

Assessing literacy in the classroom: The BEAR assessment system for SLIC

Nathaniel Brown, Indiana University
Mark Wilson, University of California, Berkeley
Amy Dray, University of California, Berkeley
Yongsang Lee, University of California, Berkeley

Annual meeting of the American Educational Research Association
New York City
March 27, 2008

The SLIC Assessment System was developed using the Berkeley Evaluation & Assessment Research (BEAR) Center approach to embedded classroom assessment, known as the BEAR Assessment System (BAS). The BAS is an integrated approach to developing assessments that provide meaningful interpretations of student work relative to the cognitive and developmental goals of a curriculum (NRC, 2001). It consists of four building blocks: progress variables, an items design, an outcome space, and a measurement model (Wilson, 2005).

Progress Variables

Progress variables are the particular concepts and skills that form the core learning goals of the curriculum. They are organized according to the principle that assessments should reflect a developmental perspective of student learning in which deeper understandings are developed from, and take the place of, earlier understandings as students progress toward higher levels of sophistication and competence (Wilson, 2005).

In the SLIC Assessment System, progress variables were derived in part from the critical literacy skills identified in past research (McDonald & Thornley, 2002, 2005) and in part from empirical research on how students performed on pilot versions of the SLIC assessments. Importantly, they map onto the scope and sequence of the curriculum itself such that they should mirror the learning and development that is expected to take place in the classroom. The SLIC Progress Variables (see Figures 1 and 2 for an example) describe the development of literacy skills as progressing from a foundation of using surface features to navigate and predict the content of unfamiliar text, to a deeper understanding of content within and across texts, to an ability to infer meaning of unfamiliar vocabulary from context, to a critical appreciation of authorial intent and the choices authors make in crafting text. These variables were developed in partnership with the San Diego Unified School District and the curriculum developers.

Items Design

The items design is a framework for designing tasks to elicit specific kinds of evidence about student knowledge that can be interpreted in terms of the progress variables. This building block is grounded in the principle that assessment should be embedded in normal classroom activity and be based upon authentic instructional tasks.

In the SLIC Assessment System, assessment items were derived from the skills and strategies of proficient readers described in the progress variables and correspond to specific teaching points and instructional activities in the SLIC curriculum. Examples of such activities include anticipating the content of a text from the text's surface features, determining the main ideas of a paragraph, and inferring an author's implicit meaning. Each of these activities is closely linked with an item type, a template for a suite of assessment items associated with different texts that each measure the same skill or strategy. An example is shown in Figure 3.

At this stage in the project, eight full-length assessments have been developed by applying these item types to two types of texts—expository and persuasive—across 4 grade levels (7th through 10th). The narrative assessment is in development and eventually the assessment system will include a pre-test, post-test, and several intermediate benchmark assessments at grades seven, eight, nine, and ten (sixteen assessments; 4 per grade level). The goal is for teachers to monitor the progress of their students throughout the year and to allow the tracking of individual students across grade levels. These assessments will be the basis for benchmarks which will not only be helpful to teachers in their classrooms, but will also be an important part of the overall evaluation of the SLIC program.

Outcome Space

The outcome space describes the qualitatively different kinds of student response elicited by the items and maps these classes of response to the levels of the progress variables, operationalizing the principle that teachers should be the primary managers of assessment in the classroom. The outcome space is the evidentiary foundation for teachers to use on a daily basis, in both formal and informal instructional contexts, to judge the progress of their students along the progress variables and make informed decisions about future instruction.

In the SLIC Assessment System, the outcome space was informed by previous research (McDonald & Thornley, 2002, 2005) and further refined by analyzing student responses to pilot versions of the assessments. Each activity described above, such as determining the main ideas of a paragraph, is a complex skill requiring attention to and the coordination of multiple considerations. Rather than relying on single clues, proficient readers consider multiple aspects of the text and cross-check their meaning; we call this applying and cross-checking multiple *tactics* in the service of a particular skill or strategy. The SLIC Outcome Space (see Figure 4) is a general rubric that describes more and less proficient application of tactics. This outcome space is applied to each item, resulting in a collection of item-specific scoring guides that share a common framework.

In 7th grade, for example, students were required to read a persuasive text extolling the benefits of exercise and describing an exercise program. Items were developed to assess student understanding of the progress goals. For example, in order to fully understand the main idea of the paragraph, the students had to be able to pull together information from several different sentences within the paragraph and discard those sentences which were irrelevant to the main idea. These tactics had been identified through earlier investigations, and the progress goals were

developed from those tactics. Once pilot data was collected, the next step of the process was to go through student responses and decide qualitatively when students reached a particular level of progress. The process was repeated across items, instruments, and grades.

Measurement Model

The measurement model provides estimates of person proficiency and item difficulty calibrated onto an interval scale using a multidimensional item response model (Adams, Wilson, & Wang, 1997), allowing us to use quantitative approaches to establishing sound standards of validity and reliability, as well as the qualitative data that the other three building blocks have generated. The creation of a calibrated, quantitative version of the progress variables allows teachers to track and map the progress of individual students and groups as they undergo instruction; this is facilitated with an integrated suite of technological tools (Kennedy, Wilson, & Draney, 2005) that allow the teacher to focus on her crucial roles as evaluator and interpreter of student work.

In the SLIC Assessment System, the measurement model allows us to both quantitatively measure the progress of students across all four years of the study along a common interval scale and to interpret those numerical estimates of proficiency in terms of the qualitative, meaningful levels of the progress variables. Current work is establishing the validity and reliability of this scale.

Figure 5 presents a visual display of the results of the 9th grade pretest and helps to unpack the idea of the measurement model. First, we show 10 of the items on the instrument along the bottom; each item response is scored along four scales: (i) disregards tactics, (ii) using a single tactic, (iii) uses multiple tactics, and (iv) thoroughly uses tactics. These scales are our progress variables, which are lined up on the left. They reflect the expectation that over time

students will move from less sophisticated to more sophisticated literacy techniques. In this case, we expect students to move from disregarding tactics entirely to being thorough in their use of tactics over the duration of the curriculum.

The students' ability levels are grouped together in an on-its-side histogram on the left of the Figure. As you can see, and as expected, the majority of the 9th grade students at pretest clump within the range of disregarding tactics. Some students do start at the point of using at least one tactic. Over time, the goal is to shift students higher on the vertical axis so that they are more likely to use the tactics emphasized by the curriculum. We expect the classroom teachers to be able to interpret this graph fairly easily and we expect these findings to inform instruction. Furthermore, we can easily map change over time as the student progresses through the curriculum and across grades. These findings will also inform the evaluation piece of the project.

Conclusion

The BEAR assessment system starts from the position that the assessment instrument is always secondary—there is always a purpose for which the instrument is needed and the context in which it is being used (Wilson, 2005). It entails defining a construct—a theoretical or practical object of interest—prior to developing any item or any method of scoring responses.

Furthermore, the process is collaborative. The development of the assessment is designed with the goals of the curriculum and teachers in mind.

For this project, assessments were developed, piloted, and designed in partnership with school district leaders and educational researchers and refined based on feedback from teachers and coaches. While effective, the process was not always easy. The BEAR system requires the identification of progress goals for the curriculum; sometimes these progress goals are more obvious to the practitioners on the team than to researchers. Pinning down “when progress has

been made” or recognizing when a student has reached particular level of ability (i.e. “using multiple tactics”) is sometimes easier said than done. The project has gone through several iterative stages of defining progress goals, creating items, and specifying what a specific response in the outcome space means substantively and how it can be reliably scored by teachers. For example, rater effects were discovered across teachers—in the pilot test, teachers had very different ways of scoring particular student responses. This has led us to re-think our progress goals, outcome spaces, and the best way to score the responses. Ultimately, however, we are confident that this will lead to an assessment system for the SLIC program that will serve the purposes of practitioners and researchers alike.

References

- Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, 21, 1-23.
- Belzer, A. (2005). Improving professional development systems: Recommendations from the Pennsylvania adult basic and literacy education professional development system evaluation [Electronic Version]. *Adult Basic Education*, 15(1), 33-56.
- Kennedy, C. A., Wilson, M., & Draney, K. (2005). GradeMap 4.1 (computer program). Berkeley, CA: University of California-Berkeley, BEAR Center.
- McDonald, T., & Thornley, C. (2002). Unlocking meaning in text: some thoughts on the literacy challenges our students face. *English in Aotearoa*(48), 54-60.
- McDonald, T., & Thornley, C. (2005). Literacy teaching and learning during the secondary years: Establishing a pathway for success to NCEA and beyond. *set: Research Information for Teachers*, 2, 9-14.
- Mizell, H. (2003). Facilitator: 10 refreshments: 8 evaluation: 0 [Electronic Version]. *Journal of Staff Development*, 24(4).
- National Research Council (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.
- Poulson, L., & Avramidis, E. (2003). Pathways and possibilities in professional development: case studies of effective teachers of literacy. *British Educational Research Journal*, 29(543-569).
- Thornley, C., & McDonald, T. (2002). Reading across the curriculum: Secondary students talk about themselves as readers. *set: Research Information for Teachers*(1), 19-24.
- Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.
- Wilson, M. & Scalise, K. (2003). Reporting progress to parents and others: Beyond grades. In J. M. Atkin & J. E. Coffey (Eds.), *Everyday assessment in the science classroom*. NSTA Press: Arlington, VA.
- Wilson, M. & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education*, 13, 181-208.

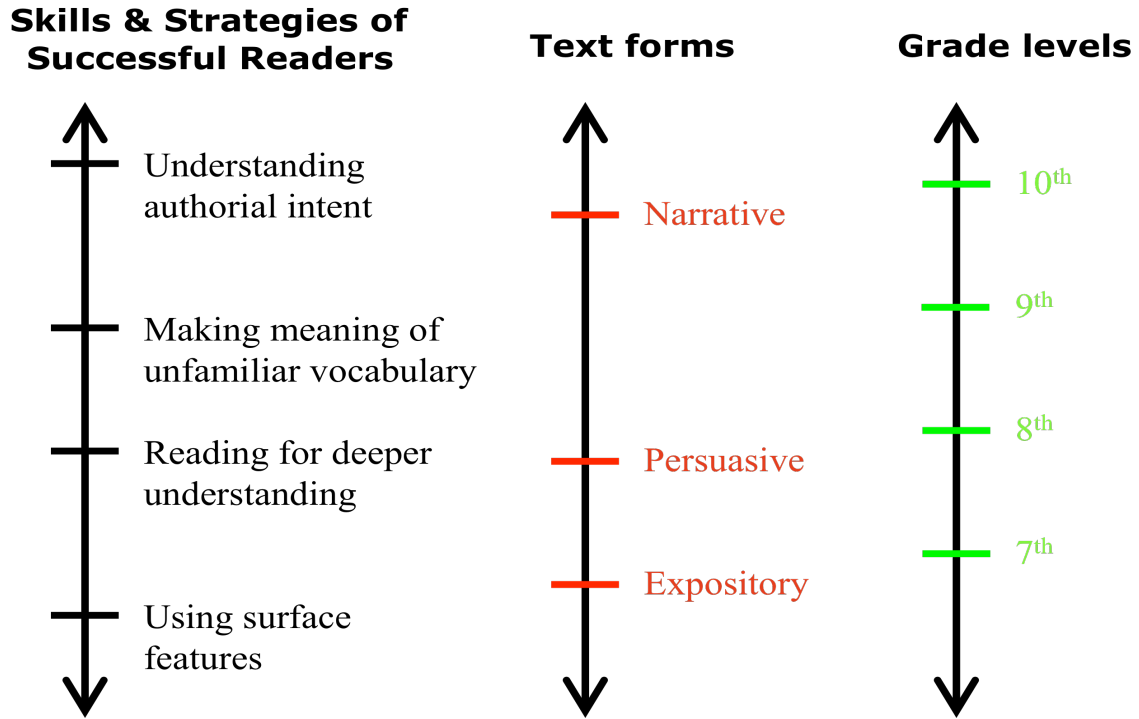


Figure 1. First SLIC building block: The Construct Map. The three scales represent three possible ways to map students' progress through the curriculum.

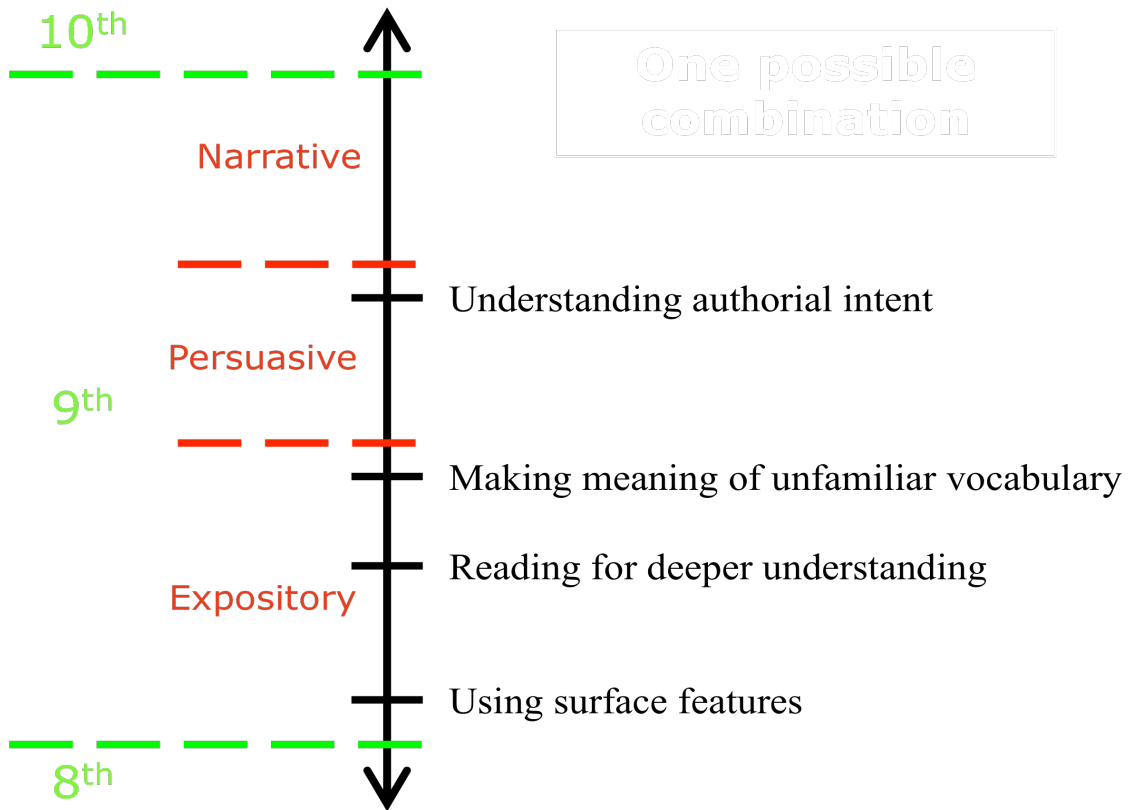


Figure 2. One possible example of how to combine the three scales in Figure 1 into a single Construct Map.

SLIC Skill or Strategy	Item Type
<p>Anticipate content from text features</p> <p>TACTICS:</p> <ol style="list-style-type: none"> 1. Read the title for a general overview 2. Read summary text, if present (e.g., sub-titles, “main idea”) 3. Read section headings to learn the topics that will be covered 4. Scan maps, charts, tables, illustrations, captions, etc. 	<p>Scan the title, [sub-titles, headings, sub-headings, illustrations and captions]. What do you think this text will be about?</p> <p>How did you figure out what this text is about?</p>
<p>Make meaning of unfamiliar vocabulary</p> <p>TACTICS:</p> <ol style="list-style-type: none"> 1. Consider local context (e.g., sentence-level, paragraph-level) 2. Consider text-level context 3. Consider word morphology 4. Consider prior knowledge 	<p>Reread paragraph [number] on page [number] that begins with, “[text]...” What does [word/phrase] mean?</p> <p>How did you figure out the meaning of this word?</p>
<p>...and ~10 others</p>	

Figure 3. Second SLIC building block: The Items Design.

T Thorough	Response has enough cross-checked, tactic-based elements to constitute a complete answer .
C Cross-Checked	Response has at least two tactic-based elements that have been cross-checked
M Multiple Tactics	Response has at least two tactic-based elements that have not been cross-checked
S Single Tactic	Response has one tactic-based element
D Disregards Tactics	Response has no tactic-based elements or is blank

Figure 4. Third SLIC building block: The Outcome Space.

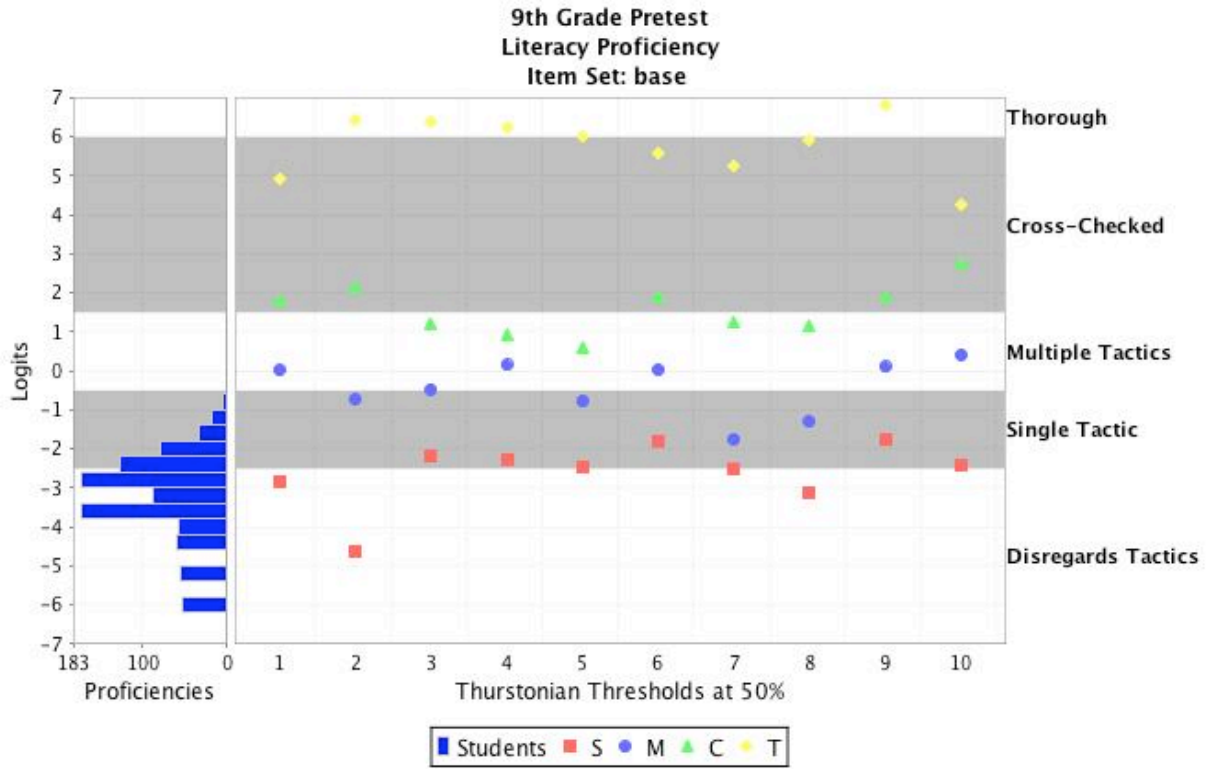


Figure 5: Output from the fourth SLIC building block: The Measurement Model.