Engagement, Alignment, and Rigor as Vital Signs of High-Quality Instruction: A Classroom Visit Protocol for Instructional Improvement and Research

Diane M. Early, Ronald D. Rogge, Edward L. Deci

The High School Journal, Volume 97, Number 4, Summer 2014, pp. 219-239 (Article)

Published by The University of North Carolina Press
DOI: 10.1353/hsj.2014.0008

Engagement, Alignment, and Rigor as Vital Signs of High-Quality Instruction: A Classroom Visit Protocol for Instructional Improvement and Research

Diane M. Early
University of North Carolina at Chapel Hill
diane_early@unc.edu

Ronald D. Rogge
University of Rochester
rogge@psych.rochester.edu

Edward L. Deci
University of Rochester
deci@psych.rochester.edu

This paper investigates engagement (E), alignment (A), and rigor (R) as vital signs of high-quality teacher instruction as measured by the EAR Classroom Visit Protocol, designed by the Institute for Research and Reform in Education (IRRE). Findings indicated that both school leaders and outside raters could learn to score the protocol with adequate reliability. Using observations of 33 English language arts (ELA) teachers and 25 mathematics teachers from four high schools, findings indicated that engagement, alignment, and rigor were all predictive of math and ELA standardized achievement test scores when controlling for the previous year's scores, although some of the associations were marginal. Students' self-reports of their engagement in school were also generally predictive of test scores in models that included perceived academic competence and observed engagement, alignment, or rigor. We discuss the importance of classroom engagement, alignment, and rigor as markers of instructional quality and the utility of the EAR Protocol as a means of assessing instructional quality.

Keywords: instructional quality, engagement, alignment, rigor, high school, standardized test scores

No Child Left Behind legislation enacted in 2002 both mandated standardized achievement tests in all states seeking federal funds and required schools and school districts to improve student test scores over time as a means of improving educational outcomes for all students (Rothman, 2012). Subsequently, Race to the Top has added to the press for improved test scores by increasingly holding individual teachers accountable for improving the scores of students in their classes (Klein, 2012). The National Research Council and the Institute of Medicine (2004) argued that the quality of teachers' instruction is the most proximal and powerful predictor of students' learning. Accordingly, considerable interest has been directed toward methods of assessing the quality of instruction in our schools as a way to monitor instruction and highlight the types of supports needed. In this article we describe one such tool for assessing instruction that provides immediate, understandable data to researchers, school leaders, and technical-assistance providers. We tested the extent to which different types of trained observers can achieve adequate inter-rater reliability, and we tested the tool's ability to predict student test scores, when a continuous scoring system is applied.

Vital Signs of Instructional Quality

The current study investigates the psychometric properties of the Engagement, Alignment, and Rigor (EAR) Classroom Visit Protocol, a tool designed by the Institute for Research and Reform in Education (IRRE) for measuring "vital signs" of instructional quality. Similar to a physician measuring blood pressure and pulse to obtain a quick picture of a person's health, this protocol was designed to identify vital signs of instructional quality that could be measured quickly and often as a way of tracking variation in the quality of instruction. Improvement in test scores is an important long-term outcome of high-quality instruction, but schools need more immediate feedback to gauge whether their efforts to improve instruction are working. The EAR Classroom Visit Protocol was intended for use by school staff, as well as outside change agents, consultants, and researchers, to meet these more immediate needs.

The Current Research

The EAR Classroom Visit Protocol was developed in 2004, and IRRE began field testing it immediately. It has been used in more than 100 elementary, middle, and high schools across the country for more than 27,000 visits (Broom, 2012). Those data, and feedback from schools that use the tool, provide preliminary indication of its utility, but it has not been used in a research study. The current study was conducted by an independent research team and aimed to (1) describe the EAR Protocol, (2) establish a system for creating continuous scores from EAR Protocol data, (3) investigate the tool's inter-rater reliability, when used by trained, external observers or by trained school and district personnel, and (4) examine the tool's ability to predict standardized test scores when using the continuous scoring system, both by itself and in conjunction with students' self-reported engagement in school and perceived academic competence.

Classroom Observational Tools for Measuring Instruction

Currently, there are a few tools available that have been found to be reliable and predictive of achievement. The Bill & Melinda Gates Foundation (2012) recently published an investigation of five such observational tools, focused on children in grades 4 through 8, namely: (1) Framework for Teaching (FFT; Danielson), (2) Classroom Assessment Scoring System (CLASS; Pianta, La Paro, & Hamre), (3) Protocol for Language Arts Teaching Observations (PLATO; Grossman), (4) Mathematical Quality of Instruction (MQI; Hill), and (5) UTeach Teacher Quality Tool (UTOP; Marder & Walkington). The five tools vary considerably in terms of their approaches and foci. For example, FFT emphasizes effective questioning in the classroom as a way of promoting intellectual engagement, and CLASS focuses on interactions between teachers and students as the basis for student learning. PLATO focuses on classroom practices in language arts. MQI focuses on math teaching knowledge and values teacher accuracy. UTOP focuses on math, science, and computer teaching and values a variety of modes of instruction, from inquiry based to direct. The Gates Foundation study concluded each of the five tools was a predictor of high-stakes state achievement test scores in math and English language arts (ELA), as well as an open-ended assessment of literacy, and a test of conceptual understanding of math.

The EAR Protocol expands the list of useful tools because it has several desirable characteristics and a somewhat different focus that may be preferable for some schools or districts. For example, the EAR Protocol can be used in all subject areas, including core subjects such as math and ELA and electives such as physical education and art. Its focus is a set of specific instruction-related experiences for students that result from what teachers do, rather than a set of teacher behaviors. Engagement (E), alignment (A), and rigor (R) are postulated to be vital signs of quality that can be improved through multiple types of professional development, but the tool does not focus on implementing specific teaching strategies. The EAR Protocol is appropriate for all school levels, including high school, whereas the Gates Foundation study focused solely on grades 4 through 8. Further, if used widely in a school or district, the protocol could provide a common language and set of descriptors to promote discussions about high-quality instruction across grades and subjects.

The EAR Protocol requires a 20-minute observation, providing enough time to obtain a clear picture of what is happening in the classroom while still being feasible for school administrators to use on a regular basis. Multiple observations of a single teacher, grade, or department are necessary for the results to be meaningful, but having this short observation period makes it usable for administrators who generally have very full schedules. The 20-minute observation stands in contrast both to the more extensive observation required by in-depth teacher evaluation tools such as the FFT and to the three- to five-minute "walk-throughs" that are popular with school personnel. Although those very brief visits may help administrators gain a picture of the general state of instruction, they are very subjective and are too brief to provide a meaningful understanding of what is taking place in an individual teacher's classroom (Downey, Steffy, English, Frase, & Poston, 2004; Protheroe, 2009). The 20-minute EAR Protocol allows for a richer sampling and a more quantitative representation of instructional quality.

Further, the EAR Protocol data are collected on a smartphone or tablet computer and uploaded immediately via Wi-Fi or docking station to a secure server. Authorized users can generate reports and graphs from an online system that aggregates observations across an individual teacher, department, grade, small learning community, school, or district. This system provides administrators with immediate feedback to quickly identify trends and changes. The information can be used for professional development and reflective conversations, as well as performance management and support at the district and school levels.

The EAR Classroom Visit Protocol

Engagement

Engagement—the first vital sign measured by the EAR Protocol—is a prerequisite for school success. It is manifested as effort and persistence and allows students to profit from challenging curricula (National Research Council and the Institute of Medicine, 2004), but an agreed upon way to measure it has been lacking. Many studies published in the past 40 years have confirmed that students who are high in intrinsic motivation are more engaged in learning that is deeper and more conceptual (e.g., Benware & Deci, 1984; Grolnick & Ryan, 1987) and perform better on heuristic, as opposed to algorithmic, tasks (Cerasoli, Nicklin, & Ford, 2014; McGraw, 1978). There is also evidence that when students have fully internalized the regulation for learning, they tend to be more engaged in learning and to perform better than when learning is controlled by external contingencies (e.g., Black & Deci, 2000; Grolnick & Ryan, 1989). Thus, intrinsic motivation and fully internalized motivation predict engagement and positive educational outcomes; together they are referred to as autonomous motivation for learning (Ryan & Deci, 2000). Importantly, research has shown that when teachers are supportive of students, interested in the material, and energized by teaching, their students are more autonomously motivated and engaged (Deci, Schwartz, Sheinman, & Ryan, 1981; Patrick, 1995).

Fredricks (2011) described three types of school engagement: behavioral, emotional, and cognitive. She pointed out that these three types of engagement are typically positively correlated, with most students reporting that they are high or low on all three. Rather than attempting to distinguish between these three types of intercorrelated engagement, the EAR Protocol focuses on measuring the observable aspects of all three types of school engagement. In the EAR Protocol, engagement is defined as students being actively involved during class. When students are engaged, they are actively processing information (listening, watching, reading, thinking) or communicating information (speaking, performing, writing) in ways that indicate they are focused on the task (Connell & Broom, 2004). Observers watch for behavioral engagement, such as participation and positive conduct; emotional engagement, such as signs of interest or boredom and reactions to the teacher and activities; and cognitive engagement, such as exertion of mental effort.

The EAR Protocol assesses classroom-level engagement by repeatedly noting the percentage of students who are on-task and the percentage actively engaged in the work. Conversations with students, when practicable, fine-tune these estimates. Once aggregated, these proportions serve as one indicator of instructional quality, based on the assumption that the extent to which students are engaged is a good marker of the extent to which the instruction is engaging. Engagement, as defined in this tool, is thus an indicator of instructional quality, rather than a characteristic of students, echoing the ideas of other researchers who argue that student engagement can be seen as a measure of the quality of the experience provided by the teacher or staff (Brophy, 2008; Smith & Hohmann, 2005) and that teachers play an important role in student engagement through task selection, classroom management, and instructional scaffolding (Fredricks, 2011).

Alignment

The second vital sign of high quality instruction is alignment, which is the extent to which the teacher is providing content that is on time and on target with what students need to learn, as specified by relevant state and local standards and assessments. Porter (2002) argued that instructional content has received little research attention despite widespread, commonsense understanding that students are most likely to learn what they are taught. He pointed out that measuring instructional content, and its alignment with materials, standards, and assessments, is important for many reasons, including accountability (e.g., parents and taxpayers), ensuring that students are exposed to a logical progression of instruction, monitoring reform and professional development efforts, and as a predictor of student achievement. Porter (2002) and Polikoff (2012) rely on a complex system to code the content of instruction (as reported by teachers), instructional materials, standards, and assessments. The system provides detailed quantitative codes for the information taught or tested and the cognitive demand on students. Once two sets of information (e.g., instruction and standards) have been coded, overlap can be assessed. This system yields important, fine-grained information but is not practical when administrators or technical assistance providers need a quick method for assessing instruction and monitoring change. Other researchers have used similar systems for quantifying the extent to which state-mandated assessment systems are aligned with state educational standards (Herman, Webb, & Zuniga, 2007; Webb, Herman, & Webb, 2007). That type of alignment is important, but differs from alignment in the EAR Protocol because its focus is the testing system, not the instruction.

The EAR Protocol defines alignment as students (1) being asked to do and actually doing schoolwork that reflects academic standards, and (2) having opportunities to master the methods used on high stakes assessments such as their state's standardized tests and college entrance exams. Alignment can be assessed in relation to district, state, or national standards and assessments. In aligned classrooms, what is being taught and what students are being asked to do are: in line with the standards and curriculum; "on time" and "on target" with the scope and sequence of the course of study; and provide students opportunities to experience high stakes assessment methodologies among other assessment approaches (Connell & Broom, 2004).

Rigor

Rigor*,* as defined in the EAR Protocol, reflects the commonsense notion that students will only achieve at high levels if that level of work is expected and inspected for all students. It was selected as the third vital sign of instructional quality for several reasons. First, the literature showed strong links between academic challenge and students' intrinsic motivation (e.g., Csikszentmihalyi, 1975; Danner & Lonky, 1981; Deci, 1975; Harter, 1978) and engagement (Shernoff, Csikszentmihalyi, Schneider, & Shernoff, 2003), and intrinsic motivation and engagement are linked to student achievement. Further, Lee and colleagues found that a rigorous high school curriculum was linked to higher achievement and lower dropout rates after controlling for students' background and past performance (Lee & Burkam, 2003; Lee, Croninger, & Smith, 1997). The EAR Protocol authors also reasoned that with rigorous instruction all students would be expected and supported to master material at levels sufficient to yield grade-level or better learning of material embodied by the standards and assessed by the state-mandated exams.

Although rigor is sometimes misinterpreted to mean schoolwork that is extremely hard or involves a lot of homework and classwork (Williamson & Blackburn, 2010), in the EAR Protocol it is intended to convey that expectations for all students are consistently high and that instructional strategies deployed by teachers ensure that the work requested optimally challenges all students to move from where they are toward higher standards. According to the EAR Protocol, in rigorous classrooms the learning materials and instructional strategies challenge and encourage all students to produce work or respond at or above grade level. All students are required to demonstrate mastery at these levels and have the opportunity for re-teaching as needed (Connell & Broom, 2004).

Students' Self-Reported Engagement in School and Perceived Academic Competence

In addition to testing whether the EAR Protocol would predict standardized achievement test scores, when the previous year's scores were controlled, the current study examined whether student reports of their engagement in school and perceived academic competence across their school experiences would add to this prediction. In contrast to classroom-level engagement as measured by the EAR Protocol—which is a measure of instructional quality—students' self-reports of their general engagement in school across all courses is a malleable, individual difference of the students. It has been shown to predict outcomes such as attendance and school dropout, as well as standardized test scores (Appleton, Christenson, Kim, & Reschly, 2006; Finn & Rock, 1997; Klem & Connell, 2004). Including it in models that also include classroom-level engagement will help remove the student individual differences so that the role of instructional quality can be seen more clearly.

Similarly, past research has established a link between perceived competence and change in students' academic performance (Connell, Spencer, & Aber, 1994; Gambone, Klem, Summers, Akey, & Sipe, 2004). Students who see themselves as "good at school" are likely to learn more in school, regardless of the quality of instruction. Thus, removing this variance from the models helps further understand the unique role of instructional quality. Including self-reported engagement in school and perceived academic competence is in line with the Gates Foundation (2012) report, which encouraged schools to use both classroom observation and student reports of their learning experiences to get a full picture of instruction in their schools, arguing that neither source of information alone is sufficient.

Method

Description of the EAR Classroom Visit Protocol

The first goal of this study was to describe the EAR Classroom Visit Protocol. This protocol is a 15-item observational tool completed by trained observers. The items appear in Table 1. The observer watches the class for 20 minutes and, while in the classroom, takes notes and makes tallies that are specific to the instrument's items. He or she then responds to the 15 items directly after the observation. Typically teachers receive multiple 20-minute observations throughout the school year to gain an in-depth picture of their instruction. The observers must be experienced educators, such as administrators, teachers, technical-assistance providers, or researchers with past classroom experience, and they must be trained in use of the protocol. For each of the 15 items, the protocol includes "visitor prompts" or reminders about what specific types of behaviors they should observe and note.

Engagement. Classroom visitors use two items to assess engagement: one measures the percentage of students who are on-task, and the second measures the percentage of on-task students who are actively and intellectually engaged in the work. For both items, trained observers walk around the classroom, inspecting students' work, watching students' facial expressions, and listening to students' conversations and responses to teachers' questions. An example of a visitor prompt reads: "Perform a 'quick scan' and estimate the percentage of students who appear to be on task: thinking, speaking, writing, making, listening." During training, observers are reminded to repeat the room scans, looking at all students, and tally the results for both "on-task" and "actively engaged" several times during the visit. Additionally, the training involves extensive conversations about how to determine which students are on-task and actively engaged. Further, classroom visitors have brief conversations with a few students who appear to be on task, if they can do so without being disruptive. Questions include: "What does your teacher expect you to learn by doing this work?" and "Why do you think the work you are doing is important?" The open-ended questions require students to explain their answers, so they cannot simply provide socially desirable responses. The conversations fine-tune the observational estimate of the percentage of actively engaged students. It is important to note that the observations are at the classroom level (not the student level) because they refer to all students in the class.

Alignment. Observers make eight binary judgments of whether the learning materials and activities, expectations for student work, and students' class work reflect relevant federal, state, and local standards, designated curricula, and high-stakes assessments. When available, observers are provided with pacing guides for the observed courses to aid their judgments about whether materials and instruction is "on time" and "on target." One visitor prompt reads: "Are students being asked to

Table 1: *EAR Protocol Descriptive Statistics (n = 2,171)*

	Item	Mean	SD
Engagement			
E1	% of students on task.	77%	21
E2	% of students actively engaged in the work requested.	63%	28
	Product of E1 * E2[1]	53%	30
Alignment			
A1a	The learning materials did$_{(1)}$ / did not$_{(0)}$ reflect content standards guiding this class.	.94	.23
A1b	The learning materials were$_{(1)}$ / were not$_{(0)}$ aligned with the designated curriculum to teach those standards.	.93	.25
A1c	The learning materials were$_{(1)}$ / were not$_{(0)}$ aligned with the pacing guide of this course or grade level curriculum.	.89	.31
A2a	The learning activities did$_{(1)}$ / did not$_{(0)}$ reflect content standards guiding this class.	.92	.26
A2b	The learning activities were$_{(1)}$ / were not$_{(0)}$ aligned with the designated curriculum to teach those standards.	.92	.27
A2c	The learning activities were$_{(1)}$ / were not$_{(0)}$ aligned with the scope and sequence of the course according to the course syllabus.	.88	.33
A3	The student work expected was$_{(1)}$ / was not$_{(0)}$ aligned with the types of work products expected in state grade level performance standards.	.72	.45
A4	Student work did$_{(1)}$ / did not$_{(0)}$ provide exposure to and practice on high stakes assessment methodologies.	.56	.50
Rigor			
R1	The learning materials did$_{(1)}$ / did not$_{(0)}$ present content at an appropriate difficulty level.	.89	.32
R2	The student work expected did$_{(1)}$ / did not$_{(0)}$ allow students to demonstrate proficient or higher levels of learning according to state grade level performance standards.	.59	.49
R3	Evaluations/grading of student work did$_{(1)}$ / did not$_{(0)}$ reflect state grade level performance standards.	.37	.48
R4	% of students required to demonstrate whether or not they had mastered content being taught.	35%	38
R5	% of students demonstrated threshold levels of mastery before new content was introduced.	12%	26

[1] E2 refers to the proportion of those students who were on task (in E1) who were actively engaged, so E1 and E2 must be multiplied together to be meaningful.

complete work that aligns with the kinds of work products expected to meet or exceed state and district grade level standards?"

Rigor. This construct is assessed with five judgments (three binary, two percentages) concerning the cognitive level of the material, the work expected, and the extent to which students are required and supported to demonstrate mastery of the content. Items concern whether learning materials and work products are appropriately challenging, whether students are expected to meet/surpass relevant standards, and whether they have an opportunity and support to demonstrate proficiency. In the visitor prompts, observers are asked to consider the level of thinking and performing required by the learning activities, as defined in Bloom's taxonomy (Bloom, Englehart, Furst, Hill, & Krathwohl, 1956). One visitor prompt reads: "Code student work expected as rigorous only if preponderance of observed work *and* work expected during classroom visit was at Intermediate Level and some of the work was at the Advanced Level."

EAR Protocol Data Collection

This study took place in four high schools from a single district during the 2008–2009 school year. The schools were relatively large, with an average student enrollment over 1,500. Over 40% of the students enrolled in these schools were Latino/Hispanic and a roughly equal percentage was non-Hispanic, White. About one-third of the students were from low-income families. EAR Protocol data were collected for multiple purposes: to support future professional development; to establish a scoring system with continuous variables that could be used for research; to investigate the tool's inter-rater reliability; and to assess the tool's ability to predict standardized test scores in math and ELA.[1]

Data were collected by three groups of individuals: (1) IRRE consultants (n = 9) who had used the tool extensively over several years and were also providing instructional support in these schools, (2) former educators hired expressly for this project who had deep knowledge of high school classroom practices but no direct connection with these schools (n = 3), and (3) school leaders such as principals, assistant principals, and instructional coaches from the participating schools (n = 21).[2] The former educators and school leaders were trained by IRRE using their standard training procedures that consist of (1) two full days of group instruction, including several classroom visits followed by scoring discussions, (2) a two- to three-week window during which those participating in the training make practice visits as teams to calibrate their scoring, and (3) two additional full days of group instruction focusing on calibration and use of the data for instructional improvement. For additional information about EAR Protocol training, see IRRE (2012).

In all, 2,171 EAR Protocols were collected during the 2008-09 school year; 416 were collected by IRRE consultants, 347 by the former educators, and 1,408 by school leaders. Table 1 presents descriptive statistics for the 15 individual indictors across the 2,171 observations. These observations, which were made in all types of courses, including math, ELA, science, history, art, and special education, were used for the Confirmatory Factor Analysis (CFA) discussed below. Only data from 10th grade math and ELA classes were used in the analyses to predict test scores because those are the subjects for which standardized test scores were available. In these predictive validity analyses we used only data from fall observations of math and ELA classes

[1] Throughout this manuscript, we use the term English language arts (ELA) to describe courses and exams focused on comprehension, reading, and writing in English. The courses included typical high school English courses, as well special education English and English courses specifically designed for English language learners that met the regular English course requirement. The exam was called "Reading."

[2] School district employees were trained by IRRE to collect EAR Classroom Visit Protocol for their own instructional improvement purposes, as well as for this study. The district decided which individuals would collect data for their purposes. Eight individuals (in addition to these 21) who worked for the district conducted EAR Visits during the year but either did not participate in any inter-rater reliability visits (n = 2) or did not appear to understand the tool based on preliminary analyses using thresholds set by IRRE (n = 6). Thus, their data were used for the districts' internal purposes only and have been excluded entirely from this research.

for two reasons. First, the exams were administered early in the spring term so spring teachers would have had little influence on scores. Second, we sought to minimize any potential effects of professional development resulting from the districts' use of the tool. The tool was introduced to this district during the fall term. District leaders were just learning to use both the instrument and the data it provides at the time of data collection so virtually no tool-based professional development with teachers took place. Thus, the predictive validity analyses included 125 observations of 33 different ELA teachers and 102 observations of 25 different math teachers.

Student Questionnaires

In the fall of 2008, all 10th grade students at the four high schools were asked to respond to an online questionnaire administered during the school day. In all, the schools enrolled 1,690 10th graders and 1,620 of these students (96%) responded to the questionnaire. However, as detailed later, the current analyses include only the 1,144 10th graders who had complete data (9th and 10th grade standardized test scores in math and/or ELA and classroom observations).

The questionnaires were developed by IRRE, based on similar items used in their past work (Akey, 2006; Klem & Connell, 2004). Two scales were of particular interest in this study: self-reported engagement in school and perceived academic competence. The measure of self-reported engagement in school asked students to respond to six items, using a four-point scale ranging from "not at all true" to "very true." Sample items include: "It is important to me to do the best I can in school" and "I pay attention in class." Cronbach's alpha on this scale in the current sample was .70 (n = 1,144, *mean* = 3.04, *SD* = 0.48). The measure of perceived academic competence included six items, using the same four-point response scale. Sample items include: "I feel confident in my ability to learn at school" and "I am capable of learning the material we are being taught at school." Cronbach's alpha was .76 (n = 1,144, *mean* = 3.23, *SD* = 0.51).

Standardized Achievement Tests

The 10th grade outcome for this study was the Arizona Instrument for Measuring Standards (AIMS) High School Exit Exam. Students in Arizona begin taking a high school exit exam in the spring of the 10th grade in ELA and math, repeating it each semester until they pass. For this study, only scores from the first administration of this exam were used. Pearson, PLC (2010) published a detailed report on the development and psychometric properties of these exams. According to that report, Cronbach's alpha for ELA is .92 and for math is .95.

Additionally, this district administers a nationally normed standardized assessment called the Terra Nova in math and ELA to all 9th graders. Brown and Coughlin (2007) found that the Terra Nova had strong internal consistency (.80-.95) and moderate to strong test-retest reliability (.67-.84). Further, they reported strong criterion and predictive validity. The district provided the research team with all available 9th grade Terra Nova and 10th grade high school exit exam scores for students who were in 10th grade in the 2008–2009 school year.

Analytic Plan

Scoring the EAR Classroom Visit Protocol. The second goal of this study was to establish a system for creating continuous scores from EAR Protocol data. IRRE has used the EAR Protocol extensively in its instructional improvement efforts. In order to give straightforward feedback to educators, IRRE has used thresholds to indicate whether a classroom is at an acceptable level on each EAR vital sign, and then calculated the percentage of classrooms within a department, grade, or school that has met the threshold on each of the vital signs. However, for research purposes, it is preferable to have continuous variables that express the full range of variance on these constructs and maximize power when analyzing associations between the quality of instruction and student outcomes. To this end, we used Analysis of Moment Structures (AMOS) to test a series of CFA measurement models to evaluate the appropriateness of various methods of creating continuous scores. We started by testing models that measured engagement, alignment, and rigor separately and included all 15 indicators. Next, we deleted some indicators, as warranted by the initial models, creating the strongest and most parsimonious measures of each of the three constructs. Last, because E, A, and R were inter-correlated, we tested a single-indicator model.

Inter-rater agreement. The third goal of the study was to establish the tool's inter-rater agreement across the different types of users, applying the continuous scoring system. To meet this goal, we calculated intraclass correlations between pairs of scores, using 388 cases where two data collectors were present during the observation.

Predictive validity. The fourth goal of the study was to investigate the relations between classroom instruction as measured by the EAR Protocol and standardized test scores, while controlling for previous test scores. We used Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) to model these data in which students were nested within sections (i.e., specific period of a specific teacher), and sections were nested within teachers. The analyses that included only E, A, or R were followed by analyses that added the students' self-reported engagement in school and perceived academic competence. This order was selected because many users of this tool will likely not have other student-level data, and we wanted to be sure the data gathered by the observational tool was predictive by itself.

Results

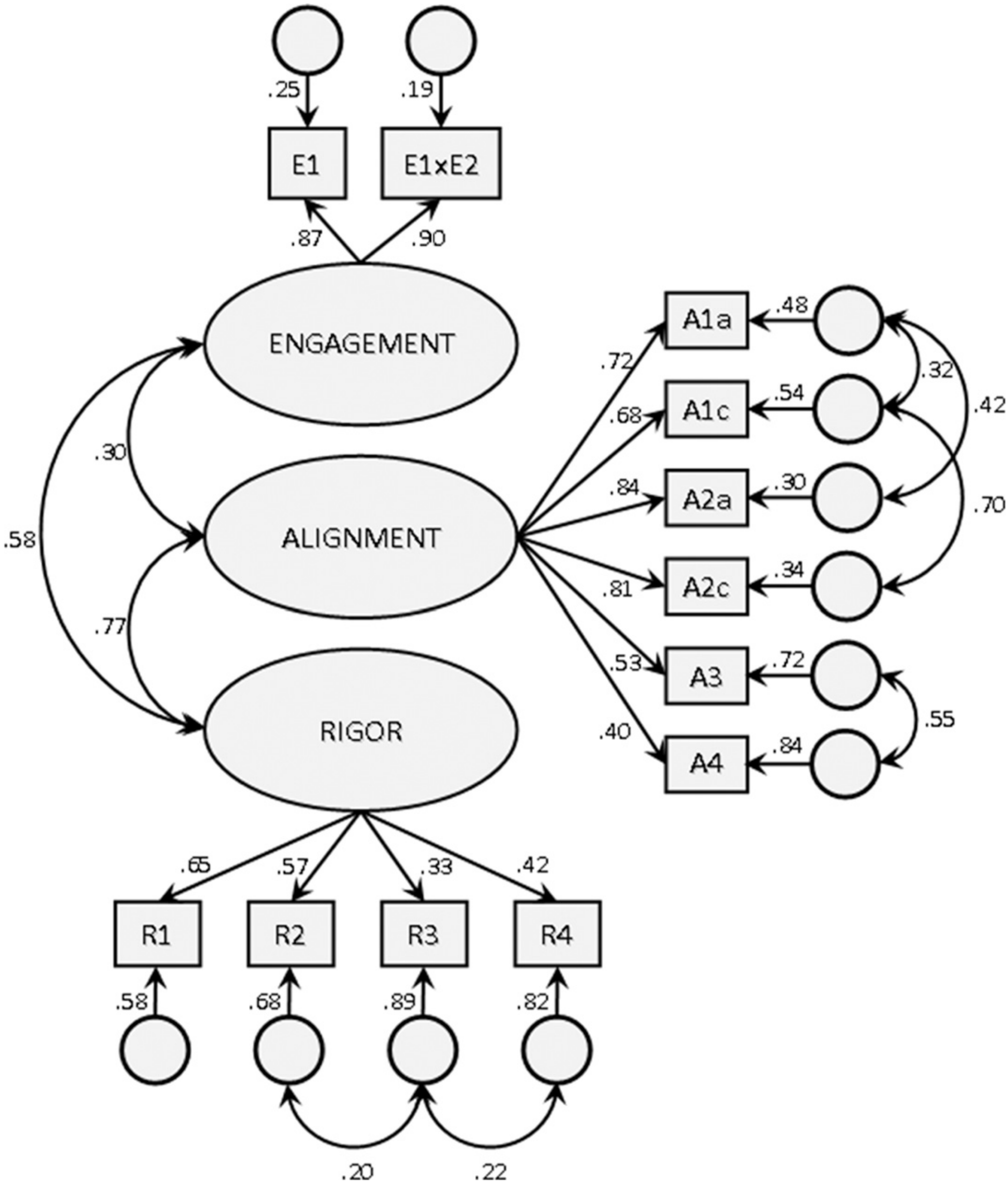
Scoring the EAR Classroom Visit Protocol

To meet the study's second goal—establishing a system for creating continuous scores from EAR Protocol data —we used the 2,171 observations to test CFA measurement models to evaluate the appropriateness of various methods of creating continuous scores. We chose to analyze these 2,171 observations as they represent all of the observations made across 271 teachers of 821 sections, across an entire academic year in four separate schools. Thus, they represent the true diversity of ratings that are likely to be obtained using this instrument.[3]

In the final CFA model (see Figure 1), the two engagement percentages were used to form latent indicators of engagement: the first indexing the proportion of students who were on task, and the second indexing the proportion of students who were both on task and actively engaged. Alignment included six dichotomous items in the final CFA. Originally, alignment had been measured by eight dichotomous items, but 88–94% of the responses were positive for the first six items (see Table 1), meaning those items provided little information and were redundant. As a result, two of those items were excluded (A1b, A2b) from the final CFA model, and the six remaining items were used to form latent indicators of alignment. For rigor, three

[3] We acknowledge that there is dependency within these observations as they are nested within 271 teachers, suggesting that the true N for these analyses might be as low as 271. We would argue that 1) observations of teachers in different classes and on different days are still likely to yield unique information for these analyses, and 2) that an effective N of as low as 271 would still support these analyses. In fact, an additional CFA based on just aggregate scores for the 271 teachers yielded highly similar path coefficients, suggesting that the dependencies within the 2,171 observations did not excessively distort the results.

Figure 1: *Final measurement model for creating continuous scores from the Engagement, Alignment, and Rigor Classroom Visit Protocol. ($\chi^2(45) = 1207$, SRMR = .090, CFI = .916; RMSEA = .109).*



of the five indicators were dichotomous and two were continuous. In order to combine the rigor indicators on a common scale, all five were standardized using estimates of population means and standard deviations from 1,551 observations conducted by the IRRE intervention team in 19 high schools in six school districts across the country between 2004 and 2010. After standardization, the five items were entered as latent indicators of rigor. Given the similarity of items within the E, A, and R indicators, six pairs of error terms were allowed to correlate to model those slight redundancies. Although this model demonstrated adequate fit, the fifth rigor indicator (R5), measuring the proportion of students who had to demonstrate mastery prior to the introduction of new material, had a notably low path coefficient (.25; suggesting that it shared only about 6% of variance with latent rigor estimate), and was quite skewed with 88% of the observations having a value of 0 (suggesting that it also offered little variance to the measurement of rigor). Thus, R5 was omitted and the model was re-run. The resulting final model had adequate fit indices ($\chi^2(45) = 1207$, $SRMR = .090$, $CFI = .916$; $RMSEA = .109$). This model indicates that the various elements of the EAR assessment can be combined to create separate continuous engagement, alignment, and rigor scores as their observed covariance pattern in the data supports that factor structure.

As indicated in Figure 1, engagement, alignment, and rigor were all correlated, so we tested whether a model with a single underlying construct would fit the data better than the model with the three latent constructs. The fit from that model was unacceptable ($\chi^2(48) = 3095$, $SRMR = .107$, $CFI = .779$; $RMSEA = .171$) indicating that a single variable would not satisfactorily represent instructional quality, so we used the three-latent-construct model.

Based on these measurement models, three scores were created. Engagement was the mean of proportion of students on task (E1) and the proportion of students actively engaged in the work (E1 X E2). The CFA results suggested there continued to be substantial shared error (method) variance among the dichotomous alignment indicators, so we dropped two additional indicators and accepted alignment as the proportion of positive answers on the four remaining (fairly distinct) dichotomous indicators (A1c, A2c, A3, and A4). Rigor was the mean of the four standardized rigor indicators (R1, R2, R3, R4), excluding R5. Across the observations, the three variables were correlated but not so highly as to be measuring the same construct (E and A $r = .32$, $p < .000$; E and R $r = .44$, $p < .000$; A and R $r = .63$, $p < .000$).

Inter-Rater Agreement

To meet the third goal of this study—investigating the tool's inter-rater reliability—a pair of observers coded 388 cases across the 2008–2009 school year. Inter-rater reliability was calculated as the intraclass correlation (one-way random, absolute agreement) between pairs of scores. After calculating continuous scores on engagement, alignment, and rigor using the scoring method described above, the single measures intraclass correlation was .76 for engagement, .71 for alignment, and .65 for rigor. Of the 388 pairs, there were 238 where the pair was made up of an IRRE consultant and a school leader. Looking just at this sub-set, the single measures intraclass correlations remained unchanged: .76 for engagement, .71 for alignment, and .65 for rigor. There were 107 observations where the pair was made up of an IRRE consultant and one of the external observers from the research team (i.e., the former educators). Looking just at this subset, the intraclass correlations were: .72 for engagement, .62 for alignment, and .67 for rigor. Thus, all ICCs fall within the "good" (.60 to .74) or "excellent" (.75 to 1.0) range (Cicchetti, 1994).

Predictive Validity

Cases available for predictive validity analyses. To meet the study's fourth goal—examining the tool's ability to predict student test scores—the subset of observations that were collected in the fall of 2008 in 10th grade math and ELA classes was used to test the predictive validity of the EAR Protocol. In math classes, 125 observations were conducted of 33 teachers teaching 57 sections (i.e., specific period of a specific teacher). On average, each math teacher was observed 3.68 times (*range* = 1 to 8, *SD* = 2.18). The full range of math classes that enrolled 10th graders was included in the analyses: special education math, Algebra I, Algebra II, College Algebra, and Honors Pre-Calculus.

In ELA classes, 102 observations were conducted of 25 teachers teaching 50 sections. On average, each ELA teacher was observed 4.08 times (*range* = 1 to 7, *SD* = 1.89). Most 10th graders were enrolled in 10th grade English, so those were the primary courses included. However, 9th grade English and special education English courses that included some 10th graders were also included. Tenth grade English courses specifically designed for English language learners (ELLs) were included along with those without that special focus.

After calculating continuous E, A, and R scores for each observation using the scoring described above, a mean E, A, and R score was calculated for each section and each teacher. Thus, although there was variation in the number of observations per section and per teacher, each is represented equally in the final data set.

Math teachers were relatively diverse (43% female; 75% White, 10% Latino, 10% Multi-Racial). They had an average of 5.09 years of teaching experience (*SD* = 1.23) and had been in their current positions 2.50 years (*SD* = 1.06) on average. The ELA teachers were less diverse: 81% female and 93% White. ELA teachers had an average of 4.94 years of teaching experience (*SD* = 1.18) and had been in their current positions 2.38 years (*SD* = 1.29) on average.

In all, 1,144 students were included in the predictive validity analyses (483 in both the math and ELA analyses, 151 in the math analyses only, and 510 in the ELA analyses only). The participating students were 51.0% female. They were 41.5% non-Hispanic White, 40.9% Latino/Hispanic, and 11.9% African American. There were 634 students available for the math analyses, meaning that they had both 9th and 10th grade math scores and their math section had been observed. There were 993 students available for the ELA analyses, meaning that they had both 9th grade and 10th grade ELA scores and their ELA section had been observed. The sample size dropped slightly when student questionnaires were added to the models (*n* = 621 for math; *n* = 975 for ELA) due to some missing student questionnaires.

Multi-level model description. We used Hierarchical Linear Modeling (HLM; Raudenbush & Bryk, 2002) to predict 10th grade exit exam scores in math or ELA, controlling for the previous year's score in the same subject, as a function of observed math or ELA classroom E, A, or R. Specifically, we built 3-level models in which between-student differences within sections, as measured by students' previous year's test scores, were modeled at level 1; within teacher (between section) variation was modeled at level 2; and between teacher variation was modeled at level 3. Math and ELA outcomes were modeled separately, using observed E, A, or R in fall math classes in the math models and observed E, A, or R in fall ELA classes in the ELA models. The equations of a typical model were:

Level 1 $\quad$ 10th grade score = $\pi_0 + \pi_1$ (9th grade score) + e
Level 2 $\quad$ $\pi_0 = \beta_{00} + \beta_{01}$ (variation in observed E, A or R across classes within a teacher) + r_0
$\pi_1 = \beta_{10}$
Level 3 $\quad$ $\beta_{00} = \gamma_{000} + \gamma_{001}$ (variation in observed E, A, or R across teachers) + u_{00}
$\beta_{01} = \gamma_{010}$
$\beta_{10} = \gamma_{100} + u_{10}$

As seen in the equations, the association between 10th and 9th grade test scores was allowed to vary across teachers (as a level 3 random effect). This allowed for different strengths of association between 9th grade and 10th grade performance across teachers as unusually effective (or ineffective) instruction by a specific teacher could reduce how strongly 10th grade performance would be predicted by 9th grade scores. In addition, average 10th grade test scores were allowed to vary across classes (as a level 2 random effect) and to vary across teachers (as a level 3 random effect). This essentially allowed the model to estimate different average levels of 10th grade performance (e.g., intercepts) for each of the sections examined and for each of the teachers examined, recognizing that certain sections and certain teachers might have contained students with stronger or weaker skills. The remaining effects were set as fixed effects as we did not expect them to vary widely across sections or teachers.

Both the predictor and outcome variables were standardized prior to running these analyses, essentially converting the HLM coefficients into standardized coefficients. Secondary models including students' self-reported engagement in school and perceived academic competence were run to investigate the predictive role of these individual student characteristics beyond that of classroom-level observations. Prior to running the models of interest, we ran fully unconditional models to better understand variation at the three levels of interest. All HLM results appear in Table 2.

Predicting math scores from EAR observations. The fully unconditional 3-level model suggested that 38.4% of the variance in 10th grade math scores was at level 1 (differences between students within the same section), 15.3% of the variance was at level 2 (differences between sections of the same teacher), and 46.3% of the variance was at level 3 (differences between teachers). Next, we ran the HLMs outlined above. As seen in Table 2, 9th grade test scores served as a strong predictor of 10th grade test scores. A 9th grade test score one standard deviation above the mean predicted a 10th grade test score .61-.62 standard deviations higher on math, suggesting a strong component of student ability and/or past instruction in 10th grade scores. After controlling for these effects, when observations of E, A, or R were separately allowed to predict residual change in standardized test scores, the results offered support for each as predictors of student achievement. As seen in the first set of columns for predicting 10th grade math scores (on the three rows labeled “Differences Across Teachers in…”), for every standard deviation higher than average a fall math teacher was rated on engagement, the model predicted his or her students scoring an average of .17 standard deviations higher on their 10th grade math tests, even after controlling for their 9th grade math scores and variations of engagement among the teacher's different sections. Similarly, for observed alignment, a +1 standard deviation difference for a teacher predicted a statistically significant +.16 standard deviation difference in students' scores, and for rigor a +1 standard deviation difference for a teacher predicted a marginally significant +.14 standard deviation difference in students' scores.

Thus, under the stringent conditions of predicting standardized math test scores in 10th grade after controlling for standardized scores one year earlier, we found evidence that each of the teaching variables of observed engagement, alignment, and rigor explained some variance. Math teachers whose instruction was more engaging, aligned, and rigorous had students who showed greater gains on standardized tests. These results further suggest that the dimensions assessed by the EAR tool capture aspects of classroom dynamics and effective instruction that lead to measurable real-world gains in learning, underscoring the utility of this instrument.

Predicting ELA scores from EAR classroom observations. The fully unconditional 3-level model suggested that 76.5% of the variance in 10th grade ELA scores was at level 1 (differences between students within the same section), 2.5% of the variance

Table 2: *Observed Engagement, Alignment, or Rigor Predicting Standardized Test Scores*

	Predicting 10th Grade MATH Scores						Predicting 10th Grade ELA Scores					
	without student questionnaire variables (n_i students = 634 n_j sections = 57 n_k teachers = 33)			including student questionnaire variable (n_i students = 621 n_j sections = 57 n_k teachers = 33)			without student questionnaire variables (n_i students = 993 n_j sections = 50 n_k teachers = 25)			including student questionnaire variables (n_i students = 975 n_j sections = 50 n_k teachers = 25)		
Predictor Variables	Coeff.	SE	df	Coeff.	SE	df	Coeff.	SE	df	Coeff.	SE	df
Observed Engagement (pseudo-R^2)	*.64*			*.68*			*.61*			*.61*		
Intercept	−0.09	0.07	31	−0.07	0.06	30	0.02	0.03	23	0.02	0.03	23
Student Self-Reported Engagement in School				0.06*	0.03	615				0.06*	0.02	969
Student Perceived Academic Competence				0.03	0.03	615				0.04^{+}	0.02	969
Previous Year's Test Score	0.62***	0.04	32	0.62***	0.04	31	0.74***	0.03	24	0.72***	0.03	24
Within Teacher Variation in Engagement	−0.03	0.08	55	−0.03	0.07	54	0.04	0.04	48	0.04	0.04	48
Differences across Teachers in Engagement	0.17*	0.07	31	0.21***	0.05	30	0.06^{+}	0.03	23	0.06^{+}	0.03	23
Observed Alignment (pseudo- R^2)	*.63*			*.66*			*.61*			*.61*		
Intercept	−0.11	0.07	31	−0.09	0.07	30	0.01	0.03	23	0.02	0.03	23
Student Self-Reported Engagement in School				0.06*	0.03	615				0.05*	0.02	969
Student Perceived Academic Competence				0.03	0.03	615				0.05^{+}	0.02	969
Previous Year's Test Score	0.62***	0.04	32	0.63***	0.04	31	0.73***	0.03	24	0.72***	0.03	24
Within Teacher Variation in Alignment	−0.01	0.07	55	−0.01	0.06	54	0.02	0.03	48	0.02	0.03	48
Differences across Teachers in Alignment	0.16*	0.08	31	0.18*	0.07	30	0.06^{+}	0.03	23	0.04	0.03	23
Observed Rigor (pseudo-R^2)	*.62*			*.65*			*.61*			*.61*		
Intercept	−0.12	0.07	31	−0.09	0.07	30	0.02	0.03	23	0.02	0.03	23
Student Self-Reported Engagement in School				0.06*	0.03	615				0.06*	0.02	969
Student Perceived Academic Competence				0.02	0.03	615				0.04^{+}	0.02	969
Previous Year's Test Score	0.61***	0.04	32	0.62***	0.04	31	0.74***	0.03	24	0.72***	0.03	24
Within Teacher Variation in Rigor	−0.05	0.07	55	−0.05	0.06	54	0.04	0.04	48	0.03	0.04	48
Differences across Teachers in Rigor	0.14^{+}	0.07	31	0.15*	0.07	30	0.10*	0.04	23	0.07	0.05	23

This table reflects twelve separate analyses: E, A, and R, for math and ELA, with and without variables from the student questionnaires. Both the predictor and outcome variables are standardized, essentially converting the HLM coefficients into standardized coefficients. $^{+}$p < .10, *p < .05, **p < .01, ***p < .001

was at level 2 (differences between sections of the same teacher), and 21.0% of the variance was at level 3 (differences between teachers). Looking at the models predicting 10th grade ELA scores (the right portion of Table 2), these ELA models suggest a strong component of student ability and/or past instruction (.73-.74). Further, as with math, we see there was evidence that the three dimensions of instructional quality were linked to student outcomes, but the results were somewhat weaker. For every standard deviation where a fall ELA teacher was rated higher than average on engagement or on alignment, the model predicted his or her students scoring an average of .06 standard deviations higher on their 10th grade standardized ELA tests (both effects marginally significant), after controlling for 9th grade ELA score and variations among the teacher's different sections. For observed rigor, a +1 standard deviation difference in teachers predicted a significant +.10 standard deviation difference in ELA scores. Thus, we see evidence that observations of teachers' instructional quality predict students' improvement in ELA, although the size of the effects was not as large for ELA as for math.

Between-section variation. In the six models without student questionnaires presented in Table 2, the between-section variation on E, A, and R within a teacher (level 2) were non-significant (see the three lines on the table labeled "Within Teacher Variation in…"). This indicates that the engagement, alignment, and rigor of a particular section was not predictive of student outcomes above and beyond that section's teacher's overall level, suggesting that the common experiences teachers are creating across different sections of their courses predict student growth in learning more than differences they create between these different sections.

Predicting math scores from EAR observations and student self-reports. The second and fourth set of columns in Table 2 summarize the HLM results after the student reports of engagement in school and perceived academic competence were included to assess their unique contribution to student learning beyond the quality of observed instruction. In the three models predicting 10th grade math scores (second set of columns for math), students' self-reported engagement in school was a significant predictor of their math test scores while perceived competence was not. Thus, higher student self-report of engagement in school was associated with slightly higher 10th grade math scores. After controlling for student reports of engagement and competence, observed E, A, and R in math classes all remained significant predictors of 10th grade math scores. Specifically, higher observed levels of a teacher's E, A, or R predicted significantly higher average 10th grade math scores in his or her students ranging from +.15 to +.21 standard deviations. These results indicate that both the observed quality of students' instructional experience and their generalized sense of engagement in school uniquely contributed to their performance on standardized assessments.

Predicting ELA scores from EAR observations and students' self-reports. The final set of columns in Table 2 suggest that higher student reports of their own engagement in school were also associated with slightly higher 10th grade ELA scores even after controlling for 9th grade test scores. Additionally, perceived academic competence was a marginal predictor of 10th grade ELA scores. Of the three observed vital signs, only observed engagement in the ELA classrooms was marginally predictive of student ELA achievement in these models, suggesting that the students' experience of engagement may have more pervasive effects across subject matter areas.

Discussion

Summary of Findings

This study had four goals: (1) to describe the EAR Protocol, (2) to devise a continuous scoring system for the protocol, (3) to establish the extent to which different groups of observers could attain acceptable levels of inter-rater reliability, and (4) to test the extent to which the EAR Protocol, when scored using the continuous system, would predict student test scores. The protocol was described in detail in the Method section and a continuous scoring system was established using CFA. Regarding the third goal, this study indicated that school and district personnel, as well as educators from outside the district, could learn to reliably use the EAR Classroom Visit Protocol. Regarding the fourth goal, observed engagement, alignment, and rigor were each positively, significantly or marginally, linked to math and ELA achievement after controlling for the previous year's test scores. When self-reports of student engagement and perceived academic competence were added to the models, self-reports of generalized engagement in school predicted achievement in math and ELA after controlling observed E, A, or R and the previous year's test scores. Students' perceived academic competence was marginally associated with ELA scores but not with math scores.

Engagement

Student engagement, measured in various ways, has often been linked to academic success (Appleton, et al., 2006; Finn & Rock, 1997; Klem & Connell, 2004). In this study, EAR Protocol observations of instruction being engaging for students predicted math achievement significantly and ELA achievement marginally after controlling for the prior year's scores. The EAR Protocol assessed the extent to which students were paying attention, doing the work requested, and appearing cognitively involved in the task. Observers watched students' behavior and facial expressions, and, when possible, had brief conversations with some students. The EAR Protocol's assessment of classroom-level engagement is a relatively quick and simple measure of a complex and fundamental construct, so its ability to predict student achievement, controlling for past achievement, adds an important tool for measuring instructional quality.

The fact that observed classroom-level engagement continues to predict students' test scores when controlling for individual student's self-reported engagement in school shows that these are two somewhat distinct types of engagement. Engagement assessed with the EAR Protocol was based on students' displayed behavior, affect, and cognition in that class. When aggregated across all students in the class, it signified the extent to which the teacher was instructing in a way that engaged the students. A student's self-report of general engagement in school, on the other hand, was an individual difference characteristic of that student. Using measures similar to the one used for this study, general engagement in school has been shown to predict standardized test scores and other important school outcomes such as attendance and school dropout (Appleton et al., 2006; Finn & Rock, 1997; Klem & Connell, 2004). Of course, these two types of engagement are related but perhaps in complex ways. For example, consistent experiences of engaging instruction across many classrooms should contribute to student reports of being generally engaged in school; but high school students who are generally engaged in school might disengage in classes where instruction is not engaging. The current study, as well as the Gates Foundation (2012) work, indicates that measuring both types of engagement are important as they account for independent variance in student outcomes.

Alignment

It is important for the instruction provided to map onto standards and assessments if we expect the instruction to make meaningful differences in students' learning and demonstrating what is desired. Indeed, Polikoff (2012) referred to alignment of instruction with standards and assessments as the "key mediating variable separating the policy of standards-based reform to the outcome of improved student achievement" (p. 341). However, established systems for measuring alignment rely on detailed coding systems (Polikoff, 2012; Porter, 2002) making them difficult to use regularly to measure change or provide feedback to teachers. Alignment, as measured by the EAR Protocol, seems to be a valuable alternative to these time-consuming systems.

Rigor

Likewise, rigor—defined as a combination of appropriate difficulty and continuous checking to ensure the students are mastering the content—is a commonsense requirement for improved outcomes. Students are more likely to be intrinsically motivated when content is challenging (Deci, 1975; Harter, 1978), and schools that provide rigorous curricula have higher student achievement after controlling for background characteristics (Lee et al., 1997). However, the field had been lacking a measure of rigor that is feasible for administrators and consultants to use regularly. This study indicates that the EAR Protocol can fill this gap and is linked to achievement.

Math versus ELA

The associations between observed E, A, and R and student achievement scores were noticeably stronger for math than ELA. Similarly, The Gates Foundation (2012) found that links between classroom observations and math scores were roughly twice as large as those for ELA. It is possible that this difference stems from ELA achievement being harder to measure than math achievement, making the ELA tests scores less reliable. The Gates Foundation (2012) report points out that most state ELA tests consist solely of multiple-choice reading items, and this was the case in the current study. This type of reading assessment may not be sensitive to the work teachers do with regard to writing and other literacy objectives. Another possibility is that ELA achievement is more influenced than math by factors outside of the teachers' control, such as language experiences at home and in other courses. Last, it may be that it is more difficult to accurately judge ELA instruction. Indeed, Polikoff (2012) argued that math standards and assessments may be more concrete and more easily understood than ELA standards and assessments. The same may hold true for classroom observations.

Limitations

Single-district study. The current study took place in four high schools within a single district. Although this district was diverse with regard to student race/ethnicity and served a fairly high proportion of low-income students, it is impossible for a single district to represent the diversity of districts in the U. S. There are several ways that the findings might have been different had the study been conducted in an array of locales. First, this district had well defined pacing guides for every course that were provided to all individuals collecting EAR Protocol data. The classroom visitors could thus easily determine if the course was on pace and teaching the district-supported content. Further, the district had been careful to base the pacing guides on the state grade-level performance standards and standardized tests, ensuring that the correct material was covered prior to the exam, at the appropriate level of rigor. Many districts in the U.S. lack such pacing guides, which would likely diminish the reliability of the A and R observations, thus decreasing their links with student outcomes. Also, all of the school leaders and former educators were from the state in which the data were collected, so they were very familiar with the state standards and assessments. Findings might have been weaker if the classroom observers had had less familiarity with the state's education system.

These data suggest that when the curriculum and pacing are clear and map well onto the tests, and when the raters are very familiar with the state's expectations, alignment and rigor can predict student scores. We do not know if the same is true when the pacing guides lack detail or are not mapped onto the tests, or when the raters are less familiar with the state standards.

Internal agreement. Another limitation was that the internal consistency of the self-report measure of engagement in school (Cronbach's alpha = .70) and the inter-rater reliabilities on the EAR Protocol (range = .62 to .76) were at the lower end of the acceptable range. This might have attenuated the strength of the correlational associations for these constructs. Thus, the results presented for self-reported engagement and the EAR Protocol scores may slightly underestimate the strengths of associations that would have been obtained with a more internally consistent scale and stronger inter-rater reliability. We maintain, however, that these levels of inter-rater reliability on the EAR Protocol are fairly impressive given the large number of observers, with different roles and backgrounds. Moreover, the fact that the EAR Protocol significantly predicted changes in standardized test scores suggests that it had sufficient inter-rater reliability to be useful.

Small effect sizes. The effect sizes presented here are in the small range (.14 to .17 SD for math; .06 to .10 SD for ELA). Nonetheless, we believe the tool is of value because it is one of only a few classroom instruction assessments shown to predict student outcomes, is appropriate for high school instruction, and can reasonably be learned and employed by school leaders and technical assistance providers in their daily work.

Different scoring systems. As noted earlier, IRRE typically uses thresholds to communicate with school leaders about the extent to which instruction is engaging, aligned, and rigorous. For this study, we created and applied a continuous scoring system to maximize variance and therefore increase our power to detect effects. This continuous scoring system is not currently in use in schools and might prove too difficult for schools to apply, so the predictive validity results presented here might not apply to the way in which districts are likely to use the data.

Next Steps

Recently, the authors of this article completed data collection for a large field trial to evaluate an intervention in high school ELA, Algebra 1, and Geometry classes. It included EAR Protocol data from 20 high schools, in five districts, in four states, with more than 500 ELA and math teachers. It also includes student questionnaires and test scores from roughly 20,000 students in those teachers' courses. That database can be used to further explore the psychometric properties of the EAR Protocol, the conditions under which it predicts student outcomes, and whether scores on the EAR Protocol can be enhanced by a targeted intervention.

Conclusions

School districts, technical assistance providers, and researchers need sound ways to assess the quality of instruction in the classroom in order to appropriately target professional development and to monitor changes that result from instructional improvement efforts. For such a system to be useful to educators, it needs to be feasible within the workdays of school personnel, provide immediate and actionable feedback, and give schools a common language with which to discuss high-quality instruction. To be useful for researchers it needs to be understandable for trained, independent observers, relatively brief, and have adequate psychometric properties.

The EAR Protocol meets all of these goals. It is feasible for both district personnel and researchers because it takes only 20 minutes, and in the current study both school personnel and outside observers learned to use it reliably. During training a lot of attention is paid to the EAR Protocol vocabulary, ensuring a common language among users. The data can provide quick and actionable feedback through IRRE's on-line report generating system. And importantly, based on this study, it appears to have predictive validity for state-administered achievement tests.

Acknowledgments

The research reported here was supported by grants from the Institute of Education Sciences, U.S. Department of Education (Grant R305R070025) and The Bill & Melinda Gates Foundation, both to the University of Rochester. The opinions expressed are those of the authors and do not represent views of any of the organizations. The authors thank the Institute for Research and Reform in Education for their collaboration in this work, and the participating students, teachers, schools, and school districts for their cooperation in the data collection efforts. Portions of this paper were presented at the November 2012 meeting of the Association for Public Policy Analysis & Management in Baltimore, Maryland. Dr. Early conducted this research as a faculty member at the University of Rochester.

References

Akey, T. M. (2006). *School context, student attitudes and behavior, and academic achievement: An Exploratory analysis*. New York City: MDRC. Retrieved from: http://www.mdrc.org/sites/default/files/full_519.pdf

Appleton, J. J., Christenson, S. L., Kim, D., & Reschly, A. L. (2006). Measuring cognitive and psychological engagement: Validation of the Student Engagement Instrument. *Journal of School Psychology*, *44*(5), 427–445. doi:10.1016/j.jsp.2006.04.002

Benware, C. & Deci, E. L. (1984). Quality of learning with an active versus passive motivational set. *American Educational Research Journal*, *21*(4), 755–765. doi: 10.2307/1162999

Bill & Melinda Gates Foundation (2012). Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains. Retrieved from: http://www.metproject.org/downloads/MET_Gathering_Feedback_Research_Paper.pdf

Black, A. E., & Deci, E. L. (2000). The effects of student self-regulation and instructor autonomy support on learning in a college-level natural science course: A self-determination theory perspective. *Science Education*, *84*(6), 740–756. doi: 10.1002/1098-237X

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook Cognitive domain*. New York, NY. David McKay.

Broom, J. (2012, November). Building system capacity to evaluate and improve teaching quality: A technical assistance providers' perspective. In J. P. Connell (Chair), *A developmental approach to improving teaching quality: Integrating teacher evaluation and instructional improvement*. Symposium conducted at the meeting of the Association for Public Policy Analysis & Management, Baltimore MD.

Brophy, J. (2008). Developing students' appreciation for what is taught in school. *Educational Psychologist*, *43*(3), 132–141. doi: 10.1080/00461520701756511

Brown, R. S., & Coughlin, E. (2007). *The predictive validity of selected benchmark assessments used in the mid-Atlantic region*. (REL 2007-No. 017). Washington DC: National Center for Educational Evaluation and Regional Assistance Institute of Educational Sciences, U.S. Department of Education. Retrieved from: http://ies.ed.gov/ncee/edlabs/regions/midatlantic/pdf/REL_2007017.pdf

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. (2014). Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin*. doi: 10.1037/a0035661

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in Psychology. *Psychological Assessment*, *6*(4), 284–290. doi: 10.1037/1040-3590.6.4.284

Connell, J. P. & Broom, J. (2004). *The toughest nut to crack: First Things First's (FTF) Approach to improving teaching and learning*. Retrieved from: http://www.irre.org/sites/default/files/publication_pdfs/The%20Toughest%20Nut%20to%20Crack.pdf

Connell, J. P., Spencer, M. B., & Aber, J. L. (1994). Educational risk and resilience in African-American youth: Context, self, action, and outcomes in school. *Child Development*, *65*(2), 493–506. doi: 10.2307/1131398

Csikszentmihalyi, M. (1975). *Beyond boredom and anxiety*. San Francisco: Jossey-Bass.

Danner, F. W., & Lonky, E. (1981). A cognitive-developmental approach to the effects of rewards on intrinsic motivation. *Child Development*, *52*, 1043–1052.

Deci, E. L. (1975). *Intrinsic motivation*. New York: Plenum.

Deci, E. L., Schwartz, A. J., Sheinman, L., & Ryan, R. M. (1981). An instrument to assess adults' orientations toward control versus autonomy with children: Reflections on intrinsic motivation and perceived competence. *Journal of Educational Psychology*, *73*(5), 642–650. doi: 10.1037/0022-0663.73.5.642

Downey, C. J., Steffy, B. E., English, F. W., Frase, L. E., & Poston, W. K. (2004). *The three-minute classroom walkthrough: Changing school supervisory practice one teacher at a time*. Thousand Oaks, CA: Corwin Press.

Finn, J. D., & Rock, D. A. (1997). Academic success among students at risk for school failure. *Journal of Applied Psychology*, *82*, 221–234. doi: 10.2307/1170412

Fredricks, J. A. (2011). Engagement in schools and out-of-school contexts: A multidimensional view of engagement. *Theory Into Practice*, *50*(4), 327–335. doi: 10.1080/00405841.2011.607401

Gambone, M. A., Klem, A. M., Summers, J. A., Akey, T. A., & Sipe, C. L. (2004). *Turning the tide: The achievements of the First Things First education reform in the Kansas City, Kansas Public School District*. Philadelphia: Youth Development Strategies, Inc.

Grolnick, W. S., & Ryan, R. M. (1987). Autonomy in children's learning: An experimental and individual difference investigation. *Journal of Personality and Social Psychology*, *52*(5), 890–898. doi: 10.1037/0022-3514.52.5.890

Grolnick, W. S., & Ryan, R. M. (1989). Parent styles associated with children's self-regulation and competence in school. *Journal of Educational Psychology*, *81*(2), 143–154. doi: 10.1037/0022-0663.81.2.143

Harter, S. (1978). Pleasure derived from optimal challenge and the effects of extrinsic rewards on children's difficulty level choices. *Child Development*, *49*, 788–799.

Herman, J. L., Webb, N. M., & Zuniga, S. A. (2007). Measurement issues in the alignment of standards and assessments: A Case study. *Applied Measurement in Education*, *20*(1), 101–126. doi: 10.1207/s15324818ame2001_6

IRRE (2012). *Measuring what matters: Effective practices, engagement, alignment and rigor classroom visit protocol. Summary of professional development, technical supports and costs*. Toms River, NJ: Author. Retrieved from: http://www.irre.org/publications/measuring-what-matters-effective-practices-engagement-alignment-and-rigor-classroom-vis

Klein, A. (2012). Obama uses funding, executive muscle to make often-divisive agenda a reality. *Education Week*, *31*(35), 1–28.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of School Health*, *74*(7), 262–273. doi: 10.1111/j.1746-1561.2004.tb08283.x

Lee, V. L. & Burkam, D. T. (2003). Dropping out of high school: The role of school organization and structure. *American Educational Research Journal*, *40*(2), 353–393. doi: 10.3102/00028312040002353

Lee, V. L., Croninger, R. G., & Smith, J. B. (1997). Course-taking, equity, and mathematics learning: Testing the constrained curriculum hypothesis in U.S. secondary school. *Education Evaluation and Policy Analysis*, *19*(2), 99–121.

McGraw, K. O. (1978). The detrimental effects of reward on performance: A literature review and a prediction model. In M. R. Lepper & D. Greene (Eds.), *The hidden costs of reward* (pp. 33–60). Hillsdale, NJ: Erlbaum.

National Research Council and the Institute of Medicine. (2004). *Engaging Schools: Fostering High School Students' Motivation to Learn*. Committee on Increasing High School Students' Engagement and Motivation to Learn. Board on children, Youth, and Families, Division of Behavioral and Social Sciences and Education. Washington, DC: National Academies Press.

Patrick, B. C. (1995). *College students' intrinsic motivation as a function of instructor enthusiasm*. Unpublished Doctoral Dissertation, University of Rochester.

Pearson, P. L. C. (2010). *Arizona's Instrument to Measure Standards. 2010 Technical Report*. Retrieved from: http://www.azed.gov/standards-development-assessment/files/2011/12/aimstechreport2010.pdf

Polikoff, M. S. (2012). Instructional alignment under No Child Left Behind. *American Journal of Education*, *118*(3), 341–368. doi:10.1086/664773

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher*, *31*(7), 3–14. doi: 10.3102/0013189X031007003

Protheroe, N. (2009). Using classroom walkthroughs to improve instruction. *Principal*, March/April, 2009.

Raudenbush, S. W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage.

Rothman, R. (2012). Laying a common foundation for success. *Phi Delta Kappan*, *94*(3), 57–61.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. doi:10.1037/0003-066X.55.1.68

Shernoff, D. J., Csikszentmihalyi, M., Schneider, B., & Shernoff, E. S. (2003). Student engagement in high schools from the perspective of flow theory. *School Psychology Quarterly*, *18*(2), 158–176. doi: 10.1521/scpq.18.2.158.21860

Smith, C., & Hohmann, C. (2005). *The youth program quality assessment validation study: Findings for instrument validation*. Ypsilanti, MI: High/Scope Educational Research Foundation. Retrieved from: http://highscope.org/file/EducationalPrograms/Adolescent/ResearchEvidence/WebFinalYouthPQA TechReport.pdf

Webb, N. M., Herman, J. L., & Webb, N. L. (2007). Alignment of mathematics' state-level standards and assessments: The role of reviewer agreement. *Educational Measurement: Issues and Practice*, *26*(2), 17–29. doi: 10.1111/j.1745-3992.2007.00091.x

Williamson, R. & Blackburn, B. R. (2010). 4 myths about rigor in the classroom. *Eye on Education*, Larchmont, NY. Retrieved from: http://static.pdesas.org/content/documents/M1-Slide_21_4_Myths_of_Rigor.pdf