ICAT: An Adaptive Testing Procedure for the Identification of Idiosyncratic Knowledge Patterns

G. Gage Kingsbury and Ronald L. Houser

Northwest Evaluation Association, Lake Oswego, OR, USA

Abstract. Traditional adaptive tests provide an efficient method for estimating student achievements levels, by adjusting the characteristics of the test questions to match the performance of each student. These traditional adaptive tests are not designed to identify idiosyncratic knowledge patterns. As students move through their education, they learn content in any number of different ways related to their learning style and cognitive development. This may result in a student having different achievement levels from one content area to another within a domain of content. This study investigates whether such idiosyncratic knowledge patterns exist. It discusses the differences between idiosyncratic knowledge patterns and multidimensionality. Finally, it proposes an adaptive testing procedure that can be used to identify a student's areas of strength and weakness more efficiently than current adaptive testing approaches. The findings of the study indicate that a fairly large number of students may have test results that are influenced by their idiosyncratic knowledge patterns. The findings suggest that these patterns persist across time for a large number of students, and that the differences in student performance between content areas within a subject domain are large enough to allow them to be useful in instruction. Given the existence of idiosyncratic patterns of knowledge, the proposed testing procedure may enable us to provide more useful information to teachers. It should also allow us to differentiate between idiosyncratic patterns or knowledge, and important mutidimensionality in the testing data.

Keywords: adaptive testing, item response theory, diagnostic testing

A wide variety of procedures have been suggested for selecting items within an adaptive test. These have ranged from branching procedures like the stradaptive test (Weiss, 1976) to procedures that maximize information (Samejima, 1977). The earlier adaptive testing procedures were designed for the measurement of ability, using content domains without appreciable variation from one question to the next. As adaptive testing moved into education and licensure, the specific content sampled within the domain became quite important. A variety of procedures have been developed to allow the adaptive test to satisfy a set of content constraints as it selects items. These approaches range from simple procedures to match a single set of constraints (Kingsbury & Zara, 1991) to procedures that meet a wide variety of constraints by creating shadow tests (van der Linden & Chang, 2003). These procedures allow the test developer to specify a particular test blueprint, even though the particular items that a test taker sees are not specified. This is a very reasonable approach in a wide variety of settings in which the construct being measured is unidimensional, but the items in the domain can be differentiated by their content characteristics.

Most of the approaches to content constraints within adaptive tests have developed from the idea that an adaptive test should have the same content controls that we find in a fixed-form test. While this is a useful idea, it limits the development of adaptive tests to the characteristics imposed by a fixed test form. It is useful to move beyond these constraints if we want to improve measurement of individual characteristics.

As we move into educational tests that examine how students grow from beginning readers to accomplished adults, the ability to isolate patterns of growth becomes extremely important. A number of authors have researched different aspects of the use of diagnostic tests to help identify the particular misunderstandings or weaknesses that a student has. For instance, Tatsuoka and Tatsuoka (1997) suggested a way of identifying an individual's location within a "rule space" to enable the identification of specific misunderstandings that the person was using in answering questions. Bart (2007) has described a method for developing items which disclose misunderstandings dependent on the particular answer that a student gives to the item. Rudner and Talento-Miller (2007) have described a statistical procedure for identifying goal areas or types of items that are causing a student more difficulty. Each of these developments has contributed to our ability to identify unique performance by test takers so that we might intervene in instruction, make suggestions for individual improvement, or consider program improvements.

If we assume that idiosyncratic patterns of knowledge exist, as described in the work of the authors above, then we should be able to design a testing process to enhance our ability to identify and measure the degree to which an

DOI 10.1027/0044-3409.216.1.40

individual's performance differs from one goal area to another. That is the purpose of this study.

For this work, we postulate that a student can have consistent and pervasive relative weaknesses (or strengths) in one or more goal areas within a domain of content. We call these individual patterns of strength and weakness *idiosyncratic knowledge patterns*. These patterns may appear due to differential instruction, differential motivation, or many other causes. It is likely that these idiosyncratic knowledge patterns can be affected by instruction, if they are accurately identified.

This study will detail a live-data study addressing the existence and stability of such idiosyncratic patterns using existing adaptive testing procedures. It will then describe an approach to adaptive testing that improves the measurement of idiosyncratic knowledge patterns.

Idiosyncratic Knowledge Patterns and Dimensionality

When students are learning a subject in school, they take a variety of paths. In mathematics, one student may have a predilection for geometry while another student may excel in solving equations. A third student might have equal skill in both concepts. All three paths may be considered idiosyncratic knowledge patterns. The third student, with homogeneous knowledge across goal areas, may be served well by homogeneous instructional practices. At the same time, the other two students may need different approaches or resources, and a teacher would benefit from having this information.

As students have an opportunity to learn the content, a common pattern of knowledge is likely to emerge. Our item response models should identify this pattern empirically, and use it to calibrate the test questions to a common scale or response space. It should be noted that the presence of idiosyncratic knowledge patterns does not imply that the response space is multidimensional. An example helps clarify this distinction:

– Consider a 30-item mathematics test made up of 10 geometry questions, 10 algebra questions, and 10 computation questions.
– If we collect data from 1000 students who have received instruction in the three goal areas, we may assess the dimensionality of the response space.
– The dimensionality of the response space describes how the population of interest (Lord & Novick, 1968) will respond to the questions on the test. This means that dimensionality is a population characteristic.
– The 1000 students from whom we collected the test data represent the population of interest, but they also represent 1000 different individuals. The things that make them unique are much more varied and pervasive than the things that make them a *population*.
– Variations in learning patterns from one test taker to another would influence how they would answer individual questions, but wouldn't create dimensions in the response space, since by definition they vary from person to person. If 20 of the students in the sample were somewhat higher performing in geometry than they were in computation and algebra, this would be useful information about the 20 students, but would not change the overall dimensionality in the population noticeably.
– For an individual with a pattern of knowledge that is different from the common pattern in the population of interest, we should see the difference reflected in reduced person fit. If the student tended to be more knowledgeable in geometry than in the other two subject areas (relative to the population of interest), person fit would decline because the individual would answer difficult geometry items correctly more often than the response model would predict.
– As the number of students with idiosyncratic knowledge patterns increases, the number of students with lower person fit would also increase. If the number of individuals with idiosyncratic knowledge patterns was high enough, this might be reflected by a reduction in the strength of the first common factor and a reduction in the consistency of item parameter estimates.

The presence of idiosyncratic knowledge patterns does not violate unidimensionality or essential unidimensionality, but it may make detection of dimensionality more difficult.

A way of conceptualizing idiosyncratic knowledge patterns is that while the item response theory (IRT) model applies to the entire population of students, some individual students have different trait levels from one goal to another. This means that for those individual students, the IRT model that describes the latent space for the population is a less-than-complete representation of student achievement.

Identifying Idiosyncratic Knowledge

Since we have described idiosyncratic knowledge as variability among goal areas within a student, we have a clear path for identifying students who have idiosyncratic patterns. For purposes of this study, we will define an idiosyncratic knowledge pattern as follows:

> An idiosyncratic knowledge pattern is one in which a student's achievement level in one goal area on a test differs by more than expected from the achievement level that the student displays on the test as a whole.

Since we can't observe the achievement levels directly, we will examine the achievement level estimates from goal areas within the example test, and compare these to the overall achievement level estimate. Because the information available in these estimates differs fairly substantially, the error of measurement for each of these estimates must be accounted for in determining whether a difference is a knowledge pattern or simply noise in the data. For this study, a goal area achievement level estimate was identified as being different from the overall achievement level estimate if the difference between the point estimates was greater than the sum of the associated standard errors. (A variety of other rules could be used, but this one has the advantage of taking into account both standard errors, which differ due to the number of items included in each goal area, the number of items represented in the overall score, and the characteristics of the items available in the item pool.)

In other words, a goal achievement level estimate differs from the overall achievement level estimate if

$$|\hat{\theta}_j - \hat{\theta}| > w_j * e_j^2 + w * e^2 \quad (1)$$

where w is the weighting of the overall standard error, and w_j is the weighting for the standard error for goal j. This definition describes the phenomenon, but doesn't suggest that it is pervasive or that it is persistent. Given a set of weights, the goal achievement level estimate will diverge from the overall achievement level estimate due to sampling variability with known frequency. To the extent that the observed frequency is greater than the expected frequency given sampling fluctuation, the phenomenon may be of interest.

Unfortunately, current adaptive testing procedures aren't optimized to identify idiosyncratic knowledge patterns, if they exist. Instead, they maximize information at the current overall achievement level estimate. Later in this study, we will describe an adaptive testing procedure optimized for identification of such patterns (the idiosyncratic computerized adaptive testing [ICAT] procedure), but first we need to determine whether they are worth identifying.

Real-Data Examination of Idiosyncratic Knowledge Patterns

The ICAT procedure described below will only be useful if (a) idiosyncratic knowledge patterns exist, and (b) these patterns show enough consistency across time to be used to make instructional decisions. In order to identify whether and to what extent idiosyncratic knowledge patterns exist and might be useful in instruction, this study examines data from operational adaptive tests. This study examines an existing set of adaptive-testing data to determine whether evidence of idiosyncratic knowledge patterns exists and whether these patterns persist across time for students.

Student Sample

A sample of 142,301 students enrolled in grades 3 to 8 in a state in the southeastern portion of the United States in calendar years 2005 to 2006 was obtained. Each student took the tests described below as a portion of a required assessment given to all students in the school district. Each student in the sample took tests in mathematics in spring of 2005 and fall of 2005, or spring of 2006 and fall of 2006. Each student then had two tests that were separated by summer vacation and a month or two of instruction. This design allowed us to identify whether idiosyncratic knowledge patterns existed and to test whether they persisted across time and instruction.

Tests

The tests used were the Measures of Academic Progress (MAP; NWEA, 2003). These tests were content-balanced adaptive tests testing five goal areas in mathematics. Each student's test consisted of 50 precalibrated items drawn using a constrained Bayesian item selection procedure. Maximum-likelihood scores were calculated for the overall test and each of the goal areas following the test. These maximum likelihood scores are used for all of the analyses in this study. The five goal areas were:
- G1. Number and Operations
- G2. Algebra
- G3. Geometry
- G4. Measurement
- G5. Data Analysis and Probability

The test drew items from a pool of approximately 1,500 items that were selected to align with the mathematics content standards in the state and sculpted to avoid inappropriate content for the students (each item was associated with one – and only one – goal area). All items were calibrated to the RIT scale, a unidimensional one-parameter logistic (Rasch) scale that has been shown to be very stable across long periods of time (Kingsbury, 2003). The RIT scale is similar to the theta scale, with a translation that gives it an practical range of scores from 120 to 300 (120 would correspond to a student just learning numbers, while 300 would correspond to a student ready for calculus).

It should be noted that these tests do not use the ICAT algorithm that is described below, but use a traditional adaptive testing approach. All analyses were conducted using final achievement level estimates. While the tests have not been optimized to measure idiosyncratic knowledge patterns, they should be adequate to identify whether such patterns exist in students and persist across time.

Identification of Idiosyncratic Patterns

In this study, the weightings in Equation 1 were set to unity. Any individual with an achievement level estimate in a goal that met the requirements of Equation 1 was identified as having idiosyncratic knowledge in that goal. This resulted in identification of idiosyncratic knowledge patterns when differences in achievement level estimates are close to 10 points on the RIT scale, which is the measurement scale used by the MAP tests. (The adaptive nature of the tests and the deep item pools result in fairly consistent standard errors across virtually all of the student achievement range. The average standard error for the overall achievement level is 2.9 RIT points for this sample, while the standard error for the achievement level estimate in each goal averages 6.8 RIT points.)

Differences larger than 10 RIT points can have instructional implications for the student and may be useful for differentiating instruction, so the choice of unit weights would have practical significance in this application. As we learn more about the nature of idiosyncratic knowledge patterns, information about their distribution and reliability may suggest different weightings for the standard errors.

Analysis

The study addressed two primary questions. First, it examined the extent to which goal achievement level estimates differ from the overall achievement level estimate within the spring test. Second, it examined whether the same students tended to have the same idiosyncratic patterns from the spring test to the test given the following fall.

To identify a goal area which was idiosyncratically high or idiosyncratically low for a student, the signed difference between the overall achievement level estimate and the achievement level estimate for each goal area was calculated. If the magnitude of this difference for a particular goal was greater than the sum of the standard errors for the two achievement level estimates, the student was identified as having an idiosyncratic score in that goal area according to Equation 1. The frequency and distribution of these idiosyncratic patterns were examined within the spring test.

To the extent that idiosyncratic knowledge patterns occur during the first testing, and reoccur during the second testing, we may be seeing evidence of a trait of the student, rather than a state of the student within a test setting. To investigate this issue, patterns of performance for the same students were studied with the spring test and the fall test in order to determine whether idiosyncratic patterns of performance persisted within students across time.

Results

The Existence of Idiosyncratic Knowledge Patterns

Table 1 shows the frequency with which students in each grade were identified as having at least one goal area in which they demonstrated idiosyncratic knowledge during the spring test. It can be seen that approximately half of the students tested had at least one goal that was identified as being substantially different than the overall achievement level estimate.

Table 1. Students with at least one goal that differed from the overall achievement level estimate during the spring test

Grade	No. students	No. idiosyncratic	% idiosyncratic
3	43506	20922	.48
4	31551	15387	.49
5	22840	11235	.49
6	22880	10925	.50
7	21524	10469	.50

If we assume that achievement is normally distributed in the population, with a standard deviation equal to the observed standard deviation in the sample, then we can calculate the expected percentage of students having an achievement level estimate in a particular goal that is identified as idiosyncratic. The rule for identifying an achievement level estimate for a goal as idiosyncratic is that it differs from the overall achievement level estimate by more than the sum of the standard errors of the estimates.

The probability of observing a difference this great or greater due to chance fluctuation is approximately .0506, given independent events. (The probability of each estimate being at least one standard error from the true value in a given direction is .159, squared for the two independent events is .0253, times 2 for the two directions in which the scores can differ, yields .0506.) Since five goal areas are included in the test, the probability of observing at least one idiosyncratic goal by chance, given independent goal scores would be .229.

However, the goal achievement level estimates are not independent. They have a strong, positive correlation of approximately .7 with the overall achievement level estimate. This positive covariance among the achievement level estimates accounts for approximately 50% of the variance and reduces the probability of observing an individual with at least one goal score that differs from the overall score by chance even further.

From Table 1, we see that in each grade our sample has 48% to 50% of the students identified as having idiosyncratic knowledge patterns, compared with the 23% or less that we should see due to chance fluctuation. Given these observations, it is unlikely that chance variation has created the number of scores that are observed in the sample data. It appears that the patterns that we observe indicate the existence of idiosyncratic knowledge patterns in at least 25% of the sample. While this percentage is likely to differ with the subject being assessed, the number of reported goal categories, and the student population, it is clear that idiosyncratic patterns do exist within our sample data.

Table 2 shows the percentage of students idiosyncratically high and idiosyncratically low in each goal area for students in each grade in the sample. It can be seen that patterns of idiosyncratic performance emerge by grade and

Table 2. Percentage of students with idiosyncratically high and idiosyncratically low achievement level estimates in each goal area (G1 to G5)

Grade	No. student	High G1	Low G1	High G2	Low G2	High G3	Low G3	High G4	Low G4	High G5	Low G5
3	43506	11.4	2.8	5.2	7.8	3.5	12.6	8.0	4.8	6.0	6.6
4	31551	9.9	3.2	4.8	8.7	3.7	12.3	7.4	5.1	8.2	5.8
5	22840	8.1	4.5	6.1	7.5	4.1	13.2	8.6	4.4	7.6	5.9
6	22880	7.3	4.9	6.5	6.5	4.4	10.2	11.3	3.1	4.3	9.2
7	21524	7.3	5.2	6.8	6.5	4.5	9.4	13.1	2.8	3.6	10.5
All	142301	9.3	3.9	5.7	7.5	3.9	11.8	9.3	4.3	6.1	7.3

by goal area. For example, Goal 3 (G3 – Geometry) has two to three times as many students with idiosyncratically low achievement level estimates as with estimates that were idiosyncratically high. This pattern is consistent across all grades. On the other hand, Goal 5 (G5 – Data analysis and probability) shows a pattern with more idiosyncratically high estimates in lower grades and more idiosyncratically low estimates in upper grades. It is certainly the case that some content in mathematic is more challenging to teach. It remains to be seen whether the patterns of student performance tend to match patterns of teacher behavior.

The Stability of Idiosyncratic Knowledge Patterns

Table 3 shows the intercorrelations of the level of surprise for each goal area in spring and fall across all grades tested. For each goal area in each test, the amount that the goal area achievement level estimate in a particular area differed from the overall achievement level estimate on the test was calculated (the level of surprise in the goal score). These values for each goal area on the spring test were then correlated with the values from the fall test. If goal area achievement level estimated tended to diverge from the overall achievement level estimate in the same direction from spring to fall, it would indicate a persistent pattern, with students being relatively high or relatively low in the same goal areas across time. None of the intercorrelations are very large, as would be expected in an essentially unidimensional trait. At the same time, the largest intercorrelations occur within the common goals across the two testing seasons (in Table 3 in **bold**).

Table 3. Intercorrelations of goal area achievement level estimate surprise level in spring and fall tests

	Fall G1	Fall G2	Fall G3	Fall G4	Fall G5
Spring G1	**.06**	–.00	–.05	–.02	.01
Spring G2	.01	**.10**	–.01	–.05	–.05
Spring G3	–.06	–.01	**.12**	–.01	–.05
Spring G4	–.01	–.06	–.02	**.10**	–.01
Spring G5	.00	–.03	–.04	–.02	**.10**

Largest intercorrelations in **bold**.

The pattern of intercorrelations seen above indicates that we are tapping into idiosyncratic knowledge, but not in a particularly reliable manner. This might be caused by the fact that if a student is struggling in an area, it is an area the teacher will emphasize (they have all of these achievement level estimates). It could also be due to the fact that the test used in this study has not been optimized for the

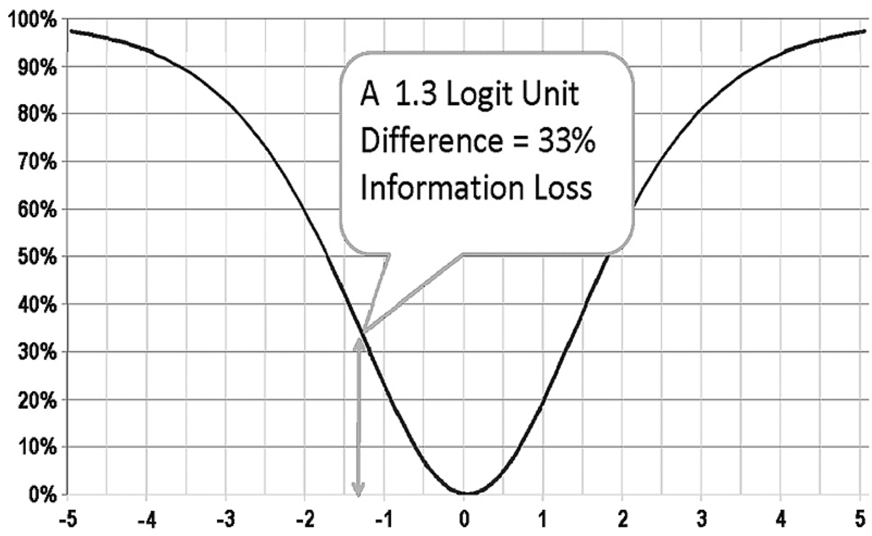

Figure 1. Information loss associated with administering items 1.3 theta units (Logits) away from true theta

identification of students with idiosyncratic knowledge patterns.

In observing the differences between observed goal area achievement level estimates and the overall achievement level estimate, the average difference for students identified as having idiosyncratic patterns was approximately 1.3 theta units. In a traditional adaptive test, this means the items in the surprising goal area are being mistargeted by approximately that amount. (Technically, all of the items in an adaptive test would be mistargeted for a student with an idiosyncratic knowledge pattern due to the adaptation. If a student had a higher performance level in one goal, the items chosen for that goal would be tend to be too easy for the student, while the other items would tend to be slightly too hard. The magnitude of the mistargeting would be dependent on the characteristics of the item pool, the number of items selected in each goal area, and the number of goals on the test.)

Figure 1 shows the information loss associated with mistargeting by 1.3 theta units in a test using the one-parameter logistic IRT model. The loss is approximately 33% for an infinitely deep item pool. Much of the process for ICAT is based on reducing this loss.

The ICAT Algorithm

To the extent that these idiosyncratic knowledge patterns are related to goal areas within the test design, we can describe an idiosyncratic knowledge CAT (or ICAT) process for identifying these patterns so that teachers may provide their students with appropriate instruction.

The ICAT algorithm that we will describe below assumes that the testing system has minimal information about the student when the test begins (the student's year in school, scores from previous tests that the student has taken, the test questions that the student has been administered in the past). As the test progresses, it moves through several phases (segments) in which the test's behavior changes, depending on the student's performance. The first segment acts much like the earliest adaptive tests, maximizing information around the provisional achievement level estimate. The second segment equalizes the information available about each goal area, finalizes the overall achievement level estimate, and identifies content areas in which the student seems to have idiosyncratic performance. The third segment concentrates on the content areas that seem to be idiosyncratic, identifying the achievement level estimates in each of the idiosyncratic goals with as much accuracy as possible.

To describe an adaptive test strategy, we need to consider the test design, the entry point, item selection, scoring, and test termination. (For those who are interested in the decisions made in the development of an adaptive achievement test, see Kingsbury and Houser, 1999). To build the ICAT procedure, we have used the MAP test design as a starting point, and added the characteristics needed to improve measurement of idiosyncratic knowledge patterns.

Many of the elements (such as the overall test design) won't change from a traditional adaptive test to ICAT. We have included a complete description of the adaptive testing process that was used to collect the example data, to allow the reader to see how the ICAT process may be integrated into an operational system. The procedure for item selection to identify idiosyncratic knowledge patterns may be described as follows, (*ICAT* denotes an element specific to ICAT):

Test Design Basic Assumptions

- The test will examine performance in a single subject area that has been shown to be essentially unidimensional in the population of interest
- The population of interest has been defined as students who have had an opportunity to learn the content included in the item pool
- The adaptive test will draw N items from an item pool of P items
- The test samples J goal areas, each represented by a unique and independent set of items
- During the test, an overall achievement level estimate and achievement level estimates for each individual goal will be calculated *(ICAT)*

Entry Point

- The initial estimate of performance for the student is defined by identifying the best information available about the student's achievement level, as follows:
  - The growth-adjusted estimate of student performance given a prior test score is used, if available;
  - The grade-level mean for the student is used next, if available;
  - A predefined overall estimate is used next, if no other information is available.
- The initial estimate will be used as the momentary achievement level estimate to select the first item for the test. The initial estimate may be adjusted downward slightly to allow a student to encounter personally easy questions at the beginning of the test.

Item Selection

Segment One (Items 1 to X)
- The first item is chosen to maximize information at the entry point, within the constraints of the item exposure process (no content-balancing process is used during Segment One).
- Items 2 through X are selected to maximize the informa-

tion available at the appropriate momentary achievement level estimate.
- As each item is evaluated, item information is calculated at the point on the measurement scale which represents the information-weighted average of achievement level estimates from the overall test and the goal the item represents (If the student has an achievement level estimate of –0.5 for the total test with test information of 6.0, and a goal area achievement level estimate of 1.0 with goal information of 2.0, the information-weighted estimate for the goal would be [–0.5*6.0 + 1.0*2.0]/[6.0 + 2.0], or –.125). *(ICAT)*
- *X* should be chosen to allow the achievement level estimate of θ to stabilize so that the standard error of measurement is approximately half of the standard deviation of achievement in the population of interest.

Segment Two (Items $X + 1$ to $X + Y$)
- For the next *Y* items, select the item that maximizes information in the goal area with the achievement level estimate with the lowest total information (This is done using the information-weighted achievement level estimates as described in Segment One).
- *Y* should be chosen to allow the standard error of measurement of each goal area achievement level estimate, θ_j, to be approximately half the standard deviation of achievement in the population of interest.
- At the end of Segment Two, perform a statistical test to identify whether the value of each θ_j estimate differs from the estimate of θ. *[ICAT]*
- At the end of Segment Two, finalize the estimate of θ. *[ICAT]*

Segment Three (Items $X + Y + 1$ to end)
- Select the next set of items to improve precision of measurement for only those goal areas identified as noticeably different from the overall achievement level estimate. *(ICAT)*
- During this portion of the test, item selection within any goal area is done using the estimate of θ_j for that goal area. *(ICAT)*
- During this Segment, the estimate of θ is not updated. *(ICAT)*

Scoring

- A final estimate of θ based on the information at the end of Segment Two is calculated and reported (this is the last point in the test at which content balancing was used).
- Estimates of θ_j for each goal area that is tested in Segment Three are calculated and reported.
- An estimate of θ that is based only on the items for goals not tested in Segment Three is calculated and reported. This score could be considered the student's overall score, but that term loses meaning in the face of idiosyncratic knowledge patterns. For a student who doesn't display an idiosyncratic pattern, a single score has a great deal of utility. For a student whose knowledge pattern is idiosyncratic, an overall score masks the character of the student's learning.
- Maximum-likelihood scoring is used for all reported scores.

Test Termination

- The test ends at the end of Segment Two if no goal has an achievement level estimate that differs from the achievement level estimate for the total test.
- During Segment Three, the test ends if the standard error of measurement for each of the goal areas still being tested falls below one-third of the standard deviation of the population.

An ICAT Example

Figure 2 shows an example of ICAT with the divergence of an achievement level estimate for a goal from the achievement level estimate for the total test. In this case, a test with three goal areas is given to a student. The first item is chosen from Goal Area 1 and administered. After the response, the achievement level estimate for the total test and the goal are exactly the same in the first column (206). The next two items are chosen from the other goals, and the overall achievement level estimate drifts down so that in the third column it reaches 179. Since no additional items have been chosen in Goal Area 1, it maintains its earlier estimate (206). Since the overall estimate has fallen and its standard error has been reduced, there is no longer overlap between the standard errors. If the testing process were active this early in the test, it would indicate the student as being idiosyncratically high in Goal Area 1. For the fourth item, the overall achievement level estimate and the Goal Area 1 achievement level estimate would be combined based on test information at the respective estimates. In a well-designed item pool, this will result in the selection of an item that is about one-quarter of the way from the overall estimate to the goal estimate (approximately 186). As the fourth item is chosen from Goal Area 1 and answered correctly, both achievement level estimates are updated. This continues as the test progresses through Segment 1 and Segment 2.

As can be seen by the figure, the unique characteristics of the student's responses move the achievement level estimates apart as the test progresses. If the student has consistent performance across the goal areas being tested, the ICAT would differ very little from a traditional adaptive test. If the student is displaying differential performance in the goal areas, the test extends to detail this idiosyncrasy. This approach is designed to allow the ICAT process to be

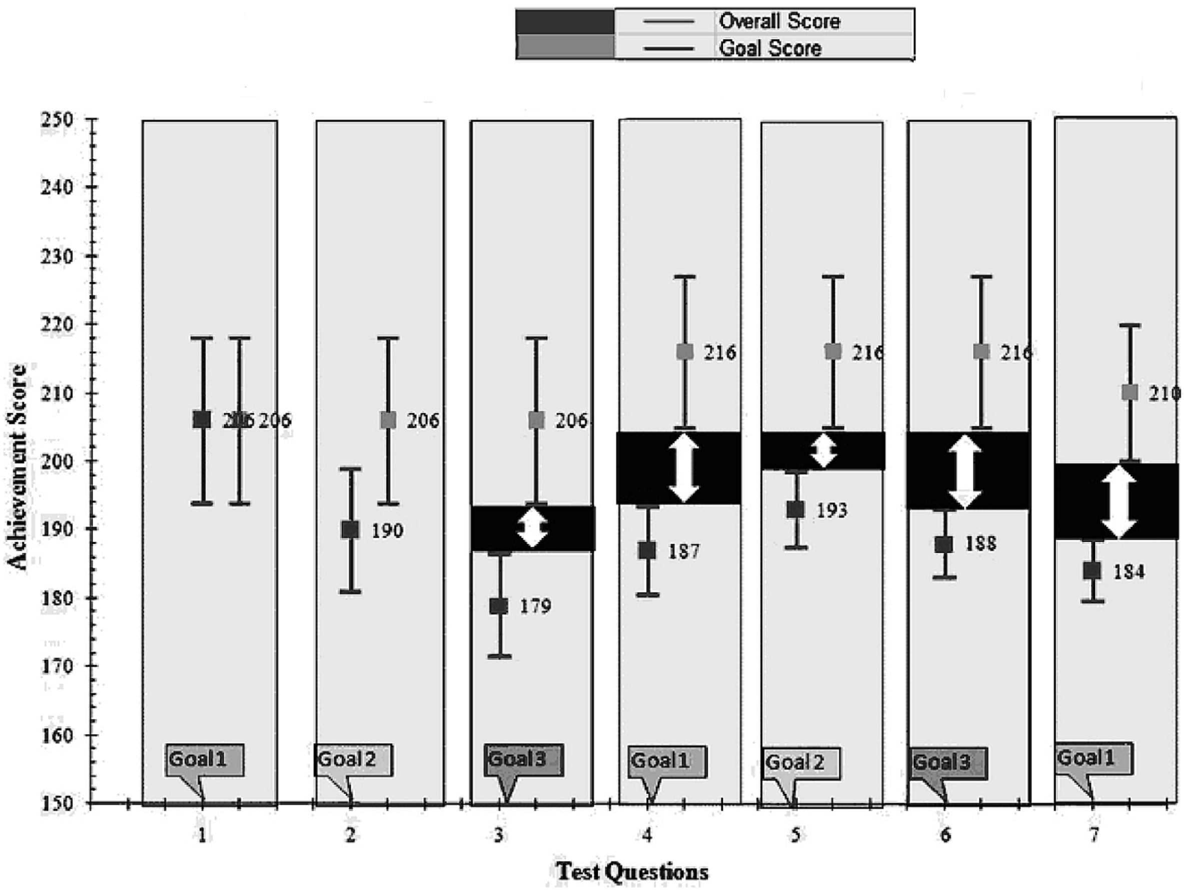

Figure 2. Overall achievement level estimate and achievement level estimate for a single goal for a student taking an ICAT test.

used in with a group of students without hurting the accuracy of the assessment for any student. While this is a design feature, the performance of the procedure in actual settings is an area of research that needs investigation.

ICAT was specifically designed for measuring student achievement in K-12 education, but it could be extended to use in almost any educational/training setting without substantial modification.

Discussion

A test can be designed to enhance the ways in which the test scores will be used. With fixed-form tests, there are limited options to enable the test to be used for different purposes. Within an adaptive test, on the other hand, we can build the purpose for the scores into the test design. This study has suggested a test design that allows us to collect more detailed information about a student's strengths and weaknesses.

Using a traditional adaptive test and an example data set, a large percentage of the students in our sample exhibited an idiosyncratic knowledge pattern beyond what we would expect due to chance. In addition, this pattern can be seen to sustain itself (weakly) across a span of 4 to 5 months. Even with a nonoptimized test design, we can capture information about patterns of knowledge. The ICAT test design should enable us to capture more information concerning each student's pattern of knowledge.

The example study here raises a number of psychometric research questions that need to be addressed prior to large-scale use of the ICAT system. After this psychometric research is nailed down, the questions may become more interesting. Live-study research with the ICAT should help us determine how persistent idiosyncratic knowledge patterns are within students. Classroom research can help us determine whether idiosyncratic patterns indicate misunderstandings, incomplete instruction, or styles of learning.

For students without an idiosyncratic knowledge pattern, ICAT will have the look, feel, and measurement characteristics of a traditional adaptive test. The information loss for these students should be quite small, since the goal achievement level estimates should hover around the total achievement level estimate. This is another research question that may be of interest to examine in simulation and live studies.

In almost every educational setting, the clarity and consistency of instruction is an important issue for the teacher and the students. The ICAT procedure should provide more information about each student's unique capabilities. Giving this information to teachers in a timely fashion should allow them to provide a better education for all of their students.

References

Bart, W. (2007, April). *What is a highly diagnostic test item? A dense item perspective.* Paper presented to the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Kingsbury, G.G. (2003, April). *A long-term study of the stability of item parameter estimates.* Paper presented to the annual meeting of the American Educational Research Association, Chicago, IL.

Kingsbury, G.G., & Houser, R.L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow, & J.B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 93–115). Mahwah, NJ: Erlbaum.

Kingsbury, G.G., & Zara, A.R. (1991). A comparison of procedures for content-sensitive item selection in computerized adaptive tests. *Applied Measurement in Education*, *4*, 241–261.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

NWEA (2003). *Technical manual for use with Measures of Academic Progress and Achievement Level Tests.* Portland, OR: Northwest Evaluation Association.

Rudner, L.M., & Talento-Miller, E. (2007, April). *Diagnostic testing using decision theory*. Paper presented at the annual conference of the National Council on Measurement in Education, Chicago, IL.

Samejima, F. (1977). The use of the information function in tailored testing. *Applied Psychological Measurement*, *1*, 233–247.

Tatsuoka, K.K., & Tatsuoka, M.M. (1997). Bug distribution and statistical pattern classification. *Psychometrika*, *52*, 193–206.

Weiss, D.J. (1976). Adaptive testing research at Minnesota: Overview, recent results, and future directions. In C.L. Clark (Ed.), *Proceedings of the First Conference on Computerized Adaptive Testing* (pp. 24–35). Washington DC: United States Civil Service Commission.

van der Linden, W., & Chang, H.-H. (2003). Implementing content constraints in α-stratified adaptive testing using a shadow test approach. *Applied Psychological Measurement*, *27*, 107–120.

G. Gage Kingsbury

Northwest Evaluation Association
5885 SW Meadows Road, Suite 200
Lake Oswego, OR 97035
USA
Tel +1 503 624-1951
Fax +1 503 639-7873
E-mail gage.kingsbury@nwea.org

