Analyzing diagnostic items:
What makes a student response interpretable?

Caroline Wylie (ETS) and Dylan Wiliam (Institute of Education, University of London)

Paper presented at the annual meeting of the National Council on
Measurement in Education (NCME), April 2006, Chicago, IL.

The work reported in this paper was supported by the Institute for Education Sciences (IES) under grant number R305K040051, although the views expressed here are solely those of the authors.

Introduction

There is widespread evidence that teachers struggle to build contingencies into their classroom practice, and thus rarely pause to take stock of student learning during a lesson (Daws & Singh, 1996, HMCI, 2004). Even when information is gathered in the form of student responses to questions, it appears that teachers have great difficulty adjusting their instruction "on-the-fly" (Black & Wiliam, 1998; Kennedy, 1999). Furthermore in mathematics and science classrooms, procedural knowledge tends to be emphasized to the detriment of conceptual understanding (Hiebert *et al*., 2005). Thus while teachers are often very well aware of the errors that students make, they appear to not to distinguish between misremembering of an algorithm and those situations where more deep-seated conceptions are driving students' incorrect responses (Black & Atkin, 1996).

In response, the *Developing and Using Diagnostic Items in Mathematics and Science* (DIMS) project (IES reference R305K040051) has created a bank of 600 items for teachers to use during mathematics and science instruction that support "on-the-fly" teacher instructional decision making, and can also assist teachers in focusing on students' conceptual understandings. The item bank is composed primarily of multiple-choice questions where it is intended that both correct and incorrect responses are *interpretable* in ways that provide useful data on the current state of understanding of each individual in a class.

The items are not intended to be assembled into quizzes or tests, but rather are to be used individually during instruction to provide "real-time" information to guide the next instructional step. The fact that the items are to be used one at a time frees the item developers from many of the constraints traditionally associated with test items. For example, when multiple choice items are used, the numbers of options can be determined by the needs of the particular item, rather than being dictated by the format of the test. Furthermore, it is possible for items to have more than one correct answer, thus reducing the probability that a correct response is obtained by guessing. For example, where an item offers students five options, and where more than one option may be correct, then the probability of obtaining a correct response by guessing is around 3% instead of 20%, as would be the case with a single-best answer rubric.

As well as providing quality information about student achievement, the items have another purpose, which is the development of teacher knowledge. Research on mathematics instruction in the elementary grades found that mathematics teachers who posed problems, questioned students regarding their problem-solving strategies, and listened to students' solution strategies had a better understanding of their students' abilities and were better able to predict performance (Carpenter, Fennema, Peterson, Chiang, and Loef, 1989). In addition, students in these classes outperformed their peers on a mathematics achievement test. The authors hypothesized that the teachers' ability to understand the processes that students were using may have helped them to adapt instruction so that more appropriate activities were provided to students, resulting in higher achievement. A subsequent study (Peterson, Carpenter, and Fennema, 1989) examining two case studies found that the teacher who engaged in this type of questioning and was able to successfully predict student performance did report using her understanding of the students' knowledge when planning and adapting future instruction.

However, beginning teachers in particular may not have yet developed skills that expert teachers possess; "that specialized amalgam of content and pedagogy that is uniquely the province of the teacher, their own special form of professional understanding" (Shulman, 1987, p. 8). This specialized knowledge connects content with teaching strategies, and includes knowing how to present knowledge in multiple ways, and also knowing how novice learners might approach a topic differently from experts. One of the goals for the DIMS project is to help teachers develop one aspect of this "pedagogical content knowledge": a greater understanding of student misconceptions.

In this paper we will briefly describe questions were developed for the DIMS project, and then we will summarize the theoretical framework developed by Bart *et al.* (1994) for single-answer correct items, and extend it to items that have multiple correct answers. In particular, we will explore how the interpretability of correct answers interacts with the space of plausible incorrect answers.

Development of DIMS questions

To keep the volume of work manageable, we focused our work on the fourth and eighth grades. In order to ensure that the items developed were relevant to the needs of the majority of teachers, prior to the development of DIMS items we conducted a systematic review of state standards. This review identified important content objectives that were contained a large number of state content standards. Our original criterion was that a content objective had to appear in the content standards of at least 25 states to be included. This criterion worked well in mathematics, due to the considerable similarity of the state standards for mathematics. However in science, there was less consistency, partly because many state standards did not allocate content standards to particular grades but rather to bands of grades (e.g., grades 3 to 5), and partly because there was much more variety in the ways the content standards for science were organized. Nevertheless, by relaxing the constraints slightly, it was possible to develop a list of content standards for fourth grade math and science (74 and 47 respectively) and for eighth-grade math and science (66 and 88 respectively) that appears to be broadly relevant to most states. This analysis then shaped the second stage of work: the identification of student "misconceptions."

Before describing this work in detail, we think it is useful to clarify what we mean by the notion of "misconception." Traditionally, misconceptions have been treated as incorrect ideas that need to be eradicated, and authors adopting a constructivist stance on learning have rightly criticized such views. For us, a misconception is any conception that is consistently activated in a particular situation, and that results in a student giving an incorrect or incomplete response. For example, if a student is asked to plot the point with co-ordinates (3, 4) and in fact plots the point with co-ordinates (4, 3) that could be as a result of a stable (mis-)conception that the first number denotes the distance upwards, and the second number denotes the distance across, or it could be as a result of guessing, and happening to get it wrong. However if the student consistently does the same thing, then that would qualify as a misconception in our definition. In this case the student does have an incorrect belief that needs to be modified (and here, telling might just work). In other cases, the misconception might be much more deeply integrated into a set of beliefs. For example a student might give 2.30 as a response to 2.3 x 10 = ? This is likely to be attributable to a conception that to multiply by ten, one adds a zero, which give the correct result for integers, but not for other numbers. Here the "misconception" arises as a result of a conception that works well in some contexts, but has been overgeneralized. A similar example in science is the conception that the world is flat. This is an entirely appropriate conception for navigating over distances of tens, or even hundreds of miles—indeed it would be perverse to attempt to work with any other conception in such cases. However, such a conception is inappropriate when dealing with distances of thousands, or tens of thousands of miles. In what follows, therefore, the term "misconception" will be used to denote a range of phenomena from incorrect beliefs (e.g., that squares are not rectangles, or that humans and dinosaurs co-existed) through incomplete beliefs (that objects with density less than one float and greater than one sink) to conceptions that are valid in some contexts but not others (e.g., that the order in which you add up a series of numbers does not matter[1]).

Starting from content areas determined to be of interest from the standards analysis, an extensive review of research literature and teaching materials was conducted (e.g., diSessa, 2000; Driver, Squires, Rushworth, Wood-Robinson, 1994; Hart, Brown, Kerslake, Küchemann, Ruddock,1985; Russell & Watt, 1990; Stavy & Tirosh, 2000). In addition we conducted a number of consultation meetings with experienced practitioners, which yielded some further ideas of student misconceptions.

The writing of the diagnostic questions was then undertaken by test developers who simultaneously attended to both the content standards and the misconceptions. Each multiple-choice item was written to address the content areas identified by the standards review in such a way that at least some of the incorrect answer choices related to student misconceptions. Not every incorrect answer is connected to a misconception, but at least one misconception is addressed by every diagnostic item, so that the items are, in a sense *distractor-driven* (Sadler, 1995). Indeed, many of the item developers found a useful way of developing these items was to start by generating a distractor relating to a misconception, then to generate the item stem, and finally to generate the key and the other distractors. A final requirement was that there were plausible “answer-choice rationales” for the any distractors not keyed to misconceptions.

The number of response options was determined by the content of the items. Some had as few as three, while others had as many as eight. Since the items are not intended to be used in a written assessment, there was no constraint to standardize the format. Thus, instructional richness drove considerations about how many distractors to use for any particular item. If no student misconceptions could be identified for a content standard then an item to address that content area was not written.

Each item is presented in two formats: one is a single page with all the information that a teacher would need, and the second page is an overhead transparency so that the item is ready for immediate classroom use. Each teacher-view page clearly identifies the student misconceptions connected to an answer choice, and has an “additional comments” space which is used for additional information with regard to the content, or a specific way in which the question might be used that differs from other questions, or additional resources the teacher might look at, or to point out a recurring theme that cuts across items and contents areas.

The item developers worked in pairs (two for math and two for science). Each item developer wrote items that were then reviewed by the other. Periodically they then met with other project staff to scrutinize the latest batch of items. Almost invariably, revisions were then made, followed in some cases by further development, and in each case by a final round of editing. Items were then printed and sent out in batches to the pilot teachers throughout the year. Feedback from the pilot teachers informed a further round of revisions.

In the classes of participating DIMS teachers, every student was provided with a set of cards with the letters A to H printed on each card. The letter on each card is in a sufficiently large font to be seen from the front of the classroom. Typically, the question was displayed using an overhead projector, students individually selected their response(s) and then on cue held up one or more cards for the teacher to see. Thus the teacher could quickly and efficiently get a sense of the entire class response to the question, and could select students to probe their understanding based on their answer choices.

DIMS Items Meet Traditional Psychometrics

The relevance of traditional psychometrics for novel applications of assessment, especially applications of assessment to aid instruction, has been debated for many years. Popham and Husek (1969) argued that classical indices of reliability were inappropriate for criterion-referenced assessments, since high-quality instruction should reduce true-score variance, thus depressing the reliability of any assessment. Indeed, Bloom (1964) regarded reduction in true-score variance as one of the main aims of instruction.

Similar considerations apply to the development of distractor-driven multiple-choice items. It has long been known that the response choices or scales provided for ratings scales influence student responses in attitude questionnaires (see for example, Sedlmeier, 2006). It is not surprising, therefore, that Narode (1987) found that the inclusion of distractors related to student misconceptions made items more difficult, and lowered item discrimination. Sadler (1998) pointed out (p. 351) that a distractor-driven multiple-choice item on the reason for day and night taken from the Project STAR Astronomy Concept Inventory—a test designed for high school astronomy—actually had a lower facility for high school students than an item testing the same knowledge but without plausible distractors was for *third-grade* students (Figure 1).

These problems are not confined to classical test theory. Within the framework of item-response theory (see, for example, Hambleton & Swaminathan, 1985), a number of simplifying assumptions are made, which may be appropriate in some contexts, but are questionable in others. First, it is assumed that as the ability of a candidate increases, the probability of a correct response increases (Haladyna, 1994). In other words, it is assumed that the curve representing this relationship—the item characteristic curve (ICC)—is strictly monotonic increasing. Data reported by Sadler (1998) shows that the item shown on the left of

Figure 1 **does not fit this assumption. For the very lowest performing students, the probability of a correct response is around 60%. However, for students just below average ability, the probability of a correct response is almost 20 percentage points *lower*; these students are the most likely students to believe that the reason for day and night is that the earth moves around the sun (Sadler, 1998). For students of average, and above average, ability, increasing ability *is* associated with increasing probability of success on this item. However, it is important to note that for more deep-seated misconceptions, it is possible that the incorrect responses are the most commonly chosen by almost all students. A second item from the STAR test discussed by Sadler (1998) asked students why summer was hotter than winter (**

Figure 2). For this item only students in the top 5% of achievement were more likely to choose the correct answer than one of the other responses. Students of average achievement were most likely to choose option A than the correct option (these students were five times more likely to choose option A than the correct answer) while students from the 80th to the 95th percentile were most likely to choose option C.

Figure 1: Two items on the reason for day and night (adapted from Sadler, 1998)

Project STAR item 1 (grades 8 to 12)		
What causes night and day?		p-value
A	The earth spins on its axis	0.66
B	The earth moves round the sun	0.26
C	Clouds block out the sun’s light	0.00
D	The earth moves into and out of the sun’s shadow	0.03
E	The sun goes round the earth	0.04

1969 NAEP third-grade item		
One reason that there is day and night on Earth is that		p-value
A	the sun turns	0.08
B	the moon turns	0.04
C	the earth turns	0.81
D	the sun gets dark at night	0.06
E	I don’t know	0.01

Figure 2: Item on the reason for the seasons (Sadler, 1998)

Project STAR item 17 (grades 8 to 12)		
The main reason for its being hotter in summer than in winter is:		p-value
A	The earth’s distance from the sun changes	0.45
B	The sun is higher in the sky	0.12
C	The distance between the northern hemisphere and the sun changes	0.36
D	Ocean currents carry warm water north	0.03
E	An increase occurs in “greenhouse” gases	0.03

This raises a second (implicit) assumption in traditional item-response theory—that all incorrect responses are equivalent. In item-response theory, the only thing that matters about a student’s response is whether it is correct or not. If the answer is correct, that is evidence that the ability of the student is likely to be higher than the difficulty of the item. If the answer is incorrect, that is evidence that the ability of the student is likely to be lower than the difficulty of the item. However, as can be seen from the discussion above of the item in not all incorrect answers are equivalent. A student who chooses option C is likely to have higher ability than a student who chooses option A. A number of studies have explored the relationship between the ability of the student and the likelihood of each of the incorrect options. In other words, as well as the ICC, it is possible to plot "item category characteristic curves" for each of the incorrect responses (this highlights the fact that the term "item characteristic curve" is actually a misnomer, since the ICC is really the characteristic curve of only the correct response). Samejima (1969) proposed a "graded model" for items in which the responses are ordered by difficulty. He suggested that such items could be modeled by a series of parallel trace lines—one for each response option— in which each trace line described the probability that a student at a particular ability level chooses that particular response or a response judged as being more difficult. However, the fact that one response option is in some sense "harder" than another (as option C in is harder than option A) does not mean that the responses are ordered in the sense required by Sanejima's model. Bock (1972) proposed a more general approach—sometimes called the nominal model—for situations in which responses are scored in two or more latent categories without any assumption of an ordering amongst the responses. As Thissen (1976) showed, this approach can increase the accuracy of a test quite considerably, especially for lower-achieving students. However, while the use of incorrect responses can increase the accuracy of ability estimates, ability estimates by themselves are of limited value in instructional decision-making because two students with similar ability may hold completely different views about the concept at issue. For effective instructional action, the teacher needs to know what each student believes about, for example, the reason for night and day, and therefore what is needed is a classification, not a measurement.

If we were dealing with situations in which a number of items could be used, then Bayesian Inference Networks (Mislevy, Steinberg & Almond, 2003) would offer an attractive framework for analysis, but the requirements of real-time instructional decision-making mean that we have to deal with individual items. For this reason, we have explored the idea of the "semi-dense" item, proposed by Bart, Post, Behr and Lesh (1994).

Framework for Semi-Dense Items

Bart, Post, Behr, and Lesh (1994) set out the properties of a semi-dense item through a hierarchy of properties that items may possess. In this paper we will illustrate this framework though a series of mapping diagrams.

Bart *et al.* (1994) begin with a definition of a cognitive rule as "any sequence of cognitive operations that produces an item response from the processing of an item stem. Cognitive rules include problem solving strategies, decision making strategies and algorithms." Thus a correct answer to a problem is generated by a cognitive rule, and so too are incorrect answers. The collection of cognitive rules that are associated with an item is described as the "cognitive microtheory" for that item. The Bart *et al.* framework establishes a series of conditions describing how the cognitive microtheory relates to item responses. They describe the characteristics of a semi-dense item in terms of five properties: response interpretability, response discrimination, rule discrimination, exhaustive set usage, and semi-density.

The properties are not mutually exclusive, but some are preconditions for others. The first three properties (response interpretability, response discrimination, and rule discrimination) form a hierarchy of increasing stringency, while the fourth property (exhaustive set usage) can be applied to the second and third property as an additional constraint.

These properties are most easily understood through a series of diagrams. In Figure 3 through Figure 7 the cognitive rules are represented on the left of the figure and the answer choices on the right. The arrowed lines illustrate the (one or more) cognitive rules that interpret each answer choice.

Figure 3 illustrates the property of response interpretability, where each response is interpretable by at least one cognitive rule. This property however does not preclude the possibility that one rule could interpret more than one answer choice, nor that two separate cognitive rules could explain one response. In addition there are cognitive rules in the microtheory connected to the item that may not explain any of the answer choices.

Figure 3: Response Interpretability

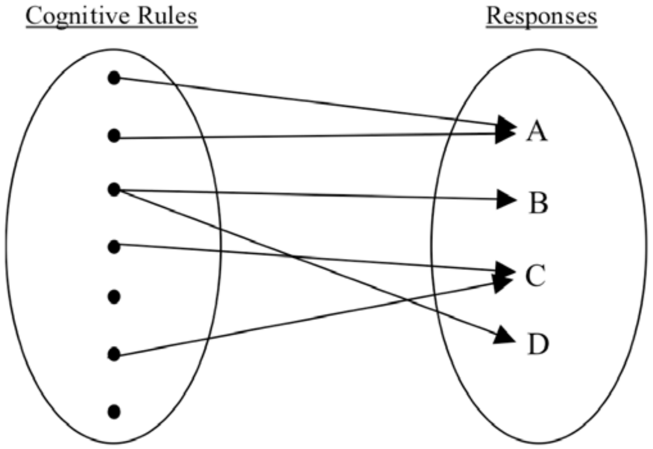
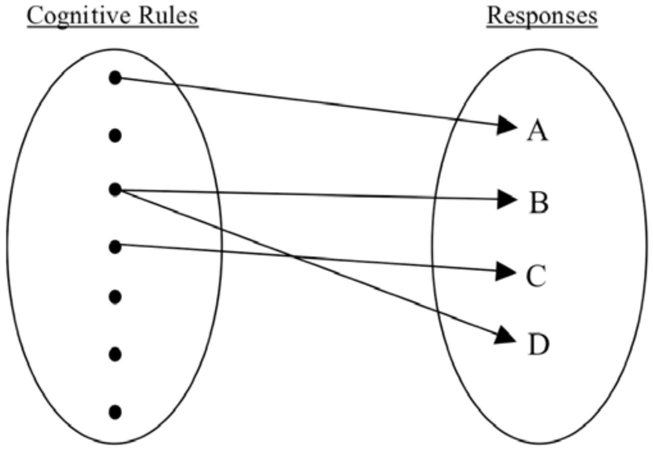


Figure 4 below illustrates the property of response discrimination. In this situation each response is interpreted by only one cognitive rule.

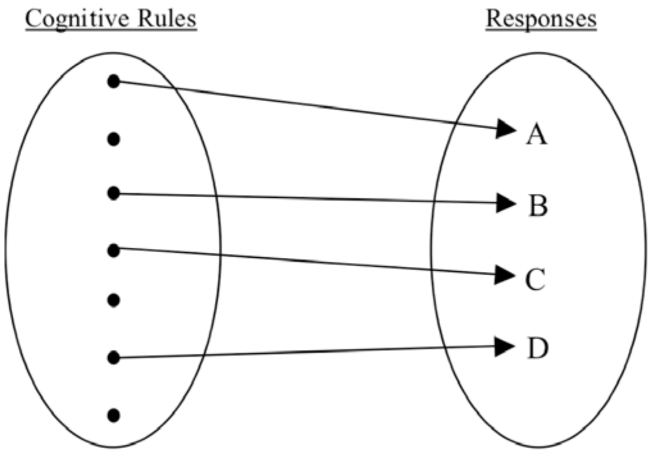
Figure 4: Response Discrimination



Comparing Figure 3 with Figure 4 responses A and C are no longer interpreted by two separate cognitive rules, although it is still possible for one cognitive rule to lead to two separate answer choices.

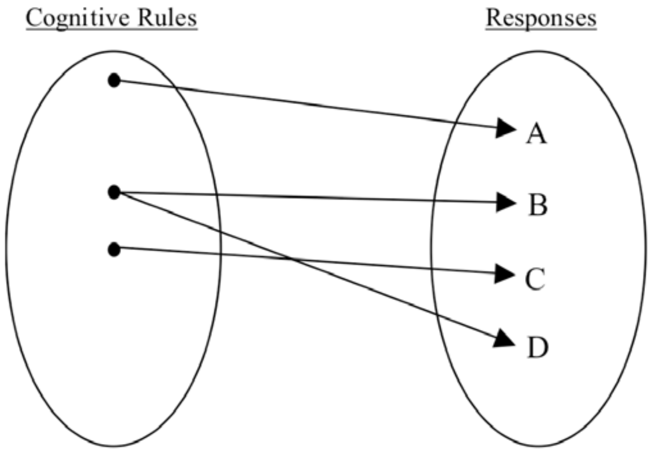
Figure 5 illustrates a further constraint on the relationship between cognitive rules and responses to the item in that in order to have rule discrimination each cognitive rule interprets only one response. Thus whereas in Figure 4 the same rule provided an interpretation for both answers B and D, that cannot be the case for an item to possess the property of rule discrimination as shown in Figure 5.

Figure 5: Rule Discrimination I



Another property that can define the relationship between the cognitive rules and item responses is that of exhaustive set usage, which means that every rule in the cognitive microtheory connected to a particular item interprets at least one response to the item. So unlike Figure 3 through Figure 5 where cognitive rules do not have to connect to a response, in Figure 6, every cognitive rule explains a response.

Figure 6: Exhaustive Set Usage

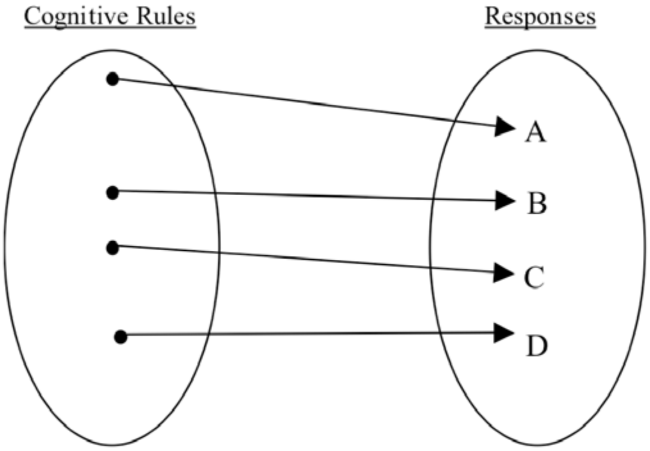


By definition, the property of exhaustive set usage cannot be applied to items that only have response interpretability, but starts with items that have response discrimination.

The final property of semi-density then describes an item that has the property of rule discrimination and exhaustive set usage, resulting in the illustration shown in

Figure 7.

Figure 7: Semi-density



In order to illustrate how the Bart *et al*. framework can be applied to items we consider two items from the DIMS project along with an example from the Bart *et al.* (1994) paper below. All three items are presented with two versions: the initial question and a second improved question that has better diagnostic properties.

Figure 8: Two versions of an algebra question

Version 1	Version 2
If $e+f = 8$, then $e+f+g =$	If $f+g = 8$, then $f+g+h =$
a) 9	a) 9
b) 12	b) 12
c) 15	c) 15
d) $8+g$	d) 16
	e) $8+g$

In version 1 of the question in Figure 8, a cognitive rule for each answer choice can be posited. A student who selects option (a) may have reasoned that since one more item has been added to the original equation, the new answer must be one more than the original answer. For option (b) the student has probably reasoned that *e* and *f* must be each equal to 4 since that is the most straightforward way to get a total of 8, and hence *g* must also be of value 4 to give a total of 12. To arrive at answer (c) one possible argument would be that *e* and *f* must be of value 3 and 5 respectively, since that is the next easiest way to achieve a total of 8, where *e* and *f* are of different values, and then since *g* is the next letter in the alphabet, and 7 is the next value in the number pattern, the total answer is 3+5+7 which is 12. However, a second explanation for why a student might decide that *g* has a value of 7 is simply that *g* is the seventh letter of the alphabet.

Version 2 of the question is similar except for the change in the letters. The explanations for why a student might incorrectly select answer choices (a) or (b) are the same as for version 1. Answer (c) can now only be selected if a student is assuming that f=3, g=5, and h=7. If a student thought that the value of *h* could be determined by counting its place in the alphabet they would assume that *h* was of value 8 which would result in the incorrect answer of 16.

To consider this question using the Bart *et al.* (1994) framework we could say that version 1 of the question has response interpretability since each response or answer choice can be interpreted by at least one cognitive rule, but since two rules provide explanations for answer (c), version 1 does not have response discrimination. And if we assume that the set of cognitive rules is exhaustive then this item could be considered to be semi-dense.

Figure 9: Two versions of a time question

Version 1

"There are two flights per day from Newtown to Oldtown. The first flight leaves Newtown each day at 9:20 and arrives in Oldtown at 10:55. The second flight from Newtown leaves at 2:15. At what time does the second flight arrive in Oldtown? Show your work."

Version 2

"There are two flights per day from Newtown to Oldtown. The first flight leaves Newtown each day at 9:05 and arrives in Oldtown at 10:55. The second flight from Newtown leaves at 2:15. At what time does the second flight arrive in Oldtown? Show your work."

The example question shown in Figure 9 illustrates how a question was modified in a very small way, but with a significant improvement in the interpretability of the answers. In the first example a student who had a solid understanding of time would subtract 9:20 from 10:55 in order to find the duration of the first flight, and then add the difference on to the departure time of the second flight to determine the second flight's arrival time of 3:50. However, a student who did not fully understand that time must be manipulated differently from base 10 problems could complete this particular problem using base 10 strategies. The only modification to version 2 of the question was an adjustment to the departure time of the first flight. As a result, a student who used a base ten approach to solve the problem might answer 2.15+(10.55-9.05)=3.65.

As a result of the modification in version 2 of this question, a teacher can be more confident that a student who answers the question correct actually understands the content, as opposed to version 1 where a student could get to the wrong answer through faulty logic.

To add a third example of a question to this discussion, let us consider the example provided in the Bart *et al.* (1994) paper in which students might arrive at the correct answer through one of six approaches. The question posed is shown in Figure 10 below.

The correct answer is 24 cents, which a student could arrive at using a variety of approaches such as unit rate, factor of change, cross multiplication algorithm, equivalent fractions, equivalence class, pair generation. An inappropriate additive model would lead a student to the incorrect answer of 12. In the original version of the question students are presented with two answer choices, namely 12 and 24. Applying the semi-density framework to version 1 of this question, it has response discrimination since each answer choice is interpretable by one or more cognitive rule, but because the correct answer is interpretable by six cognitive rules the question does not have response discrimination, and thus is does not possess any of the other properties of semi-density. In version a teacher can identify which reasoning process the student used based on answer selection.

Consider now the three questions presented in terms of the potential impact on a teacher's instructional decision making. The first question presented in Figure 8 started out with an incorrect answer that could be explained by two cognitive rules, both of which were incorrect cognitive rules. The original version of the question in Figure 9 had a correct answer that could be explained by a correct cognitive rule but also by an incorrect cognitive rule, and finally version 1 of the question excerpted from the Bart *et al.* paper (Figure 10) had a correct answer that could be described by six correct cognitive rules. While version 1 of all three of these questions had only response discrimination, from an instructional perspective some questions were less in need of refinement than others.

Figure 10: Two versions of an arithmetic question

Version 1

Ann and Kathy each bought the same kind of bubble gum at the same store. Ann bought two pieces of gum for six cents. If Kathy bought eight pieces of gum, how much did she pay?

(a) 12

(b) 24

Version 2

Ann and Kathy each bought the same kind of bubble gum at the same store. Ann bought two pieces of gum for six cents. If Kathy bought eight pieces of gum, how much did she pay?

(a) 24 cents, because 8×3=24 or 8 pieces × 3 cents per piece=24 cents.

(b) 24 cents, because 4×6=24.

(c) 24 cents, because 2/6=8/x and 2x=48 for which x=24.

(d) 24 cents, because 2/6=8/x and 2/6x4/4=8/24.

(e) 24 cents, because 2/6=4/12=6/18=8/24.

(f) 24 cents, became 2/6=4/12=8/24.

(g) 12 cents, because 2+4=6 implies that 8+4=12

Figure 11: Description of three items

Question	Version 1 of three questions
e+f=8	An incorrect answer explained by two inappropriate cognitive rules
Travel time	A correct answer explained by both an appropriate and an inappropriate cognitive rule
Cost of candy	A correct answer explained by many correct cognitive rules

In order to make informed instructional decisions that will engage students in deeper learning, the critical distinction is between students who are operating with a correct or an incorrect cognitive rule: it is less critical to distinguish between correct cognitive rules. To put it another way, in instruction, it is far more serious to assume that students do understand something when they don't, than to assume they don't understand something when they do.

Thus of the three examples provided although they all only had response discrimination, in terms of instructional use, the question illustrated in Figure 10 was the one that was least important to modify to increase its diagnostic power.

In terms of directing student learning specifically, being able to distinguish between incorrect cognitive rules may help the teacher guide the next instructional step. Thus, a teacher would have been better informed using the improved algebra question that was illustrated in Figure 8, since her instructional strategy could vary according to how students were imputing a value for *h*.

However, the item that was most essential to correct was the time question shown in Figure 9 since in the original version, the correct answer could be achieved using both a correct and an incorrect cognitive rule. Some students who would give the correct answer could in fact be false positives, and pose greater instructional risk than false negatives.

Conclusion

The Developing and Using Diagnostic Items in Mathematics and Science (DIMS) project has developed a series of 600 items for fourth and eighth grade mathematics and science that teachers can use individually to support real-time instructional decision making. Since the items are designed to be used individually, much of the framework of traditional psychometric theories is not particularly useful for analysis. For this reason, we have begun to explore the relevance of the concept of a semi-dense item proposed by Bart, Post, Behr and Lesh (1994) as a theoretical tool for analyzing the DIMS items. The idea of a semi-dense item provides a useful starting point, but for items that support instruction, we argue that a modification of the Bart et al framework is required. For an item to support instructional decision-making, the key requirement is that in no case do incorrect and correct cognitive rules map on to the same response. If this property is met, then we would argue that the five properties identified by Bart *et al.* are less important. If this property is not met, then we would argue that the five properties identified by Bart *et al*. are irrelevant.

References

Bart, W. M., Post, T., Behr, M. J., & Lesh, R. (1994). A diagnostic analysis of a proportional reasoning test item: An introduction to the properties of a semi-dense item. *Focus on Learning Problems in Mathematics,* **16**(3), 1-11.

Black, P. J., & Atkin, J. M. (Eds.). (1996). *Changing the subject: innovations in science, mathematics and technology education*. London, UK: Routledge.

Black, P. J., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles Policy and Practice,* **5**(1), 7-73.

Bloom, B. S. (1964). *Stability and change in human characteristics*. New York, NY: John Wiley & Sons.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more latent categories. *Psychometrika,* **37**, 29-51.

Carpenter, T. P., Fennema, E., Peterson, P. L., Chiang, C., Loef, M. (1989). Using Knowledge of Children's Mathematics Thinking in Classroom Teaching: An Experimental Study. *American Educational Research Journal,* **26** (4): 499-531.

Ciofalo, J., & Wylie, C. E. (2006). Using diagnostic classroom assessment: one question at a time. *Teachers College Record,* Date Published: January 10, 2006. http://www.tcrecord.org ID Number: 12285, Date Accessed: 8/7/2006 11:49:49 AM

Daws, N., & Singh, B. (1996). Formative assessment; to what extent is its potential to enhance pupils' science being realized ? *School Science Review,* **77**(281), 93-100.

diSessa, A. (2000). Changing Mind: Computers, Learning, and Literacy. MA: MIT Press

Driver, R. A., Rushworth, S. P., & Wood-Robinson, V. (1994). *Making sense of secondary science: Research into children's ideas.* London and New York: Routledge Falmer.

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

Hart, K., Brown, M., Kerslake, D., Küchemann, D., Ruddock, G. (1985). *Chelsea Diagnostic Mathematics Tests: Teacher's Guide.* NFER-Nelson. Berkshire, UK.

Havil, J. (2003). *Gamma: exploring Euler's constant*. Princeton, NJ: Princeton University Press.

Hiebert, J., Stigler, J. W., Jacobs, J. K., Givvin, K. B., Garnier, H., Smith, M., Hollingsworth, H., Manaster, A., Wearne, D., & Gallimore, R. (2005). Mathematics teaching in the United States today (and tomorrow): results from the TIMSS 1999 Video Study. *Educational Evaluation and Policy Analysis*, **27**(2), 111-132.

Kennedy, M. (1999). The role of preservice education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as the learning profession* (pp. 54-85). San Francisco, CA: Jossey-Bass.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston, MA: Kluwer Academic Publishers.

Her Majesty's Chief Inspector of Schools. (2004). *Annual report 2003-2004*. London, UK: Her Majesty's Stationery Office.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: interdisciplinary research and perspectives*, **1**(1), 3-62.

Narode, R. (1987). Standardized testing for alternative conceptions in basic mathematics. In J. D. Novak (Ed.), *Second international seminar on Misconception and Educational Strategies in Science and Mathematics* (Vol. 1, pp. 222-333). Ithaca, NY: Cornell University Press.

Peterson, P.L., Carpenter, T., & Fennema, E. (1989). "Teachers' Knowledge of Students' Knowledge in Mathematics Problem Solving: Correlational and Case Analyses." *Journal of Educational Psychology*, **81**(4): 558-569.

Popham, W. J., & Husek, T. R. (1969). Implications of criterion referenced measurement. *Journal of Educational Measurement*, **6**(1), 1-9.

Russell, T., Watt, D. (1990). *Primary SPACE (Science Processes and Concept Exploration) Project Research Reports: Evaporation and Condensation*. Liverpool University Press.

Sadler, P. M. (1992). *The initial knowledge state of high school astronomy students.* Unpublished Ed.D. dissertation, Harvard University.

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.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph*, **18**.

Sedlmeier, P. (2006). The role of scales in student ratings. *Learning and Instruction*, **16**(401-41), 415.

Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57*, 1-22.

Stavy, R., & Tirosh, D. (2000). *How students (mis-)understand science and mathematics: Intuitive rules.* New York: Teachers College Press.

Thissen, D. M. (1976). Information in wrong responses to the Raven's progressive matrices. *Journal of Educational Measurement,* **13**(3), 201-214.

Wylie, E. C., & Wiliam, D. (2006). *Diagnostic questions: is there value in just one?* Paper presented at the Annual Meeting of the National Council on Measurement in Education held at San Francisco, CA.

Notes

[1] While the order of addition is irrelevant for any *finite* list of numbers, it turns out that the terms of some infinite sums (called conditionally convergent) can be re-arranged to yield a different total. In fact, Bernhard Riemann showed that if an infinite sum can be re-arranged to give a different sum, it can be re-arranged to give any sum whatsoever (Havil, 2003).