; however, this application of the dichotomous Rasch model in conjunction with MDDMC items is relatively unexplored (Herrmann-Abell, 2011, 2014). The purpose of this study is to explore the diagnostic information about student understanding of physical science concepts provided by a Rasch-based analysis of MDDMC items. This study

Correspondence to: Stefanie A. Wind, e-mail: wind.stefanie@gmail.com
An earlier version of this paper was presented at the International Objective Measurement Workshop in Philadelphia, PA, March 2014.
Contract grant sponsor: National Science Foundation.
Contract grant number: 0918618.
The contents of this article are solely the responsibility of the authors and do not necessarily represent the official views of the National Science Foundation.
Supporting Information is available in the online issue at wileyonlinelibrary.com.

examines the relationship between misconceptions and a measure of student achievement as a method for targeting curriculum development and professional development within the framework of design-based implementation research (Penuel & Fishman, 2012). Implications for research, theory, and practice are discussed. *Sci Ed* **99:**721–741, 2015

INTRODUCTION

An extensive literature in the learning sciences argues that misconceptions (also known as preconceptions or alternative, naïve, or intuitive understandings) play an important role in the science learning process (National Research Council [NRC], 2007). Science students are not blank slates; rather, they bring to their science learning experiences an array of beliefs and knowledge based on their previous education and experience of the physical world. Although research on early conceptual development demonstrates that the understanding of basic physical properties and phenomena emerges in infancy (Baillargeon, 2004; Spelke, Katz, Purcell, Ehrlich, & Breinlinger, 1992), by the time students enter school they are likely to have developed conceptions of the physical world that are inconsistent with formal scientific understanding (Caramazza, McCloskey, & Green, 1981; Ioannides & Vosniadou, 2002). For example, although even very young children have an intuitive understanding of force, this understanding is different from the concept of force embodied by Newtonian mechanics. Furthermore, misconceptions of scientific phenomena are robust and tend to be highly resistant to change through instruction (Chi, 2005)—so much so that they often persist into adulthood in spite of formal science education (Kikas, 2004). Thus, science education researchers suggest the identification of student misconceptions and the assessment of students' evolving conceptual understanding as key objectives for science educators.

One method for exploring student misconceptions is through the use of assessment items designed to reveal known misconceptions in a domain. Specifically, misconceptions distractor-driven multiple-choice (MDDMC) items can be constructed such that distractors target known misconceptions for a particular domain that have been identified through empirical research. As a result, both correct and incorrect responses provide diagnostic information about student understanding (Herrmann-Abell & DeBoer, 2011, 2014; Sadler, 1998). Although multiple-choice (MC) items are often criticized as an insufficient measure of higher order thinking skills (Klassen, 2006) and susceptible to test-wiseness and guessing (Lee & Winke, 2012; Rogers & Bateson, 1991; Roznowski & Bassett, 1992), examination of student responses to MDDMC items has been shown to provide insight into student understanding. As pointed out by Sadler (1998), carefully constructed distractors can be seen as "windows into childrens' ideas" and "markers of progress" that reveal information about student understanding which may be useful for targeting instruction/curriculum. He observed that assessments constructed with MDDMC items "combine the richness of qualitative research with the power of quantitative assessment, measuring conceptual change along a single uniform dimension" (Sadler, 1998, p. 265). Furthermore, student responses can be examined over time to identify changes in patterns of misconceptions.

Although previous research on the development of MC items has emphasized the importance of constructing distractors that are "plausible and discriminating," (Haladyna, 1999; Haladyna & Rodriguez, 2013, p. 91), the use of items whose distractors are designed to reflect distinct misconceptions based on empirical evidence of conceptual development is less common. In other words, there is a difference in the diagnostic information that is provided by MC items whose distractors have been purposefully developed to be

plausible by item writers or curriculum developers, and MC items whose distractors draw upon evidence of conceptual development based on domain-specific empirical research. In this study, we examined student responses to a set of MDDMC items before and after a course of instruction to explore changes in their understanding of specific concepts in physical science.

MODELING STUDENT RESPONSES TO MDDMC ITEMS

A variety of methods can be used to examine student responses to MC items whose distractors have been developed to provide diagnostic information, such as MDDMC items. For example, a variety of item analysis techniques based on descriptive and inferential statistics that are often associated with classical test theory (CTT) have been used to explore patterns of responses to MC items. As pointed out by Haladyna and Rodriguez (2013), CTT techniques for examining MC item distractors have not changed much since their introduction around the 1950s (Guilford, 1954; Lord, 1952) and are limited when low frequencies are observed for one or two answer choices—a common occurrence in MC item assessments.

An alternative to CTT methods for examining patterns of responses to MDDMC items is the use of models based on item response theory (IRT). IRT models are used to predict and explain student responses to individual items in terms of a latent variable. Whereas CTT techniques are based on the total score scale, which is not linear, IRT models are used to calculate estimates of student achievement and item difficulty on a linear (interval-level) scale. In the context of MDDMC items, IRT models can be used to examine the functional relationship between student responses and model-based estimates of achievement on a latent variable.

Examination of previous research indicates that a variety of IRT methods are available for gathering diagnostic information about student conceptions based on their answers to MC items. One approach involves treating the distractors as a "hierarchy of correctness" and analyzing student responses using a partial-credit model (Andrich & Styles, 2009; Asril & Marais, 2011; Smith, 1987). This method is appropriate when the misconceptions that serve as distractors contain aspects of the correct answer, such that they can be scored using a rating scale. Wilson (1992) proposed a related method that makes use of the ordered partition model (OPM). The OPM is particularly useful when used within a construct-modeling framework (Wilson, 2005), because the model incorporates a scoring scheme that relates each answer choice to a specific location on the construct map that represents a progression along a latent variable, with the correct answer receiving the highest score (Wilson, 2008). Because the OPM requires that score values be assigned to each answer choice, this approach is similar to a partial-credit treatment of MC item answer choices. The distinguishing feature of the OPM is that multiple answer choices can share the same score value. Similarly, Wang (1998) proposed a Rasch-type distractor model that estimates separate difficulty parameters for each distractor in MC items. The major benefit of the distractor model is its ability to examine model-data fit issues associated with distractors to inform item revision.

IRT methods have also been proposed that are appropriate for MDDMC items whose answer choices cannot be ordered in terms of "correctness" or progression along a latent variable. One approach within this category was illustrated by Herrmann-Abell and DeBoer (2011, 2014) based on Rasch measurement theory. The major benefit of Rasch measurement theory is that it is based on the principles of *invariant measurement.* In other words, Rasch models facilitate the interpretation of student achievement and item difficulty within a single frame of reference, such that inferences about student achievement do not depend

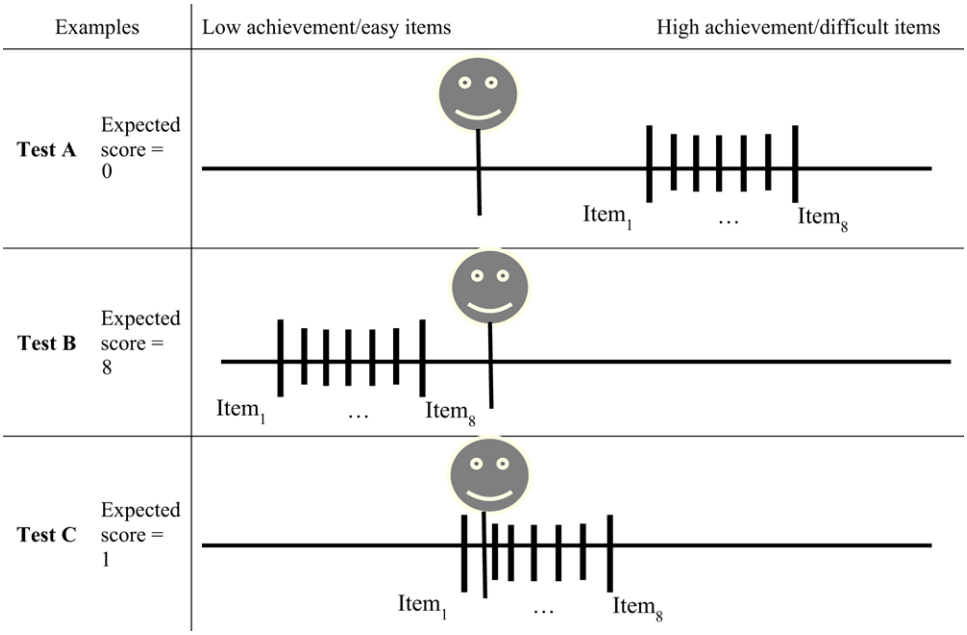

Figure 1. Invariant measurement illustration.
Note that the student has the same achievement level on all three tests.

on item characteristics and inferences about items do not depend on student characteristics (Wright & Masters, 1982).

Figure 1 illustrates the importance of separating items and persons to interpret results from an achievement test. Three separate tests and one student are used for the illustration, with each test illustrated separately using a horizontal line. The horizontal line represents the latent variable, or construct, that is being measured by the test, with locations farther to the right indicating higher achievement and higher item difficulty, and locations farther to the left indicating lower achievement and easier items. The person's location is the same on each test.

Test A is a difficult test for the example student, where the difficulty of the items exceeds the student's level of achievement, and the expected score is zero. Test B is an easy test for the example student, where the student's achievement exceeds the difficulty of all of the items, such that the expected score is eight. Test C includes one item that can be correctly answered by the student, and seven that are too difficult, so the expected score is one. This example illustrates three different expected scores for a single student that suggest three different achievement levels, depending on the characteristics of the items. As pointed out by Wright and Masters (1982):

> If the meaning of a test score depends on the characteristics of the items … we must "adjust" their score for the effects of the particular test items from which that particular score comes. This adjustment must be able to turn test-bound scores into measures of person ability which are test-free. (p. 6)

When data fit the Rasch model, these challenges related to sample dependency for the interpretation of student achievement and item difficulty are avoided. Specifically, estimates of student achievement are separated from the specific sample of items, and estimates of item

difficulty are separated from the specific sample of students. Furthermore, the estimates of student achievement and item difficulty are placed on a single linear (interval-level) scale, such that the same "ruler" is being used to make comparisons among students and items, regardless of which students or items are considered in the comparison. To consider the quality of a particular measurement, the Rasch model provides estimates of the error associated with each item calibration and person measure, along with indicators of model-data fit for each item and person. The Rasch model can also be used to examine "traditional" indicators of measurement quality, including reliability, for persons and items.

The ability to turn "test-bound scores into measures" is a major motivation for the adoption of Rasch modeling by many national and international assessments including the National Assessment of Educational Progress (NAEP) and Program for International Student Assessment (PISA) as well as the American Association for the Advancement of Science (AAAS) item bank, which was the major source for the assessment items used in this study. Furthermore, models based on Rasch measurement theory are frequently used to measure cognitive and affective constructs in the context of science education (Liu & Boone, 2006). For example, the dichotomous Rasch model has been applied to MC science assessments (e.g., Boone & Scantlebury, 2006). Rasch models for rating scale data (Andrich, 1978; Masters, 1982) have been used to measure affective variables related to science (Sondergeld & Johnson, 2014) including constructs such as socioscientific decision-making strategies (Eggert & Bogeholz, 2009) and self-efficacy (Boone, Townsend, & Starver, 2010).

Hermann-Abell and DeBoer applied the dichotomous Rasch model (Rasch, 1960/1980) to estimate student and item locations on a single linear continuum that represents a construct, or latent variable. For each item of interest, they used graphical displays to illustrate the proportion of students selecting each answer choice along the range of student achievement estimates. The resulting displays provide diagnostic information that describes the relationship between student achievement levels and the popularity of misconceptions that are included in MDDMC items. This approach goes beyond what would be obtained using proportions of students selecting each distractor at the pre- and posttest time points because it provides information about the degree to which each answer choice (misconception) is attractive to students at different levels of achievement. Furthermore, the use of the Rasch model is desirable in that estimates of student achievement can be described separately from item difficulty estimates, and that these estimates are on an interval-level scale. The current study applies and extends the Rasch-based distractor analysis methodology illustrated by Herrmann-Abell and DeBoer (2011, 2014) using MDDMC items that address physical science concepts for eighth-grade students using a pre- and posttest design.

PURPOSE

The purpose of this study is to explore diagnostic information about student understanding of physical science concepts provided by a Rasch-based analysis of MDDMC items. Furthermore, the study demonstrates the use of the Rasch-based MDDMC methodology as a tool for examining changes in student misconceptions following instruction via an experimental physical science curriculum.

Two research questions guided this study:

1. What does a Rasch analysis of MDDMC items reveal about student understanding and misconceptions of physical science concepts?
2. How can a Rasch analysis of MDDMC items be used to examine changes in student misconceptions of physical science concepts over time?

PROCEDURES

This study uses data from an assessment of physical science concepts that was designed as part of a larger design-based implementation research (DBIR) program (Penuel & Fishman, 2012; Penuel, Fishman, Cheng, & Sabelli, 2011). In contrast to efficacy research focused primarily on measuring treatment effects of an intervention, DBIR is an iterative approach that typically focuses on investigating and improving the implementation of educational interventions. As such, DBIR researchers often use analyses to iteratively refine interventions and to gain a deeper understanding of the conditions under which particular innovations can be expected to be effective (Harris, Phillips, & Penuel, 2011; Penuel & Fishman, 2012). Thus, a major goal for these assessments was to use the results to inform ongoing curriculum development, teacher professional development, and research on curriculum implementation.

The following section describes the procedures used to address the two guiding research questions.

Instrument

To examine student misconceptions before and after a unit of instruction via the experimental science curriculum, pre- and postassessments were developed using MDDMC items. The preassessment was administered before the beginning of the unit, and the postassessment was administered within 1 week of the end of the unit. The assessments shared 16 common MDDMC items, which were aligned to physical science domains that correspond to the learning goals of the science curriculum. The MDDMC items included in the physical science assessment were drawn from several sources, including original items from subject matter experts involved in the development of the science curriculum, released items from the National Assessment of Educational Progress Questions Tool (NAEP; U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, 2015), and items from the AAAS Item Bank developed by Project 2061.

Although each of the items was constructed using the MDDMC format, this study focused on items from the AAAS Project 2061 item bank. Items in the AAAS Project 2061 item bank were emphasized because of the empirical basis for answer choices that reflect student misconceptions (DeBoer et al., 2008). The AAAS Project 2061 items have been developed as part of a multiyear National Science Foundation–funded grant aimed at developing MDDMC items aligned with the science content standards described in *Benchmarks for Science Literacy* (AAAS, 1993) and the *National Science Education Standards* (NRC, 1996). Additional information about the development of the AAAS item bank is available in Herrmann-Abell and DeBoer (2011).

Participants

Participants included a sample of 403 eighth-grade students enrolled in three public middle schools where the science curriculum is currently implemented. One of these three schools is located in a rural community, and two schools are located in suburban communities outside a major city in the southeastern United States. Participating schools vary with respect to student ethnicity, socioeconomic status, and students' prior science achievement; however, the demographics of participating students were generally representative of each school community. Demographic information for each school is presented in Table 1. The sample included all students for whom parental consent and student assent forms were obtained, which represented 86% of students in participating classrooms.

TABLE 1
Demographic Data for Participating Schools

Characteristic	School 1 ($N = 164$)	School 2 ($N = 217$)	School 3 ($N = 89$)
School setting	Rural	Suburban	Suburban
Free/reduced lunch (%)	80	65	16
Underrepresented minorities in STEM (%)	52	67	25
Average class size	25	31	18

Data Analysis

Data analyses for this study consisted of two major steps. In Step 1, the dichotomous Rasch model was used to explore the psychometric properties of the physical science assessment, including estimates of student and item locations on the physical science construct and indices of model-data fit. In Step 2, independent analyses of the pre- and postadministrations of the physical science assessment were used to create graphical displays that illustrate student response patterns for both administrations of the assessment. Data analyses were conducted using the Winsteps computer program (Linacre, 2014).

Step 1: Dichotomous Rasch Model

The first step in the data analysis was the application of the dichotomous Rasch model to estimate student achievement and item difficulty on the pre- and postassessments. When the Rasch model is applied, person and item raw scores undergo a nonlinear transformation that creates a scale onto which persons and items can be mapped that is more likely to have equal units than the original total-score scale. The raw-score transformation is used to estimate person logits (B_n) that represent person achievement, and item logits (D_i) that represent item difficulty, in terms of the latent variable. The Rasch model can be stated in log-odds form as follows:

$$\mathrm{Ln}\,[\mathrm{P}_{ni} = 1/\mathrm{P}_{ni} = 0] = \mathrm{B}_n - \mathrm{D}_i \tag{1}$$

where B_n is the location of student n on the construct (i.e., student ability), D_i is the location of item i on the construct (i.e., item difficulty), and $\mathrm{Ln}[P_{ni} = 1/P_{ni} = 0]$ is the log of the odds for a correct response ($X = 1$) rather than an incorrect response ($X = 0$) by student n on item i.

When data fit the Rasch model, invariant measurement is achieved: person ability (B_n) may be estimated without the influence of the effects of item difficulties (D_i), and item difficulty (D_i) may be estimated without the effects of person abilities (B_n).

In this study, the Rasch model is applied independently to data from the pre- and postassessments. The results from analyses using the Rasch model include estimates of individual student and item locations on the logit scale. Because the estimates for students and items are on the same scale, a visual display called a *variable map* can be constructed that presents a graphical illustration of the logit-scale locations for students and items. Other Rasch literature has used the terms *Wright map* or *item-person map* in reference to the variable map (Wilson, 2011). In the context of this study, the variable map represents an operational definition of the physical science construct. In addition to logit-scale locations for persons and items, additional statistics can be calculated based on the Rasch model

results that are used to verify the appropriateness of the variable map as a representation of the physical science construct. Additional Rasch-based statistics that are examined in this study include separation statistics and indices of model-data fit for students and items.

Separation Statistics. Separation statistics are used to describe the degree to which differences among individuals, and items are realized in a measurement procedure. The *reliability of separation statistic* based on Rasch models indicates how well individual elements, such as students or items, can be differentiated from one another. The reliability of separation statistic for students is comparable to Cronbach's alpha coefficient because it reflects an estimate of true score to observed-score variance. However, Cronbach's alpha and the Rasch reliability of separation for student statistics are slightly different because alpha is based only on the *assumption* of linear measures and the Rasch-based reliability of separation statistic is based on a linear, interval-level scale when good model-data fit is observed. For items, the reliability of separation statistic describes the spread, or differences, in the difficulty of the MDDMC items. In addition to the reliability of separation statistic, a *chi-square statistic* (χ^2) can be calculated to determine whether the differences among logit-scale locations for students and items are statistically significant.

Model-Data Fit Statistics. In the context of the assessments examined in this study, model-data fit statistics can be used to identify individual students or items that do not match the expectations of the Rasch model. This study uses two fit statistics that are calculated in the Winsteps computer program (Linacre, 2014): standardized outfit mean square error (*MSE*) and standardized infit *MSE* statistics. Because it is unweighted, the outfit statistic is useful because it is particularly sensitive to "outliers," or extreme unexpected observations. Infit *MSE* statistics are weighted by statistical information for a particular facet; as a result they are not as sensitive to extreme outliers. Both unstandardized and standardized versions of these statistics are calculated in the Winsteps computer program. However, because the sampling distribution for the unstandardized *MSE* statistics is not known, researchers are encouraged (e.g., Smith, 2000; Smith, Schumacker, & Bush, 2000) to use the standardized versions of these statistics when considering fit to the Rasch model. As a result, this study reports the standardized versions of the statistics.

Correspondence Between Pre- and Postmeasures. Although it is possible to estimate item difficulty and student achievement in a single Rasch analysis (cf. Monsaas & Engelhard, 1991; Wright, 2003), independent calibrations of the results are better suited to the purpose of this study. Following Smith (1997), separate analyses were conducted for the pre- and postassessment time points as a method for discerning changes in patterns of misconceptions between the two administrations—evidenced through changes in response patterns. As a result, Equation (1) was applied independently to each administration, and a comparison of results from pre- and postassessment distractor analyses (described below) was used to consider changes in student misconceptions following instruction.

Step 2: Distractor Analysis

The second step in the data analysis for this study was a distractor analysis based on the results from Step 1. This distractor analysis technique for MDDMC items was adapted from the method illustrated by Herrmann-Abell and DeBoer (2011) in the context of a high school chemistry assessment and by Herrmann-Abell and DeBoer (2014) in the context

of K–12 assessments of energy concepts. For each item, the distractor analysis involves the use of graphical displays that illustrate the relationship between estimates of student achievement (B in Equation (1)) and the proportion of students selecting a particular answer choice (A, B, C, or D) at each time point (pre- and postassessment). Specifically, distractor analysis plots are created by plotting estimates of student achievement (low to high) based on the Rasch model along the x-axis, and the proportion of students selecting each answer choice along the y-axis. Additional details about the interpretation of distractor analysis plots are provided in the Results section.

These distractor analysis plots can be seen as diagnostic tools that summarize patterns of student misconceptions, changes in student conceptions over time, and the overall functioning of an item. The major source for items in this study was the AAAS item bank, whose items were developed to reflect relevant physical science misconceptions. As a result, these plots provide a visual display that illustrates the popularity ordering of misconceptions associated with each MDDMC item that was observed among students with varying levels of achievement. Along the same lines, examination of the plots associated with the pre- and postadministration of an MDDMC item can be used to identify changes in student conceptions following the program of instruction. For example, the plots may reveal that a particular misconception is prevalent among students within a certain range of logit-scale locations on the preassessment, and that this misconception decreases or persists on the postassessment. Changes in student conceptions are evident when the probability for selecting an answer choice among students at a particular achievement level varies between the pre- and postassessments. Furthermore, the distractor analysis plots provide an additional diagnostic tool that can be used to evaluate the degree to which answer choices are functioning as intended. For example, in a well-functioning item, the trace line for the correct answer choice should be monotonic with the latent variable—that is, the correct answer choice should be most attractive to students with the highest achievement measures.

In this study, distractor analysis plots were examined for each item at both the pre- and postassessment administrations. Results from the distractor analysis were considered in terms of relevant literature related to student misconceptions in physical science and the implications for curriculum and professional development within the context of DBIR. To the extent that distractor analyses reveal changes in student conceptions or persistent misconceptions even after students participate in activities intended to build understanding of physical science concepts, they may inform refinements or supplements to curricula, suggest potential topics for future teacher professional development, and guide research on curriculum implementation.

RESULTS

This study used two major steps to explore student response patterns to physical science MDDMC items over time. In this section, results from each step are summarized, followed by a discussion of the results in terms of the research questions that were used to guide the study.

Results for Step 1: Dichotomous Rasch Model

The first step in the data analysis involved the use of the dichotomous Rasch model to estimate achievement measures and item difficulty calibrations on the logit-scale based on student responses to the MDDMC items before and after a course of instruction. Table 2 displays summary statistics from the dichotomous Rasch model for the preassessment (Panel A) and the postassessment (Panel B). To provide a frame of reference for

TABLE 2
Summary Statistics from the Dichotomous Rasch Model: Pre- and Postassessments

	Panel A: Preassessment		Panel B: Postassessment	
Statistics	Student	Item	Student	Item
Measure				
M	–0.49	0.00	0.40	0.00
SD	0.78	0.70	1.08	0.75
N	482	16	422	16
Standardized outfit				
M	0.00	–0.10	0.10	0.00
SD	1.00	1.60	0.90	2.20
Standardized infit				
M	0.00	–0.20	0.10	0.00
SD	0.90	1.60	0.80	2.20
Separation statistics				
Reliability of separation	0.44	0.98	0.64	0.97
Chi-square	715.0*	796.8*	918.3*	573.6*
Degrees of freedom	481	15	421	15

*$p < .05$

interpreting student achievement measures, the item difficulty calibrations were centered at zero (mean set to zero). On the preassessment, results indicated that the overall logit-scale locations for students were lower than the logit-scale locations for the items (*mean B* = –0.49, *SD* = 0.78; *mean D* = 0.00, *SD* = 0.70); this suggests that the items were generally difficult for the students prior to instruction. On the postassessment, the opposite was true: The overall logit-scale locations for students were higher on the logit scale (*mean B* = 0.40, *SD* = 1.08; *mean D* = 0.00, *SD* = 0.75). Because the focus of this analysis is on changes in response patterns to items, the significance of changes in student achievement and item difficulty locations between the pre- and postassessments is not explored further here.

Table 2 also summarizes the results from model-data fit analyses for the pre- and postadministrations of the physical science assessment. As can be seen in the table, acceptable fit to the dichotomous Rasch model was observed at both time points, with mean standardized outfit and standardized infit *MSE* statistics near their expected value of 0.00 for items on the preassessment (*mean* standardized outfit *MSE* = –0.10, *SD* = 1.60; *mean* standardized infit *MSE* = –0.20, *SD* = 1.60) and students (*mean* outfit *MSE* = 0.00, *SD* = 1.00; *mean* infit *MSE* = 0.00, *SD* = 0.90). On the preassessment, results indicate high reliability of separation statistic for items (*Rel* = 0.98), and a moderate reliability of separation statistic for students (*Rel* = 0.40). Significant differences were observed among the logit-scale locations for items (χ^2 (15) = 685.1, $p < .001$) and students (χ^2 (485) = 715.0, $p < .001$). Panel B of Table 2 reports results from the Rasch analysis of student responses at the postassessment. As can be seen in the table, values of standardized outfit and standardized infit statistics indicate acceptable model-data fit for items (*mean* standardized outfit = 0.00, *SD* = 2.20; *mean* standardized infit = 0.00, *SD* = 2.20) and students (*mean* outfit = 0.10, *SD* = 0.90; *mean* infit = 0.10, *SD* = 0.80). Similar to the preassessment, significant differences were observed among the logit-scale locations for items (χ^2 (15) = 573.6, $p < .001$) and students (χ^2 (485) = 918.3, $p < .001$), with high reliability of separation

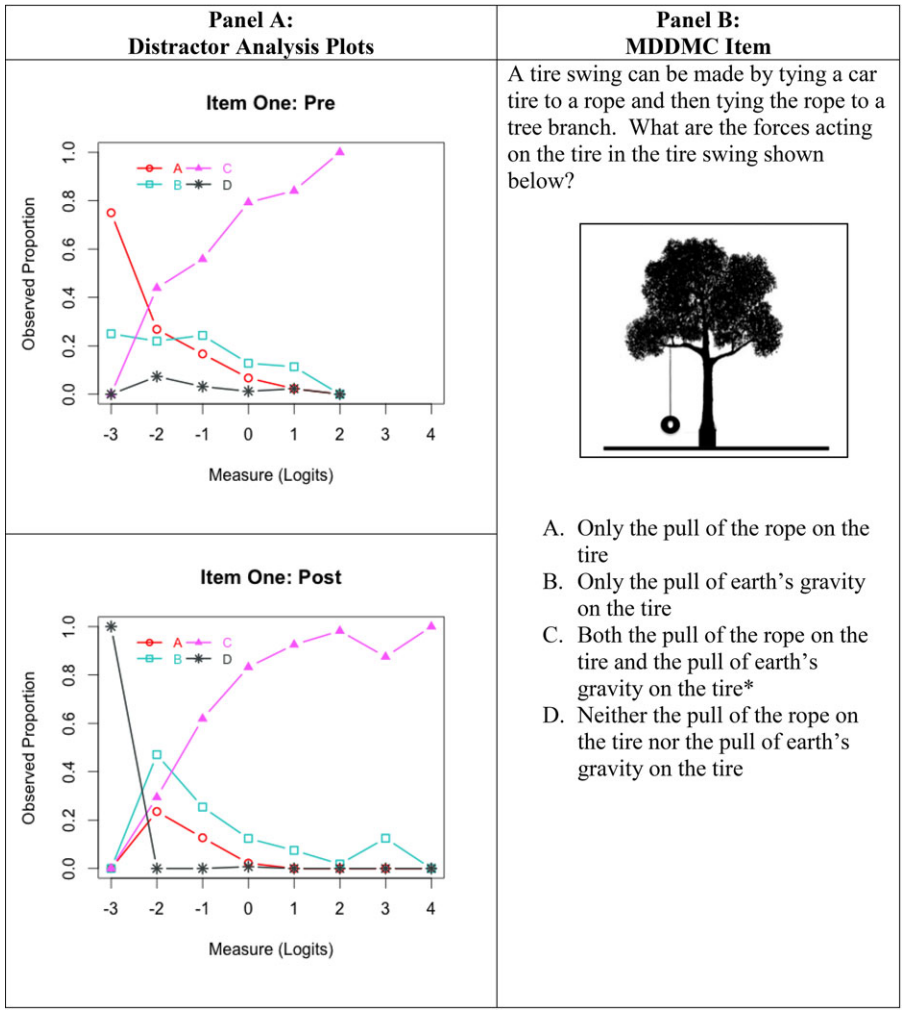

Figure 2. Illustrative distractor analysis plot: Item 1.
Note that this item is adapted from item number FM054002 from the AAAS item bank. The correct answer choice is C (marked with an asterisk). Detailed information about this item, including misconception references may be accessed at http://assessment.aaas.org/items/FM054002#/0.

statistic for items ($Rel = 0.97$), and a moderate reliability of separation statistic for students ($Rel = 0.67$).

Results for Step 2: Distractor Analysis

The Winsteps computer program (Linacre, 2014) was used to obtain the frequency of students selecting each answer choice (A, B, C, or D) along the range of student achievement estimates on the logit scale at each time point that were used to create pre- and postassessment distractor analysis plots for each item. In this section, the distractor analysis methodology is illustrated using four MDDMC items. The patterns observed among these four illustrative items reflect the overall patterns that were observed among the MDDMC items examined in this study. Then, the overall findings from the distractor analysis are summarized.

Figure 2 illustrates the distractor analysis procedure used in this study with the results for Item 1 on the pre- and postassessments. Panel A displays the distractor analysis plots for the pre- and postassessment responses to Item 1,[1] and Panel B displays Item 1 as it appeared in the assessments. The distractor analysis plots illustrate the relationship between student achievement measures (*B*) on the *x*-axis and the proportion of students selecting an answer choice (A, B, C, or D) on the *y*-axis. After Rasch estimates of student achievement on the logit scale were obtained from the Winsteps computer program (Linacre, 2014), student achievement estimates on the logit scale were rounded to the nearest integer value (–3 to 4). Then, the frequency of students selecting each answer choice was obtained for each value. At each point on the scale, the proportion of students selecting each answer choice was calculated by dividing the frequency of students who selected a given answer choice by the total number of students observed at each point on the scale. The distractor analysis plots were created by plotting the proportion of students selecting answer choices A, B, C, and D (*y*-axis) across the range of student achievement measures at each time point (*x*-axis). Accordingly, the *y*-axis values indicate the relative popularity of each answer choice for students with different levels of achievement (*x*-axis).

For Item 1, examination of both the pre- and postassessment distractor plots reveals that the correct answer choice (C) was most the most popular choice among students with high achievement on both administrations. This result can be seen by examining the proportion of students selecting answer choice C across the range of achievement measures. The plots for Item 1 indicate that answer choice C becomes increasingly frequent as student achievement measures increase. This finding, that the proportion of students selecting the correct answer choices increases as achievement levels increase, provides evidence that the item is functioning as expected. Furthermore, examination of the pre- and postassessment distractor plots reveals the proportion of students who were attracted to the misconceptions associated with the incorrect answer choices for Item 1 (A, B, and D), given estimates of achievement. Specifically, the distractor plot for the preassessment suggests that the misconception associated with answer choice A was prominent among students with low achievement measures. On the postassessment, the prominence of answer choice A decreased among low-achieving students, and the proportion of students selecting answer choices B and D increased among students with low measures.

As part of the curriculum's unit on force, students have numerous opportunities to learn about the various forces that may be acting on an object and to reason about forces by drawing and interpreting free-body diagrams. It is therefore expected that, having completed these exercises, higher achieving students would have found this item relatively easy at posttest. At the same time, previous work on students' use of free-body diagrams suggests that, while effective for promoting conceptual understanding, the use of free-body diagrams may be challenging for some students (Mattson, 2004; Van Heuvelen, 1991). For example, one study investigating students' use and understanding of free-body diagrams by undergraduate students in an algebra-based physics course found that high-achieving students readily used free-body diagrams to help solve problems and to evaluate their work, but lower achieving students tended to use free-body diagrams as a problem-solving aid or not at all (Rosengrant, Van Heuvelen, & Etkina, 2009). Thus, the apparent increase of misconceptions associated with answer choices B and D among lower achieving students may reflect the difficulty of learning to use free-body diagrams during

[1]In examining the distractor analysis plots for the pre- and postassessment, it is important to recognize the lack of equivalence in the interpretation of logit scale estimates of student achievement at each time point (e.g., Linacre and Wright, 1989). The comparison between pre- and postassessments in this study focuses on response patterns related to item distractors. No anchoring or other equating procedures were used to transform achievement measures to a common scale between time points.

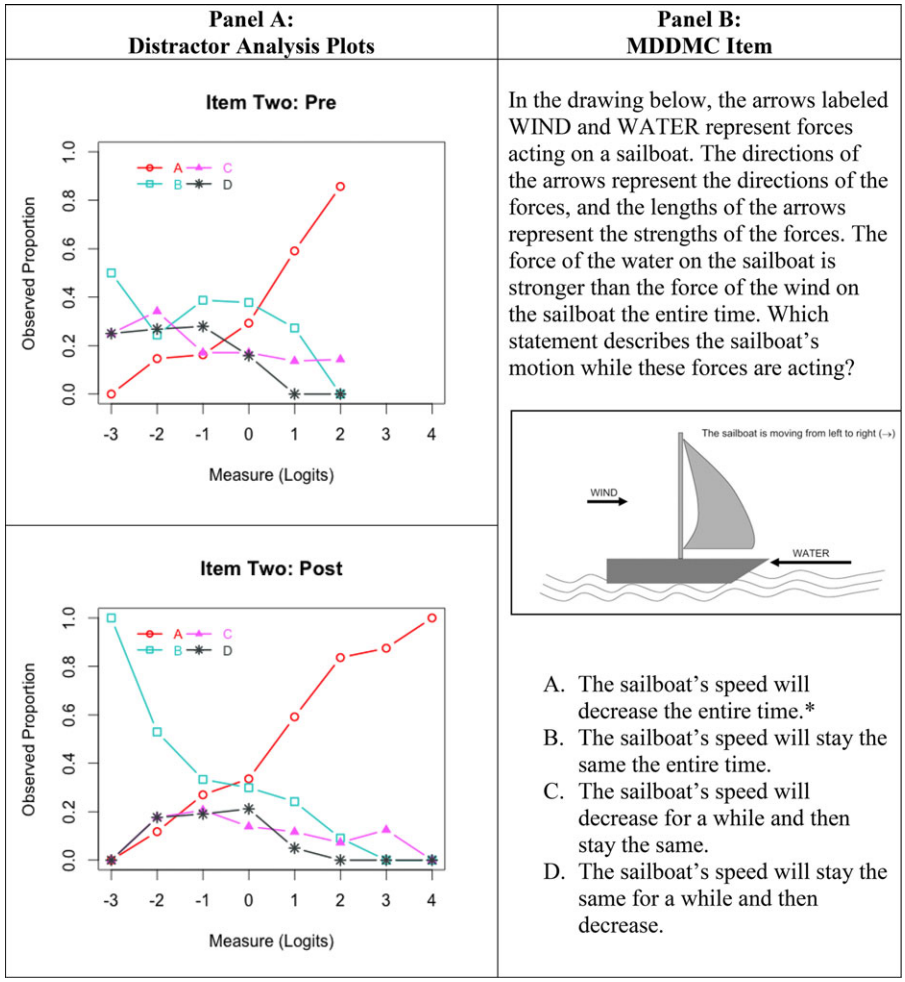

Figure 3. Illustrative distractor analysis plot: Item 2.
Note that this item is adapted from item number FM011005 from the AAAS item bank. The correct answer choice is A (marked with an asterisk). Detailed information about this item, including misconception references may be accessed at http://assessment.aaas.org/items/FM011005#/0.

instruction or unsuccessful attempts to apply their experience with free-body diagrams when answering Item 1.

A different pattern of changes between the pre- and postassessment response patterns is illustrated in Figure 3, for Item 2 on the pre- and postassessments. The plots shown in this figure are interpreted in the same manner as in Figure 2. As was seen for Item 1, the correct answer choice (A) is most popular among students with high achievement measures on both the pre- and postassessments. Furthermore, examination of Panel A reveals that response patterns changed between the pre- and postassessments for Item 2. On the preassessment, all three misconceptions (answer choices B, C, and D) displayed distinct "peaks" at different locations along the logit scale. These peaks indicate the level of achievement within which a particular answer choice was most attractive. The distractor plot for the postassessment indicates that the prominence of misconceptions associated with answer choices C and D decreased following instruction, whereas the misconception associated with answer choice B increased among students with low achievement measures.

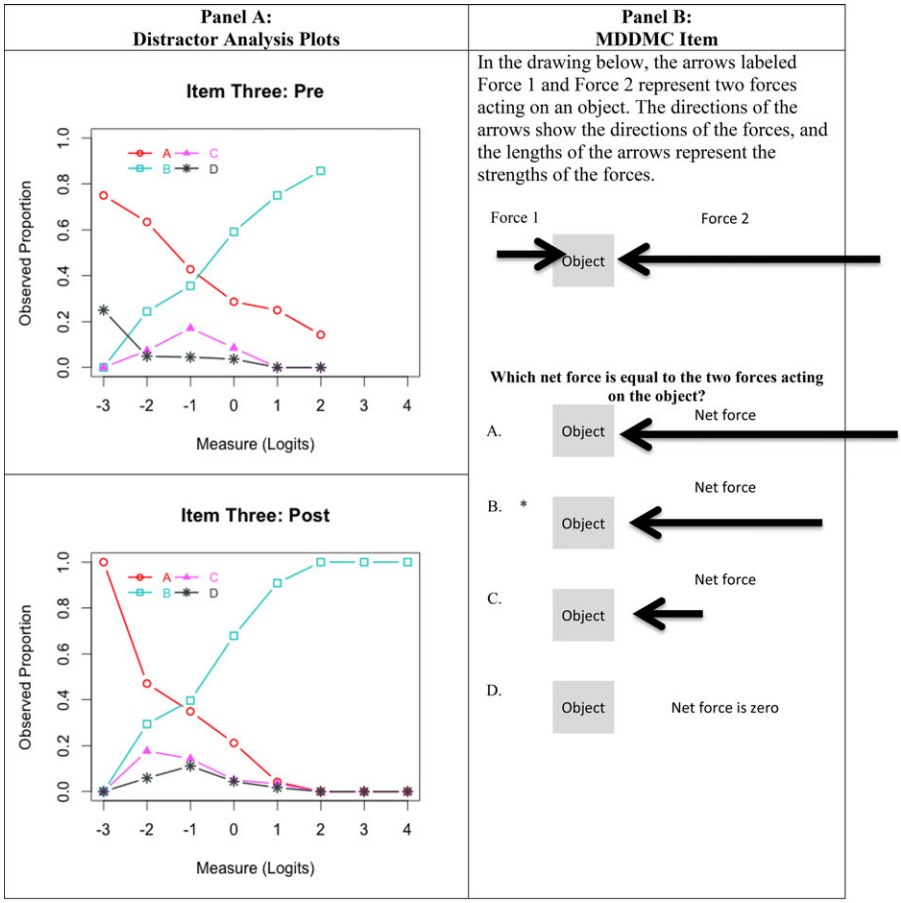

Figure 4. Illustrative distractor analysis plot: Item 3.
Note that this item is adapted from item number FM074005 from the AAAS item bank. The correct answer choice is B (marked with an asterisk). Detailed information about this item, including misconception references may be accessed at http://assessment.aaas.org/items/FM074005#/0.

Another important but difficult concept within the unit is the relationship between force and acceleration. Specifically, students engage in a variety of learning activities intended to illustrate how changes in the balance of forces (e.g., an increase or decrease in friction) results in a change in the object's motion. The changing patterns in response to Item 2 suggest that while students at various levels of achievement held misconceptions about this concept prior to curriculum enactment, these misconceptions may have been resolved for many students, and particularly for students in the mid and upper ranges of achievement, through their participation in curriculum activities.

A third type of change in response patterns between the pre- and postassessments is illustrated in Figure 4 using Item 3. For this item, the correct answer choice (B) was most popular among students with high achievement measures, and the misconception associated with answer choice A was most prominent on both the pre- and postassessments. Interestingly, the correct answer choice displayed a "peak" in popularity among students with achievement measures between approximately –2 and –1 logits on the preassessment that was not observed on the postassessment. This finding indicates that, for students with relatively low achievement (achievement estimates between about –2 and –1 logits), the

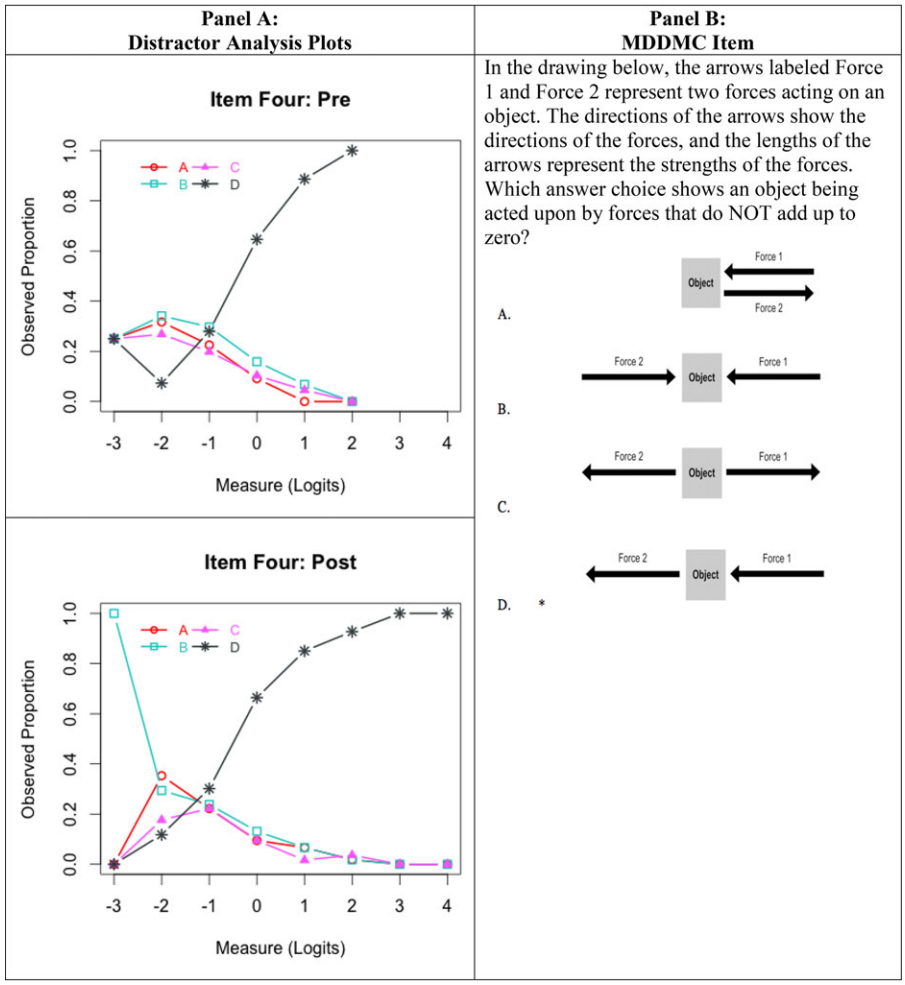

Figure 5. Illustrative distractor analysis plot: Item 4.
Note that this item is adapted from item number FM075004 from the AAAS item bank. The correct answer choice is D (marked with an asterisk). Detailed information about this item, including misconception references may be accessed at http://assessment.aaas.org/items/FM075004#/0.

correct answer choice was the most frequent selection. The relative ordering of the other two misconceptions (answer choices C and D) remained relatively stable following instruction. Similar to Item 1, students' response patterns on Item 3 may reflect the difficulty many students experience interpreting free-body diagrams, and specifically, using such diagrams to determine net force. That students rarely chose answer choices C and D at either pre- or posttest suggests that these options may not be well aligned to known misconceptions about net force. Additionally, the length of the force arrows in the diagram and answer choices for Item 3 may have been confusing for some students to compare visually to reason about the magnitude of the net force.

Figure 5 illustrates a fourth pattern of changes in student misconceptions between the pre- and postassessments using Item 4 from the physical science assessment. As was observed in the other illustrative distractor analyses, the correct answer choice (D) is monotonic with student achievement measures. On the preassessment, the misconceptions associated with answer choices C and B were most prominent among students with low

achievement measures. On the postassessment, the proportion of students with low measures who selected answer choice A increased, and the proportion of students who selected answer choice C decreased. One possibility that may account for changing response patterns on this item is the relative frequency with which each of the scenarios depicted by the answer options occurs within the curriculum. Specifically, students have multiple opportunities to observe and reason about forces acting on an object in a manner similar to that illustrated by options B, C, and D in comparison to the scenario depicted by option A, which is not illustrated within the curriculum.

Overall Summary of Distractor Analysis Results

Overall, findings from the distractor analysis for the 16 MDDMC items examined in this study suggested that the items displayed a variety of patterns of changes in student misconceptions between the pre- and postassessments. All of the distractor analysis plots for the pre- and postassessments indicated that the correct answer choice was monotonic with the latent variable—that is, the proportion of students selecting the correct answer choice increased as achievement measures increased. As suggested by the changes in the overall item difficulty estimates between the pre- and postassessment, the correct answer choice was more frequently observed on the postassessment.

It is important to note that the interpretation of patterns illustrated in the Rasch-based distractor analysis plots depends on issues related to sample size, the frequency with which each distractor is selected, and the rounding techniques used to produce the plots. The plots shown here were produced by calculating the observed proportion of students selecting each answer choice, given their estimated location on the logit scale rounded to the nearest integer. In terms of sample size issues, the distractor analysis for each of the MDDMC items examined in this study included examination of the frequencies of students who selected each answer choice. Because the focus of this study is on discovering overall patterns of misconceptions for diagnostic purposes, findings that small subsets of students selected a given answer choice were not seen as problematic. As also noted in Herrmann-Abell and DeBoer (2011, 2014), the overall patterns illustrated in the distractor analysis plots were viewed as useful diagnostic information to guide targeted and differentiated instruction to small groups of students.

SUMMARY AND DISCUSSION

This study employed a Rasch-based analysis of MDDMC items as a method to gather diagnostic information about student understanding of physical science concepts, as well as changes in student conceptions following a course of instruction. Student responses to MDDMC items that were included in pre- and postadministrations of a physical science assessments were explored in-depth using distractor analysis techniques that examined the relationship between estimates of student achievement from the Rasch model and observed patterns of responses that signaled misconceptions. In this section, findings are summarized and discussed as they relate to the two guiding questions for this study.

Research Question 1: What Does a Rasch Analysis of MDDMC Items Reveal About Student Understanding and Misconceptions of Physical Science Concepts?

The first research question for this study asked: *What does a Rasch analysis of MDDMC items reveal about student understanding and misconceptions of physical science concepts?* To address this question, this study used the dichotomous Rasch model to explore the

relationship between estimates of student achievement (B) and the probability for selecting answer choices that reflect specific physical science misconceptions. Results from the application of the Rasch model to data from the pre- and postadministrations of the physical science assessment suggested adequate fit to the model.

In general, findings from the distractor analysis suggested that the MDDMC items examined in this study were functioning as expected at both the pre- and postassessment administrations of the physical science assessment, with the correct answer choice being the most probable response for high-achieving students. In addition, results from the distractor analysis suggested that students hold a variety of misconceptions about physical science concepts and that the prominence of different misconceptions varies across the range of the latent variable. The analyses illustrated in this study highlight the usefulness of MDDMC items as diagnostic tools for identifying student misconceptions. Furthermore, the combination of distractor analysis with a measurement framework based on invariant measurement provides additional insight into the prevalence of these misconceptions among students with varying levels of achievement. As pointed out by Wilson (2008), the combination of a construct-driven measurement approach (i.e., Rasch measurement theory) has two major benefits:

> (a) it emphasizes how the misconceptions can be seen as positive manifestations of the cognitive level of sophistication of the student rather than as mere 'errors,' and (b) it makes available the strengths and possibilities of statistical modeling to the misconception analysis. (p. 86)

Through the combination of a measurement model with invariance properties and distractor analysis techniques, the methods employed in this study can be seen as a diagnostic method for examining student conceptions that provides additional insight about student conceptions beyond their estimates of achievement. By providing a more nuanced account of changing student conceptions over the course of curriculum implementation, such analyses may be of particular interest to curriculum developers and educators working to develop materials and instructional activities aimed at addressing persistent student misconceptions in physical science. For example, distractor analyses indicating the persistence of a misconception intended to be addressed by the curriculum suggest a need to revisit and perhaps revise how that particular concept is featured within the curriculum. Similarly, persistent misconceptions may signal a need for additional professional development to address gaps in teacher content knowledge or issues with curriculum implementation related to specific concepts.

Research Question 2: How Can a Rasch Analysis of MDDMC Items Be Used to Examine Changes in Student Misconceptions of Physical Science Concepts Over Time?

The second research question for this study asked: *How can a Rasch analysis of MDDMC items be used to examine changes in student misconceptions of physical science concepts over time?* To address this question, distractor analysis techniques were used to examine the functional relationship between estimates of student achievement and the prevalence of misconceptions before and after a course of instruction. The results from pre- and postassessment distractor analyses were considered side by side for each MDDMC item. In general, findings from the pre–post comparisons suggested that the probability that students selected a misconception rather than the correct answer choice decreased after instruction.

CONCLUSIONS AND IMPLICATIONS

This study extends previous research in several ways. First, it contributes to the literature on student misconceptions in physical science. Specifically, the study illustrates new methods that science education researchers and practitioners could use to gain insight into changes in students' conceptual understanding of physical science phenomena between pre- and postassessments. The graphical displays illustrated in this study provide a clear method for visually examining patterns in student conceptions that may be more readily interpreted by a range of audiences than previous tabular, graphical, and statistical summaries of distractor analysis results that may rely on advanced smoothing techniques or nonlinear measures of achievement. Combined with other indicators of student achievement at additional time points, the methods illustrated here may be able to provide insight about student learning trajectories over the course of a program of instruction. Furthermore, this study adds to previous distractor analysis research using Rasch measurement theory. Specifically, the study extends the methodology proposed by Herrmann-Abell and DeBoer (2011, 2014) by presenting a graphical distractor analysis method that can be used to explore changes in student response patterns across multiple time points. When used with MDDMC items at multiple time points in a program of instruction, this method provides diagnostic information regarding the popularity of various misconceptions at different achievement levels that can be used to inform instruction and professional development. When interpreting the results from this study in terms of substantive conclusions about student conceptions, it is important to note that the items used to illustrate the distractor analysis techniques may not perfectly reflect relevant misconceptions for all three incorrect answer choices. Additional research should explore patterns of student responses to MDDMC items using additional items whose distractors are clearly aligned to known misconceptions.

Finally, this study illustrates Rasch modeling as a promising tool for DBIR. The results of this study will inform refinements to the project-based science curriculum delivered to participating students. For example, persistent misconceptions among lower achieving students on certain items suggest a need to revisit curriculum materials to provide additional scaffolding and support for key concepts that may be particularly difficult, such as using free-body diagrams to reason about forces acting on an object. Thus, by applying the Rasch model to identify patterns of student misconceptions and better visualize changes in student conceptions following a course of instruction, the methods described in this study can inform science curriculum development, teacher professional development, and can be used to improve science teaching practices.

APPENDIX: PRE- AND POSTASSESSMENT CORRESPONDENCE

Pre–Post Analysis

This study used separate calibrations of item difficulty at the pre- and postassessment time points. As pointed out by Embretson (1991), separate calibrations are desirable in that they may avoid potential issues of multidimensionality due to the influence of modifiability of item difficulties as a result of the differential effects of instruction on individual items. To verify stability of the measures before and after instruction, the correspondence between item difficulty measures was examined using a correlation analysis and a comparison of item measures using the Wright and Stone (1979) separate calibration *t*-test (Smith, 1997). When interpreting the results from a comparison between pre- and postassessments, it is important to note that large sample sizes and multiple time points are required to gain a complete sense of changes in item difficulty (e.g., differential item functioning). As a

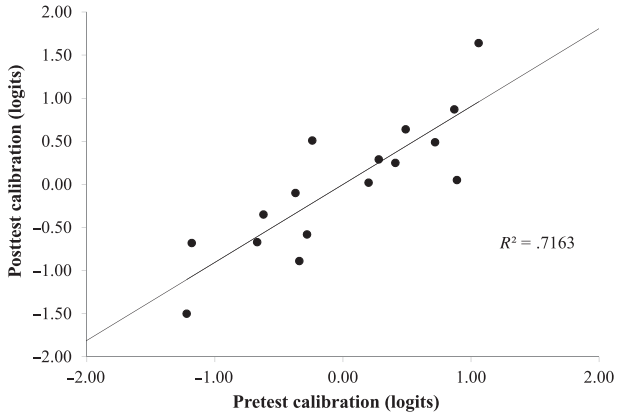

Figure A1. Correspondence between pre- and postassessment item calibrations.

result, this study did not focus on the magnitude of differences in item difficulty between administrations. Rather, the focus is on changes in response patterns to assessment items of interest related to a specific curriculum.

Results

A paired *t*-test for achievement measures (B) revealed that the overall difference between the pre- and postassessment means was highly significant ($p < .001$), indicating that there was a positive shift in achievement measures following instruction. The interpretation of this finding depends on the correspondence between the latent variable that was measured by both assessments. Following Smith (1997), the correspondence between pre- and postassessment item difficulties was examined to gauge the comparability of the construct at both time points. The correlation between the pre- and postassessment item calibrations is .85 ($p < .001$), and a bivariate plot of the item calibrations is given in Figure A1. A comparison of mean item difficulty values between the pre- and postassessments revealed that the overall difference between item difficulty estimates was highly significant ($p < .001$, $d = 1.10$). The results of the *t*-test comparison (Wright & Stone, 1979) of the individual pre- and postassessment item calibrations indicated that five of the item locations changed significantly following instruction ($t \geq |2|$). This result is higher than would be expected based on chance alone based on the 0.05 Type I error rate for these comparisons, which would result in 0.80 item pairs with a t value greater than $|2|$.

REFERENCES

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

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561–573.

Andrich, D., & Styles, I. (2009). Distractors with information in multiple choice items: A rationale based on the Rasch model. In E. V. Smith Jr. & G. E. Stone (Eds.), Criterion referenced testing: Practice analysis to score reporting using Rasch measurement models (pp. 24–70). Maple Grove, MN: JAM Press.

Asril, A., & Marais, I. (2011). Applying a Rasch model distractor analysis. In R. F. Cavanagh & R. F. Waugh (Eds.), Applications of Rasch measurement in learning environments research (Vol. 2, pp. 77–100). Rotterdam, The Netherlands: Sense Publishers.

Baillargeon, R. (2004). Infants' reasoning about hidden objects: evidence for event-general and event-specific expectations. Developmental Science, 7(4), 391–414.

Boone, W. J., & Scantlebury, K. (2006). The role of Rasch analysis when conducting science education research utilizing multiple-choice tests. Science Education, 90(2), 253–269.

Boone, W. J., Townsend, J. S., & Starver, J. (2010). Using Rasch theory to guide the practice of survey development and survey data analysis in science education and to inform science reform efforts: An exemplar utilizing STEBI self-efficacy data. Science Education, 95(2), 258–280.

Caramazza, A., McCloskey, M., & Green, B. (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about trajectories of objects. Cognition, 9, 117–123.

Chi, M. T. H. (2005). Commonsense conceptions of emergent processes: Why some misconceptions are robust. Journal of the Learning Sciences, 14(2), 161–199.

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

Eggert, S., & Bogeholz, S. (2009). Students' use of decision-making strategies with regard to socioscientific issues: An application of the Rasch partial credit model. Science Education, 94(2), 230–258.

Embretson, S. E. (1991). Implications of a multidimensional latent trait model for measuring change. In L. Collins & J. Horn (Eds.), Best methods for the analysis of change (pp. 184–201). Washington, DC: American Psychological Association.

Guilford, P. (1954). Psychometric methods. New York: McGraw-Hill.

Haladyna, T. M. (1999). Developing and validating multiple-choice test items (2nd ed.) Mahwah, NJ: Erlbaum.

Haladyna, T. M., & Rodriguez, M. C. (2013). Developing and validating test items. New York: Routledge.

Harris, C. J., Phillips, R. S., & Penuel, W. R. (2011). Examining teachers' instructional moves aimed at developing students' ideas and questions in learner-centered science classrooms. Journal of Science Teacher Education, 23.

Herrmann-Abell, C. F., & DeBoer, G. E. (2011). Using distractor-driven standards-based multiple-choice items and Rasch modeling to investigate hierarchies of chemistry misconceptions and detect structural problems with individual items. Chemistry Education Research and Practice, 12, 184–192.

Herrmann-Abell, C. F., & DeBoer, G. E. (2014). Developing and using distractor-driven multiple-choice assessments aligned to ideas about energy forms, transformation, transfer, and conservation. In R. F. Chen and A. Eisenkraft (Eds.), Teaching and learning of energy in K–12 education (pp. 103–133). Heidelberg, Germany: Springer.

Ioannides, C. H., & Vosniadou, S. (2002). The changing meanings of force. Cognitive Science Quarterly, 2(1), 5–62.

Kikas, E. (2004), Teachers' conceptions and misconceptions concerning three natural phenomena. Journal of Research in Science Teaching, 41, 432–448.

Klassen, S. (2006), Contextual assessment in science education: Background, issues, and policy. Science Education, 90(5), 820–851.

Lee, H., & Winke, P. (2012). The differences among three-, four-, and five-option-item formats in the context of a high-stakes English-language listening test. Language Testing, 30(1), 99–123.

Linacre, J. M. (2014). Winsteps (Version 3.80.1) [Computer software]. Chicago: Winsteps.com.

Linacre, J. M., & Wright, B. D. (1989). The "length" of a logit. Rasch Measurement Transactions, 3(2), 54–55.

Liu, X., & Boone, W. (2006). Introduction to Rasch measurement in science education. In X. Liu & W. J. Boone (Eds.), Applications of Rasch measurement in science education (pp. 1–22). Maple Grove, MN: JAM Press.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

Masters, G. N. (1982). A Rasch model for partial credit scoring. Psychometrika, 47(2), 149–174.

Mattson, M. (2004). Getting students to provide direction when drawing free-body diagrams. Physics Teacher, 42, 398.

Monsaas, J. A., & Engelhard, G., Jr. (1991). Examining changes in the home environment with the Rasch measurement model. In G. Engelhard Jr. & M. Wilson (Eds.), Objective measurement: Theory into practice (Vol. 3, pp. 127—140). Norwood, NJ: Ablex.

National Research Council (1996). National Science Education Standards. Washington, DC: National Academy Press.

National Research Council. (2007). Taking science to school: Learning and teaching science in grades K–8. Committee on Science Learning, Kindergarten Through Eighth Grade. In R. A. Duschl, H. A. Schweingruber, & A. W. Shouse (Eds.), Board on Science Education, Center for Education. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Penuel, W. R., & Fishman, B. J. (2012). Large-scale science education intervention research we can use. Journal of Research in Science Teaching. 49(3), 281–304.

Penuel, W. R., Fishman, B. J., Cheng, B. H., & Sabelli, N. (2011). Organizing research and development at the intersection of learning, implementation, and design. Educational Researcher, 40(7), 331–337.

Rasch, G. (1960/1980). Probabilistic models for some intelligence and attainment tests. Copenhagen , Denmark: Danish Institute for Educational Research. Expanded edition, Chicago: University of Chicago Press, 1980. (Original work published 1960).

Rogers, W. T., & Bateson, D. J. (1991). The influence of test-wiseness on performance of high school seniors on school leaving examinations. Applied Measurement in Education, 4(2), 159–183.

Rosengrant, D., Van Heuvelen, A., & Etkina, E. (2009). Do students use and understand free-body diagrams? Physics Education Research, 5, 1–13.

Roznowski, M., & Bassett, J. (1992). Training test-wiseness and flawed item types. Applied Measurement in Education, 5(1), 35–48.

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(3), 265–296.

Smith, R. M. (1987). Assessing partial knowledge in vocabulary. Journal of Educational Measurement, 24(3), 217–231.

Smith, R. M. (1992). Applications of Rasch measurement. Chicago: MESA Press.

Smith, R. M. (1997). Pre/post comparisons in Rasch measurement. In M. Wilson, G. Engelhard Jr., & K. Draney (Eds.), Objective measurement: Theory into practice (Vol. 4, pp. 297–312). Greenwich, CT: Ablex.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. Journal of Applied Measurement, 1(2), 199–218.

Sondergeld, T. A., & Johnson, C. C. (2014). Using Rasch measurement for the development and use of affective assessments in science education research. Science Education, 98, 581–613.

Smith, R. M., Schumacker, R. E., & Bush, J. J. (2000). Examining replication effects in Rasch fit statistics. In M. Wilson & G. Engelhard Jr. (Eds.). Objective measurement: Theory into practice (pp. 303–317). Stamford, CT: Ablex Publishing Corp.

Spelke E. S., Katz, G., Purcell, S. E., Ehrlich, S. M., & Breinlinger, K. (1992). Early knowledge of object motion: Continuity and inertia. Cognition, 51, 131–176.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics (2015). National Assessment of Educational Progress (NAEP) Questions Tool. Retrieved from http://nces.ed.gov/nationsreportcard/itmrlsx/

Van Heuvelen, A. (1991). Learning to think like a physicist: A review of research-based instructional strategies. American Journal of Physics, 59, 891.

Wang, W. (1998). Rasch analysis of distractors in multiple-choice items. Journal of Outcome Measurement, 2(1), 43–65.

Wilson, M. (1992). The ordered partition model: An extension of the partial credit model. Applied Psychological Measurement, 16(4), 309–325.

Wilson, M. (2005). Constructing measures: An item response modeling approach. Mahwah, NJ: Lawrence Erlbaum Associates.

Wilson, M. (2008). Cognitive diagnosis using item response models. Journal of Psychology, 216(2), 74–88.

Wilson, M. (2011). Some notes on the term: "Wright map." Rasch Measurement Transactions, 25(3), 1331.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis: Rasch measurement. Chicago: MESA Press.

Wright, B. D., & Stone, M. (1979). Best test design: Rasch measurement. Chicago, IL: MESA Press.

Wright, B. D. (2003). Rack and stack: Time 1 or Time 2 vs. post-assessment. Rasch Measurement Transactions, 17(1), 905–906.