Using technology to advance understanding:

The calibration of CLAS, an electronic developmental scoring system

Presented at the NERA conference in Trumbull CT,
October 20, 2017

Theo L. Dawson, *Lectica, Inc.*

Overview

- The development and calibration of CLAS (Computerized Lectical Assessment System)—a 20 year journey
  - About CLAS
  - The dimension—hierarchical complexity (Lectical Level)—to which CLAS is calibrated
  - The human scoring systems—Lectical Assessment System, Rubric Scoring System
  - The ongoing research behind CLAS—Lexication—which produces the curated taxonomy of meanings (the Lectical Dictionary) used to create CLAS algorithms
  - Calibration results for the LRJA (Lectical Reflective Judgment Assessment)

This paper describes the development and calibration of CLAS, an automated developmental scoring system that measures the growth of understanding. CLAS's algorithms leverage a curated taxonomy of meanings called the Lectical Dictionary—developed through an analyst-machine-learner collaboration—to determine how test takers appear to have constructed meanings over time.

About CLAS

- CLAS differs from other solutions to automated scoring in several ways:
  - It is grounded in a strong theory of learning and development.
  - Its scores are diagnostic and rich in meaning, in that they are tied to richly described evidence-based learning sequences.
  - It can readily be calibrated to track the development of understanding in a wide range of subject areas.
  - It is the core of a platform for delivering many subject-specific assessments designed to support the development of deep understanding and skill.

Context

- The creation of accurate and reliable electronic scoring systems for texts has been called the "holy grail" of educational assessment (Whittington & Hunt, 1999).
- Most are created through the computational analysis of essays that have either been scored by humans with a variety of conventional rubrics or through the analysis of “graded” texts—texts that have been determined through expert opinion to be suitable for specific school grades (Zupanc & Bosnic, 2015).
- They focus primarily on aspects of texts that lend themselves to computational analysis —such as sentence length, syntax, morphology, punctuation, size of vocabulary, word length, associations between words, and the presence of targeted vocabulary.
- Some of their developers claim to measure constructs like meaning (Pearson Education, 2010; Valenti, Neri, & Cucchiarelli, 2003), coherence (Shermisand & Hamner, 2013), development (Smith, 2009), and effective writing (Rich, Schneider, & D’Brot, 2013)
- Measurement claims are difficult to evaluate due to a lack of clarity about the meaning of these terms within the context of measurement systems and the paucity of published evidence.

The instrument:

Lectical Reflective Judgment Assessment (LRJA)

We have developed a number of instruments that are scored with the Lectical Assessment System. One of these is the LRJA.

The LRJA (form 1)

- The LRJA examines students' reasoning about **inquiry** and **evidence**, the **quality of information and evidence**, and the **nature of knowledge**. It includes questions like the following:
  1. Some scientists think that violent TV shows are bad for children. Others think some violent TV shows are okay. Which group of scientists do you think is right? Please explain.
  2. How would you decide which group of scientists was right? Please explain.
  3. If you were one of the scientists who thought violent TV was bad for children, what could you do to convince the other group of scientists that you were right? Please explain.
  4. How is it possible that the two groups of scientists have such different ideas? Please explain.
  5. Is it possible to know for sure if violent television is bad for children? Please explain.

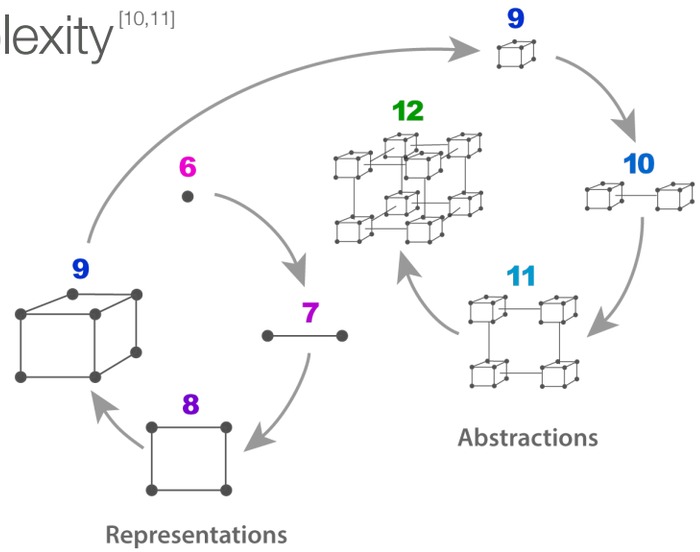
http://devtestservice.org/pdf/2012_1015_NERA.pdf

The developmental dimension:

Hierarchical complexity

We have taken a different approach. We sought to measure a theoretically derived and and empirically grounded dimension called hierarchical complexity.

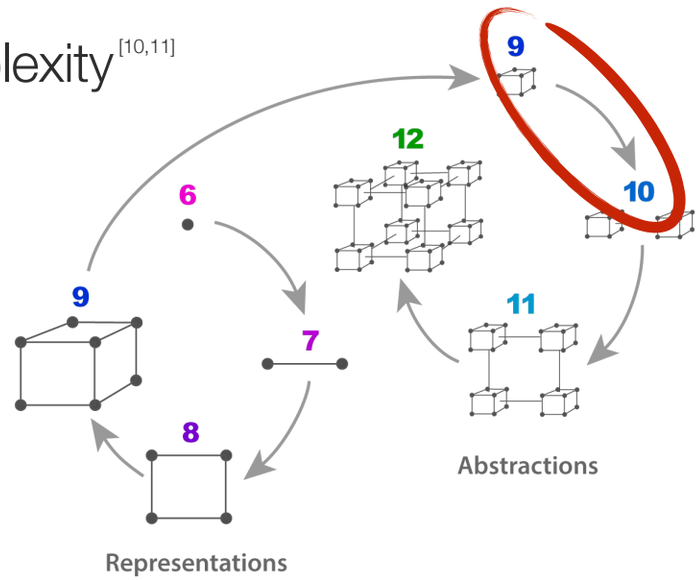
Hierarchical Complexity[10,11]



In the cognitive-developmental tradition, learning is viewed as the construction of increasingly sophisticated understandings—of the physical and social world, and of ourselves. Each new level integrates and builds upon the knowledge of the preceding level, resulting in levels of increasing hierarchical complexity.

Hierarchical Complexity[10,11]

- There are 13 levels—0-12.
- Levels 9 and 10 are most common between grades 4 and 12.



The evolution of our scoring systems

1 Score 1,325 assessments with the Lectical Assessment System (LAS)*

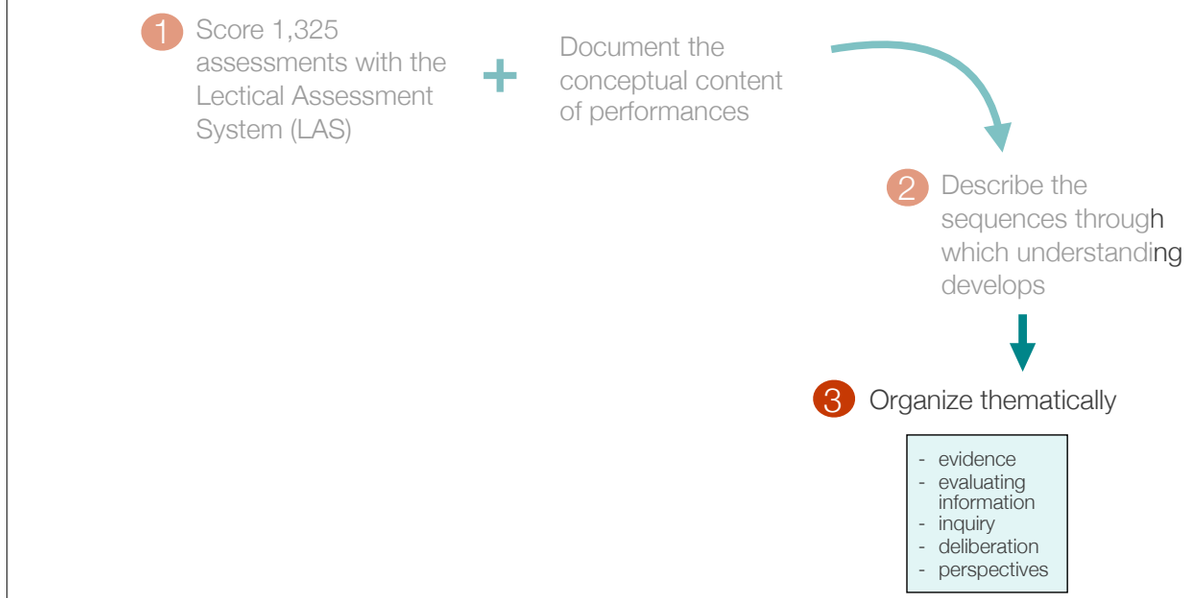
\+

Document the conceptual content of performances

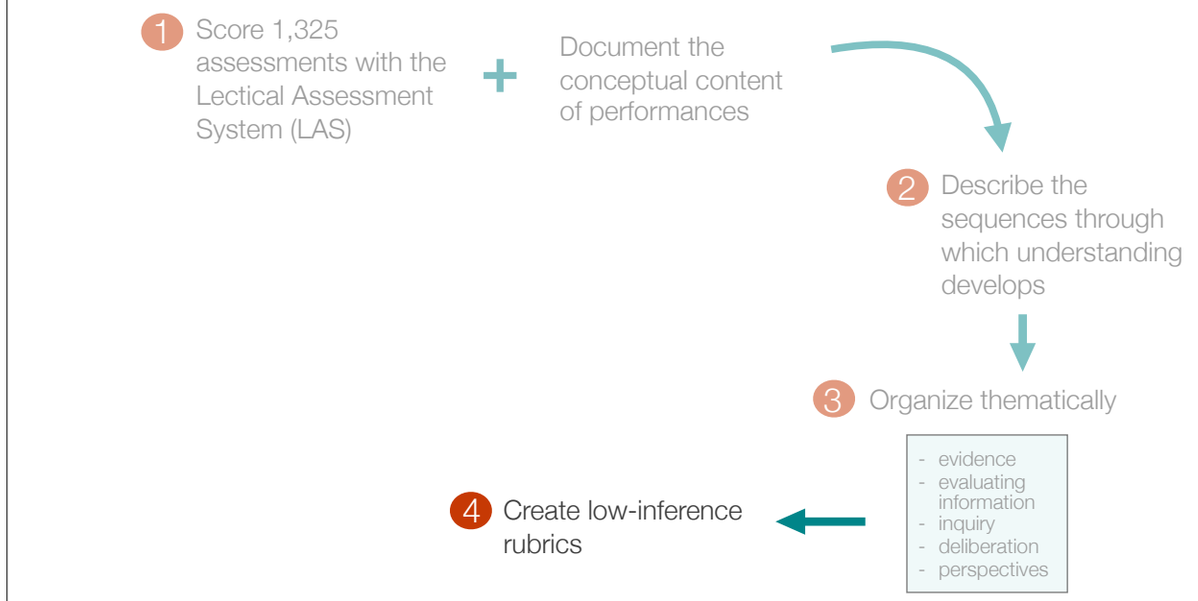
*Later, these were also scored with rubrics.

Learning to score with the LAS takes several years—not practical for large projects. To make it possible to score large numbers of LRJAs, we created a rubric scoring system. First, Certified Lectical Analysts scored 1,325 LRJA performances representing roughly equal numbers of students who were in or had completed 4th through 8th grade, with smaller Ns for grades 9-11. As analysts scored, they documented the conceptual content of the assessment performances. These performances were later also scored with rubrics.

This resulted in a developmentally curated database of “meanings” which was used to conduct a rational reconstruction of dozens of sequences through which particular meanings developed.



These sequences were then organized by theme and construct.



And converted into low-inference scoring rubrics.

Low-inference rubrics

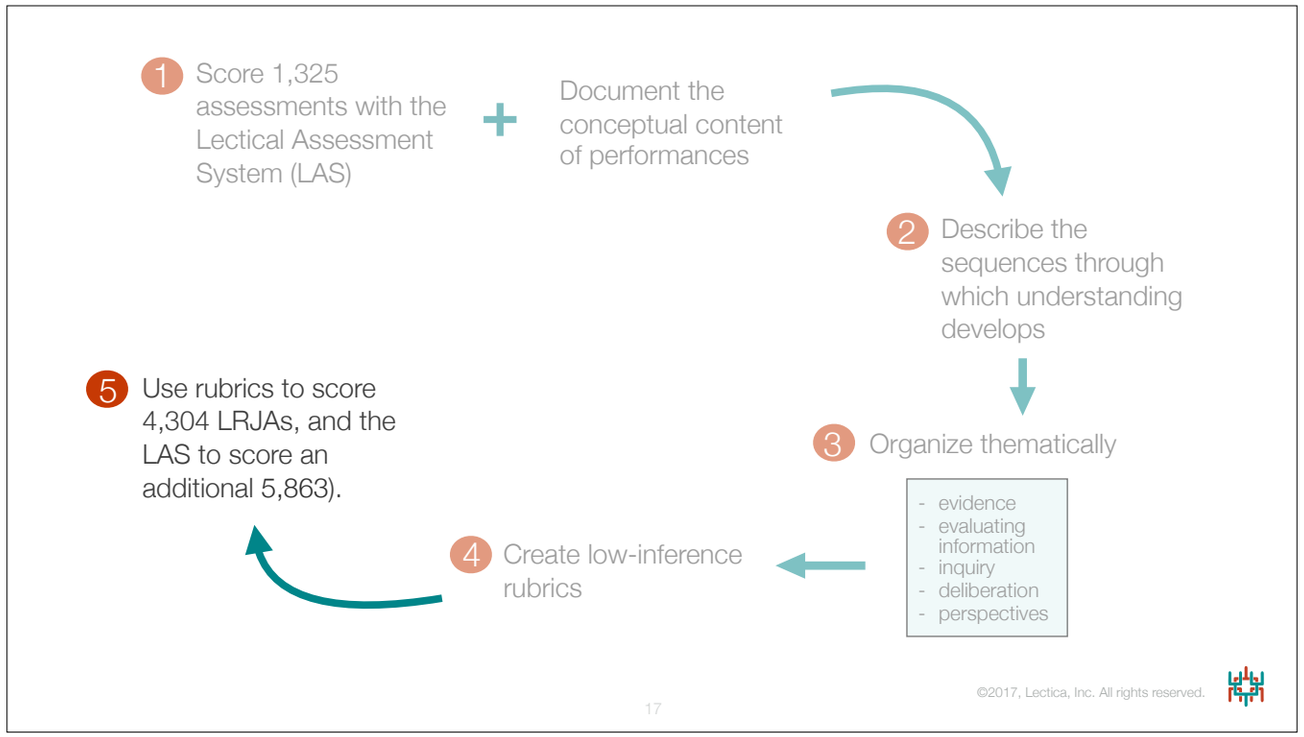
- Rubrics presented a series of specific ways in which students might understand a particular construct.
- Each selection corresponded to a Lectical *phase.*
- In the LRJA, rubrics focused on aspects of inquiry, evidence, truth, knowledge, reasoning, and conflict.
- There were 5-8 clusters of thematically related rubrics per question (1–3 were generally used).
- Each question received an average score based on rubric selections.

Question 1

Some scientists think that violent TV shows are bad for children. Others think some violent TV shows are okay. Which group of scientists do you think is right? Why?

"Both groups are correct because the kids that get too much violent TV might be more violent (by influence) than those who don't get as much violent TV. Being more violent is rather dangerous for the community, and the person him/herself."

Phase	In order to make a decision, you should…
8d	think (think hard) or ask parents (teachers, etc.).
9a	think about what you have learned from your parents (or in school), or think about what has happened to you (or someone you know).
9b	think about your own opinion (or what you like, think, believe, have seen) or the opinions of others.
9c	use your thinking skills or trust your thoughts, find reasons, think about why people think or feel the way they do, or think about the facts.
9d	think about what makes sense, possible outcomes (consequences), or what you already know from life experience.
10a	analyze the evidence you have collected, understand the reasoning behind a claim, use common sense or logic, or understand (get) both perspectives.
10b	make an educated guess or a rough estimate; think about your own values; put yourself in the other person's shoes, compare perspectives or evidence, or look for relationships.
10c	look for conflicting evidence, consider similarities and differences or the pros and cons of each position, or avoid personal bias.
10d	try to remain impartial or objective, consider multiple factors (as causes), weigh the results, or look at the big picture.



Then, trained coders rubric-scored 4,304 LRJAs and the Certified analysts LAS scored an additional 5,863 assessments. This process took 7 years. The LRJA rubrics were not, in fact, more efficient than scoring with the LAS. But building them allowed us to learn a great deal about conceptual development on targeted LRJA constructs, and led us to the development of CLAS, supports for scoring with the LAS, and a much more efficient method for developing learning sequences.

Reliability & validity

Lectical Assessment System

- Measures the same underlying dimension as several other longitudinally validated developmental scoring systems (https://youtu.be/W0PDUK3tcN0).
- Rasch person separation reliability ranges from .91 to .97.*
- Human raters are required to maintain an agreement rate of 85% within .20 of a level (continuously monitored).

LRJA Rubric Scoring System

- Based on research into learning sequences conducted with the Lectical Assessment system
- For a sample of 3,754 rubric-scored LRJAs, Rasch person separation reliability = .91
- Human raters are required to maintain an agreement rate of 85% within .20 of a level (continuously monitored).

*https://lecticalive.org/about/hierarchical-complexity

To do so, we first spent several years developing an accurate and well-validated human scoring system called the Lectical Assessment System, which is used to assign scores on the Lectical Scale (https://lecticalive.org/about/skill-levels), and a rubric scoring system for the LRJA based on research conducted with the Lectical Assessment System and a set of methods called developmental maieutics (https://lecticalive.org/about/developmental-maieutics).

Automating the Lectical Assessment System

Lectical Dictionary

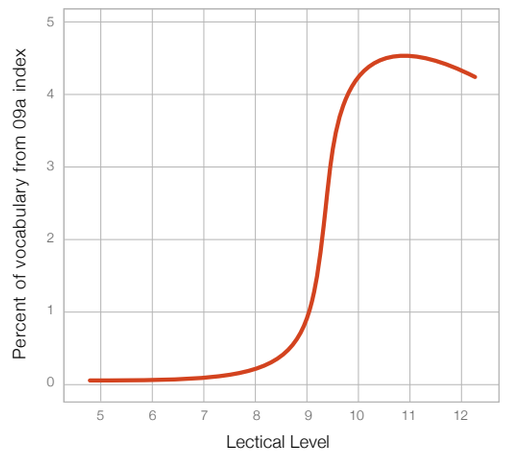
- In cognitive developmental theory (Piagetian tradition), each new level builds upon the previous level.
- New meanings build upon and transform earlier meanings.
- This development is represented in the way we put together words to express meanings.
- We are building a curated taxonomy of meanings through a process called **lexication**.
- Words and short phrases—Lectical Items—are assigned to phases (1/4 of a level) based on the phase at which their simplest possible meaning is likely to become useful.
- Over 180,000 Lectical Items are in the Dictionary.

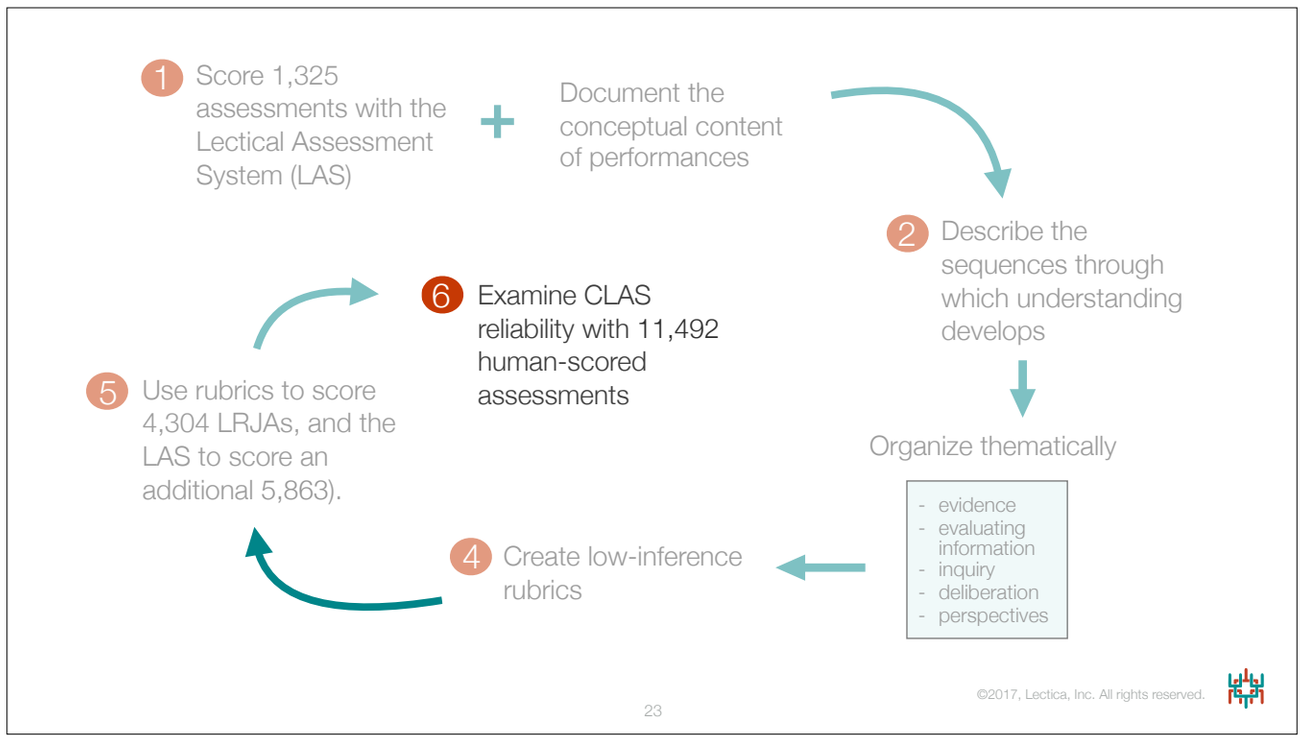
Example—evidence theme

Phase	Lectical items	Description
09b	evidence, fact, data, show evidence, have evidence, proof	Information you can show to prove something
09c	good evidence, more evidence, better evidence, accurate evidence	More is better than less, and it can be better or worse
09d	real, hard, solid, or proper evidence, explain evidence, collect, find evidence	The kind of information you must find and explain to get people to agree with you
10a	kind of evidence, source of evidence, evidence behind, compare evidence	Information that can be used to support arguments if it is the right kind or comes from a good source; can be compared to help you make a decision
10b	reliable evidence, scientific evidence, quality of evidence, unbiased evidence, impartial evidence, evaluate evidence	Information that is more or less reliable, valid, or biased; should be evaluated before you use it to make a decision

CLAS

- CLAS examines the relative distribution of Lectical Items across phases within individual performances
- And matches them to predicted distributions in a discriminant function based algorithm.





Split half reliability analysis

- Sample characteristics
  - Grades 4–12
  - Economically and ethnically diverse
  - Over 50 schools in US and Canada
- The construction and test samples were selected with the random number function in Excel.
- Construction sample
  - 5761 human scored LRJAs
- Test sample
  - 5731 human scored LRJAs

Results of split half discriminant analysis

- Construction sample
  - Human scores agreed with CLAS scores 85% of the time within .20 of a level.
- Test sample
  - Human scores agreed with CLAS scores 85% of the time within .20 of a level.

Discussion

- There have been many calls for tests that can evaluate deep understanding (as opposed to correctness). CLAS essentially captures the development of understanding/meaning as expressed directly in students' written performances.
- Reliability of LRJA 4–12 CLAS scores is adequate for research and low stakes, formative use of assessments. (High stakes use is justified if growth trajectory is what is observed.)
- CLAS produces confidence scores that identify many of the problems that lead to inaccurate scores. These do not yet catch all problematic performances.
- The Lectical Dictionary is more than the basis for a scoring system. It will become an authoritative taxonomy of the development of meaning/ understanding that can be used in a wide variety of contexts.

Notes

- We are on the verge of automating the creation of many Lectical Items, based on patterns we have observed in existing data. Within the next couple of years, we think CLAS will be running the show, with human vetting.
- CLAS is a general scoring system. Right now it is being used to score 8 different assessments, most of which are adult assessments focused on ethics, pedagogy, leadership, self understanding, and decision-making. K-12 assessments for science and social studies are under development.
- CLAS does more than produce Lectical Scores. It produces a great deal of detailed information about performances that is used to populate reports and advance research.
- We have used the Lectical Dictionary to create a developmental spell checker, and ask students to use it to check their spelling if more than 3% of words are misspelled.

Acknowledgements

The work reported in this paper/presentation would not have been possible without the contributions of the entire Catalyzing Comprehension for Discussion and Debate research team, the collaborating districts and school personnel, and the willingness of teachers and students to participate in assessments, classroom observations and recordings, and other data collection procedures. *The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant* R305F100026 *to the Strategic Educational Research Partnership Institute as part of the Reading for Understanding Research Initiative. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.*