ISSN 2278-7690

Measuring Teachers' Instruction with Multilevel Item Response Theory

Ben Kelcey
University of Cincinnati, 2840 Bearcat Way Cincinnati, Ohio, United States
ben.kelcey@gmail.com

ABSTRACT

The purpose of this study was to describe an approach for measuring teachers' uses of instruction as it relates to students' achievement through classroom observations. Despite significant work on the substantive content of observation systems chronicling teachers' instruction, literature has largely relied on simple counts of instructional features or the average of quality indicators to describe teachers' instruction. However, such coarse summaries generally do not reflect current theories of instruction, prior empirical evidence, and the framework of most observation systems. The approach presented in this paper builds on evidence that teachers' instruction varies across lessons and that instructional features or quality indicators do not necessarily contribute equally to our understanding of effective instruction. To align theory, data and methods, this study applied multilevel item response theory to the study of early literacy instruction as it relates to students' achievement. This model provided a more complex, but more precise and theoretically grounded, view of instruction by linking components of instruction theory to model parameters. Empirical results suggested that multilevel item response models encouraged precision in the specification of theory, data collection, and models that is absent in simpler models.

Indexing terms/Keywords

Multilevel item response theory, measurement invariance, multiple group measurement models, instruction, classroom observations, teacher knowledge, student achievement

Academic Discipline And Sub-Disciplines

Education

SUBJECT CLASSIFICATION

Applied statistics, psychometrics, measurement, education, teaching, learning

TYPE (METHOD/APPROACH)

Psychomteric analysis, observational study

Council for Innovative Research
Peer Review Research Publishing System
Journal: International Journal of Research in Education methodology
Vol. 7, No.1
ijremeditor@gmail.com
www.ijrem.com

Introduction: Valid measurement of teaching is essential to the advancement of quantitative research in education and understanding teacher effectiveness (Raudenbush & Sadoff, 2008). The recent emphasis on teacher effectiveness by federal education initiatives in the United States such as Race to the Top, district teacher evaluation, and compensation programs and funding agencies has underscored the centrality of teaching in advancing education (Measures of Effective Teaching, 2012). Although in principle valid assessment of the impact of teachers or educational interventions on students' cognitive development can be had without the direct measurement of teaching (e.g., value-added models), the results do little to explain the actions effective teachers take to produce student gains thereby eliminating inroads to teacher improvement. The measurement of teaching addresses critical questions involving the mechanisms through which teachers' practice shapes students' development and requires researchers to make precise their theories of teaching and interventions and helps them to mount specific and focused empirical investigations of these theories in falsifiable ways (Raudenbush & Sadoff, 2008).

There is a long history of quantitative assessments of classroom processes, however, this line of inquiry has largely been atheoretical and has produced inconsistent relationships among teacher, teaching and student variables (Hoffman, 1991). For instance, reviews of the literature on literacy instruction have suggested that research in this area has largely been based on theories of literacy (e.g., comprehension of texts) rather than theories of teaching literacy (Hoffman, 1991). Despite consensus that the nature and quality of teachers' instruction should impact their students' achievement, researchers are still far from understanding the characteristics of instruction that distinguish more and less effective teachers.

To this end, direct classroom observation has emerged as a key measurement strategy for understanding teacher effectiveness and has become a central feature in the advancement of teaching (Gitomer, 2009). Classroom observation, carried out for the purpose of studying instructional practices as they relate to students' achievement, offers a promising way to unpack the black box of teaching and arrive at a deeper understanding of effective practice. Recent studies have developed a number of classroom observational systems which measure teaching by recording or rating the quality of features or strategies thought to reflect critical dimensions of effective teaching (e.g., Pianta & Hamre, 2009). Targeted features are thought to be instances of underlying and unifying teacher quality constructs rather than single disjoint features.

Although much attention has been placed on the content development of classroom observation systems, the measurement of the data they produce has received far less attention. Research capturing observed differences in instruction has predominantly relied on simple counts of measured features and has further tended to collapse these counts across multiple observations using simple averages (e.g., Cirino, Pollard-Durodola, Foorman, Carlson, & Francis, 2007). Yet, reviews of classroom observation research have noted that describing observed instruction using simple counts and averages is largely atheoretical and has produced inconsistent relationships among teacher, teaching and learning variables (Hoffman, 1991).

Within most theoretical frameworks guiding these observations systems, instruction is framed as more than simple counts of enacted features. Effective instruction requires skillful blends of strategies, such that teachers' instructional strategies and actions build on each other, are designed for a specific purpose, and are responsive to various contextual factors (McGhie, Underwood, & Jordan, 2007; Pacheco, 2009). Such orchestration suggests that the patterns of instructional features hold information beyond their simple frequencies of use.

Comprehensive theories of teaching also suggest that instruction is a dynamic process that varies intentionally from one lesson to the next and that understanding this variation is central to understanding and improving the quality of teachers' instruction (Pacheco, 2009; Rimm-Kaufman & Hamre, 2010). Theoretical frameworks for instruction have delineated lesson to lesson variation in instruction within a teacher to understand how and why a teacher establishes a structure or pattern across all lessons and why, in any given lesson, a teacher's current choice of strategies may deviate from his/her established pattern (Soar & Soar, 1979). For example, research has indicated that teachers' vary in their use of instructional strategies depending on the purpose of a lesson (Stodolsky, 1990).

The dynamic nature of instruction suggests that the patterns of instructional strategies teachers' choose in a specific lesson and the persistence use of these strategies across lessons potentially provide significant insight into the teaching profiles of more and less effective teachers (Carlisle, Kelcey, & Berebitsky, 2013). However, empirical analyses of instruction have largely stood in opposition to such theoretical frameworks because they have typically collapsed indices of instruction across observations ignoring within teacher variability. This approach leaves us with no basis for understanding the dynamics and variability of instruction because within-teacher variation is lost. The problem, then, is finding approaches to capture both within teacher variation and teachers' stable preferences for certain teaching strategies.

Purpose

The purpose of this study was to improve the measurement of instruction as it relates to students' achievement by applying analytical methods better suited to the structure of classroom observations and the theory of instruction in early elementary literacy. Of particular relevance in this study is how teachers' choices of instructional strategies within and across lessons and how these choices associate with their students' achievement. As a preliminary investigation this paper applies multilevel item response theory to the measurement of teachers' engagement in teacher-directed instruction (e.g., coaching) in the area of early literacy.

Teacher-directed instruction or active instruction describes teachers' use of instructional strategies and actions that promote literacy concepts and development in ways that require active processing by students. Teacher-directed instruction characterizes the nature in which teachers deliver literacy content to students and is described by the strategies teachers take to ensure effective learning and practice of literacy skills (Seidel & Shavelson, 2007). In previous work synthesizing research in literacy instruction, teacher-directed instruction has been highlighted as a key component of effective instruction. In particular, instructional regimes that skillfully blend teacher-directed strategies such as modeling or asking evaluation questions are thought to engage students in higher-level thinking so that they make connections between new and prior knowledge (Taylor, Pearson, Peterson, & Rodriguez, 2003). This study looks, in part, to extend this line of inquiry by measuring teachers' uses of teacher-directed instruction in early elementary literacy lessons as they relate to students' literacy achievement.

The paper addresses seven research questions: (1) Do teachers’ uses of instructional features contribute equally to their instructional profiles in early elementary literacy lessons? (2) To what extent is the variation in teachers' instruction attributable to persistent differences among teachers versus differences among lessons within teachers? (3) To what extent are instructional strategies invariant across grades two and three? (4) To what extent does the topic of a lesson associate with teachers' lesson-specific deviations in teacher-directed instruction? (5) To what extent does teachers' knowledge associate with teachers' persistent levels of teacher-directed instruction? (6) To what extent does teachers' stable engagement in teacher-directed instructional strategies explain students' achievement gains in literacy? (7) To what extent do multilevel item response models improve our ability to test theories about teachers' choices of instructional strategies over simpler methods? If we are to develop a science of effective teaching, we need measurement tools that allow us to precisely test our theories of teaching. To this end, the utility of multilevel item response models is examined through the lens of being able to precisely test features underlying theories of effective instruction as they relate to students' achievement.

Below, the paper first outlines the sample used in this study and then describes the classroom observation system and measures used in the study. The paper goes on to explore the application of multilevel item response models to the measurement of teacher-directed instruction. In turn, the paper finishes by examining links between teacher-directed instruction and students' literacy achievement.

Method

Sample

The sample used in this study was drawn from classrooms which were in Reading First schools in the Midwest region of the United States; to qualify for funding, participating school districts met criteria for high levels of poverty, and the districts usually selected the schools with high poverty and low literacy achievement to participate in the Reading First program. The work reported in this study focused on a subpopulation of this original group. In particular, the sample consisted of 1638 lessons taught in 87 early elementary classrooms drawn from 19 different schools across 6 districts. Of the 87 teachers who participated in the study, 44 taught second grade and 43 taught third grade; 19% were non-White, 11% had a master's degree in reading, and the average years of experience was 13. Classrooms had on average 23 students, of which roughly 45 percent were minority, 21 percent were in special education, and over three quarters were eligible for free/reduced lunch.

Measures

Instruction. A major issue in the study of instruction is clarifying the mechanisms through which effective instruction operates. As a result, a formative step in measuring instruction is identifying observable features that are theoretically reflective of a dimension and can be reliably assessed through observations. Our conceptualization draws heavily on earlier work describing the salience of instructional strategies teachers take to provide effective instruction (Pressley et al., 2003). Prior research has provided support for links between student cognitive development and teachers' use of instructional strategies that promote active student engagement (Taylor, Pearson, Peterson, & Rodriguez, 2003; Seidel & Shavelson, 2007).

To describe the extent to which teachers engaged in forms of teacher-directed instruction within literacy lessons, four instructional strategies were identified on the basis of a review of the literature (Seidel & Shavelson, 2007; Taylor, Pearson, Peterson, & Rodriguez, 2003). The measured instructional strategies included the following: telling, modeling/coaching, asking questions for evaluation, and providing practice or review activities. The selected strategies were by no means exhaustive in describing teacher-directed instruction or in describing effective teaching as a whole. There are many other qualities of effective teaching and the literature has suggested that literacy is achieved from a combination of teacher-directed instruction and scaffolded opportunities. Rather, measured instructional strategies were chosen to emphasize the structure and delivery of teaching and learning and represent common strategies teachers' use to place differing levels of cognitive demand on students (Seidel & Shavelson, 2007).

Although extant literature has demonstrated evidence supporting the implications of individual instructional strategies, there is a need for clearer articulations of higher-level theories of teaching and teacher-directed instruction (Douglas, 2009). For instance, theory suggests that effective instruction involves use of high and low cognitive demand instructional strategies but there is little empirical evidence supporting such combinations with teacher-directed instruction (Pressley et al., 2001). For these reasons, there is a need for methods that provide empirical evidence toward how strategies differentially relate to the teacher-directed instruction construct.

Although prior research has tied these four strategies together and indicated their value, the same research has also suggested that they differ in the amount of cognitive demand they place on students. For instance, modeling and asking questions for evaluation are thought to engage teachers and students interactively in higher-level thinking about text so that students actively make deep connections with its meaning and their prior knowledge (Taylor, Pearson, Peterson, & Rodriguez, 2003). As a result, these types of interactions tend to place increased cognitive demand on students. In contrast, evidence has suggested that simply telling students information requires less cognitive demand because it requires lower levels of active participation and processing by students (Taylor, Pearson, Peterson, & Rodriguez, 2003). Accordingly, theory suggests that strategies form a certain involvedness scale whereby use of more cognitive demanding strategies requires skillful levels of teacher-directed instruction.

To carry out the study, observations were conducted using the Automated Classroom Observation System for Reading (Carlisle, Kelcey, Berebitsky, & Phelps, 2011). This system was specifically designed to have a narrow focus on the delivery and structure of early literacy instruction. Observations of classroom instruction were coded in real time using a tablet computer by trained observers. Observers underwent multiple training sessions and practice visits to classrooms and final inter-rater agreement across all possible fields and combinations exceeded 87%. To capture teachers' uses of teacher-directed instruction observers recorded teachers' dichotomous use of each of targeted instructional strategy for each observed lesson.

Teachers' Reading Knowledge. Research in the area of early literacy has argued that teachers' knowledge about teaching language and literacy is a critical factor in the quality of their instruction (Snow, Griffin & Burns, 2005). This literature has argued that this knowledge must be anchored in not only language and literacy knowledge but also in an understanding of how students' develop literacy skills (Moats, 2009; Snow, Griffin & Burns, 2005). Similar research in other areas, such as mathematics teaching, has progressively demonstrated that this type of knowledge guides teachers' instruction in ways that make them more effective (Hill et al., 2008). However, the evidence identifying a relationship between teachers' knowledge and practice in literacy has been less consistent than in other substantive areas (Moats, 2009). This paper advances this line of inquiry by drawing on a new generation literacy knowledge measure designed to assess the types of content problems that teachers encounter in practice.

The measure of knowledge used in this study was the Teachers’ Knowledge of Reading and Reading Practices (TKRRP; Carlisle, Kelcey, Rowan, & Phelps, 2011; Kelcey, 2011). The measure consisted of 22 items focused on focused on oral language, reading, and writing activities in the domains of word reading (e.g., phonemic awareness, letter sound relationships) and comprehension (e.g., fluency) and student learning situated in classroom practices. The items assessed both academic and pedagogical knowledge by situating items within classroom and student scenarios that early elementary teachers encounter. Previous psychometric work on the measure indicated that its item response theory reliability was approximately 0.76 (Carlisle, Kelcey, Rowan, & Phelps, 2011). Teachers in this sample were scored as part of the larger Michigan Reading First sample using over 1000 teachers. Scores were constructed using a two-parameter logistic item response model with a mean of zero and standard deviation of one.

Achievement. As measures of current and prior student achievement, analyses drew on two of the Iowa Tests of Basic Skills (ITBS) reading subtests. The first was the reading comprehension standard score which requires students to select responses to questions that follow short passages. The second subtest was the vocabulary measure. This measure assesses students' breadth of vocabulary by having them relate spoken words to pictures. Test reliability for both of these subtests in both grades exceeds 0.85 (Hoover, Dunbar, & Frisbe, 2003).

Analytic Approach

Model for Instruction. To measure instruction, data were analyzed using a multilevel item response model (Wang & Wilson, 2005). Let *Y_{ijk}=1* if teacher *k* employs strategy *i* during lesson *j* and *Y_{ijk}=0* if not so that

$$\log[\frac{P(Y_{ijk}=1\,|\,\theta_k,\gamma_{jk})}{1-P(Y_{ijk}=1\,|\,\theta_k,\gamma_{jk})}]=a_i(\theta_k+\gamma_{jk}-b_i) \qquad (1)$$

where a_i, is the discrimination parameter for action *i*, θ_k is teacher *k*'s stable level of teacher-directed instruction across all lessons, γ_{jk} is teacher *k*'s lesson-specific deviation for lesson *j*, and b_i is the difficulty parameter for action *i*. Both θ_k and γ_{jk} were assumed to have a normal distribution centered at zero and the scale of γ_{jk} was fixed to have a variance of one. Maximum marginal likelihood estimates were obtained using numerical integration with 20 rectangular quadrature points per dimension.

Under this representation, the measured instructional strategies are viewed as items, with the a_i describing the strength of the strategy-construct relationships and b_i describing strategy *i*'s required cognitive effort or difficulty. θ_k are viewed as each teacher's persistent level of engagement with teacher-directed instructional strategies across all lessons and are estimated using the between teacher variation. In contrast, γ_{jk} are seen as lesson-specific deviations and are estimated using the within teacher lesson variation. The stability of teachers' instructional choices can then be summarized using the intraclass correlation (ICC)

$$ICC = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_\gamma^2} \qquad (2)$$

where σ_θ^2 is the teacher level variance and σ_γ^2 is the lesson level variance.

Valid comparisons of teachers' use of teacher-directed instructional strategies necessitate the properties of the measured strategies to be invariant across subgroups so as to form a common measurement scale (Kelcey & Carlisle, 2013). An important feature of these data is that they are split across grades two and three. Although prior literature has identified the value of our measured strategies as they relate to students' achievement within several early elementary grades, it has not directly compared teachers' use of teacher-directed instruction across grades and has not assessed the extent to which these strategies function similarly by grade. To this end, the invariance of empirical relations between strategies and the latent dimension across grades was examined by contrasting the fits of the model in equation (1) with a model that allowed strategy (item) parameters to differ by grade

$$\log[\frac{P(Y_{ijk}^g = 1 | g, \theta_k^g, \gamma_{jk}^g)}{1 - P(Y_{ijk}^g = 1 | g, \theta_k^g, \gamma_{jk}^g)}] = a_i^g(\theta_k^g + \gamma_{jk}^g - b_i^g) \; (3)$$

where *g* indicates the grade level of teachers' classrooms and remaining notation is the same as in equation (1).

Comprehensive theories of teaching also describe effective instruction as a process that is responsive to the purpose of a lesson and informed by the knowledge a teacher brings to the classroom (Pacheco, 2009; Rimm-Kaufman & Hamre, 2010). One feature potentially relevant to lesson variation in early literacy instruction is the topic of a lesson (Kelcey & Carlisle, 2013). Early literacy instruction is generally thought to include five main topics: phonics, fluency, writing, comprehension and vocabulary (National Reading Panel, 2000). As an exploratory analysis, the extent to which the variation in teachers' use of teacher-directed instructional strategies across lesson is, in part, explained by lesson topic was investigated by linking equation (1) with an explanatory component using latent regression

$$\gamma_{jk} = \beta_{0k} + \sum_{s=1}^{S} \beta_s X_{sjk} + u_{jk} \qquad (4)$$

where X_{sjk} represents four indicator variables for five lesson topics for lesson *j* in teacher *k* with β_s as the coefficient for covariate *s,* and u_{jk} represent the adjusted lesson deviations for lesson *j* in teacher *k* conditional upon *X*. Similarly, the association between teachers' literacy knowledge and their uses of teacher-directed instruction was assessed by expand equation (1) so that

$$\theta_k = \pi_{00} + \pi_T T_k + r_k \qquad (5)$$

where T_k represents the knowledge level of teacher *k* with coefficient π_T and r_k represent the adjusted teacher levels of teacher-directed instruction for teacher *k* conditional upon grade. In formulating the analysis as a multilevel item response model, the approach potentially provides specific ways to test the mechanisms of our theory. We can first assess the extent to which patterns of teachers' choices of instructional strategies hold important information by evaluating the equality of the discrimination parameters, a_i, across strategies. Disparities among the discrimination parameters suggest that strategies differentially reflect teacher-directed instruction and that patterns of the strategies hold information beyond simple counts of strategies. Similarly, using the difficulty parameters, we can empirically examine the extent to which teachers' use of strategies align with the theoretical expectations concerning the cognitive demand strategies require. If the difficulty parameters, b_i, are related to the cognitive demand strategies require, results should demonstrate that strategies requiring higher cognitive demand are more difficult.

Third, we can describe the stability of teachers' uses of teacher-directed instruction by examining the intraclass correlation coefficient. Fourth, we can examine whether we can place grade two and three classrooms on a common scale measuring teacher-directed instruction using the given strategies. Finally, through the latent regression we are able to explore the extent to which teachers' use of teacher-directed instruction varies as a function of lesson topic and their knowledge.

Model for Achievement. To understand the value of teacher-directed instruction, the associations between teachers' regular uses of teacher-directed instruction and their students' achievement gains in reading comprehension and vocabulary were assessed. Student achievement for each subtest outcome was modeled using separate hierarchical linear models (Raudenbush & Bryk, 2002) with adjustments for students' prior achievement on both subtests such that

$$Y_{jk} = \pi_{0k} + \sum_{p=1}^{n=2} \pi_p X_{pjk} + \varepsilon_{jk} \qquad (6)$$

where Y_{jk} is the current ITBS-comprehension or vocabulary subtest score for student *i* in classroom *j*, π_{0k} is the average student score adjusted for the ITBS prior achievement subtest scores in comprehension and vocabulary, *X*, and π_p are the corresponding coefficients for the prior achievement variables while ε_{jk} has a normal distribution with mean zero and variance σ^2. At level two, the adjusted average achievement, π_{0k}, was modeled as a function of the expected a posteriori estimates of teachers' stable use of teacher-directed instruction derived from the measurement model in equation (1)

$$\pi_{0k} = \beta_{00} + \beta_{01}\theta_k + r_{0k} \qquad (7)$$

where β_{00} is the average adjusted achievement level, and β_{01} is the association between teachers' increased use of teacher-directed instruction (θ_k) and achievement. Finally, r_{0k} is the random effect of teacher *k* and has a normal distribution with mean zero and variance τ.

Results

The results from the two-parameter multilevel item response model first suggested that the measured instructional strategies differed in terms of their difficulty (Table 1). The most difficult action was modeling/coaching, which was followed by providing practice or review activities and telling and the easiest was asking evaluation questions. Similarly, the results demonstrated that the discrimination parameters among the strategies differed (Table 1). For instance, the instructional strategy of 'coaching/modeling' loaded more heavily on the teacher-directed instruction dimension than did the 'asking evaluation questions' strategy. The disparity in discrimination parameters suggests that strategies relate differently to teacher-directed instruction in ways that may privilege patterns over counts. At the same time, the low discrimination value for the asking evaluation may question the fit of this strategy to teacher-directed instruction.

Table 1

Instructional strategy (item) parameters for teacher-directed instruction

Instructional Strategy	Proportion of lessons strategy was used	χ^2 Test of Item Fit	Difficulty (SE)	Discrimination (SE)
Telling	0.69	3.20	-1.15(0.10)	0.86(0.10)
Modeling/ Coaching	0.59	2.70	-0.33(0.09)	2.81(0.46)
Asking questions for evaluation	0.70	3.12	-1.43(0.11)	0.68(0.09)
Providing practice/review activities	0.72	2.58	-1.04(0.10)	1.23(0.13)

To more formally assess the extent to which strategies differentially reflected the underlying latent trait, the fits of a one- and two-parameter multilevel item response models were contrasted by holding the discrimination parameters, a_i, constant across items in equation (1). The likelihood ratio test (LRT) and information criteria indicated the two-parameter model outperformed the one-parameter (Table 2) suggesting that strategies discriminate differently.

The second research question investigated the stability of teachers' instructional choices across the observed lessons. The results of the variance decomposition indicated a strong dependency among teachers' choices of instructional strategies within the same lesson. About 65% of the variation in observed instruction was attributable to differences among lessons within a teacher while only 35% was attributable to stable differences among teachers.

Viewed from a perspective of local item/strategy dependence, this variance decomposition suggested substantial local dependence of strategies within lessons. Comparisons of the model in equation (1) with a model which constrained γ_{jk} =0 in equation (1) also indicated that the model with lesson-specific effects fit the data better (Table 2). Similar results were found in assessing the Q_3 statistic (Yen, 1984). The resulting Q_3 statistic for equation (1) was 0.08 from the expected value whereas the Q_3 statistic was 0.40 when dropping γ_{jk} in equation (1). Such dependence suggests simpler models ignoring the fact that teachers' choices of strategies are situated within lessons (i.e., assuming γ_{jk} =0 in equation (1)), may violate the local item independence assumption (Yen, 1984). Additional item-fit assessments also suggested that the model fit well (Table 1).

Table 2

Model fit indices

Fit Index	Multilevel 1pl	Multilevel 2pl with $\gamma_{jk}=0$	Multilevel 2pl	Multilevel multigroup 2pl
-2LL	7588	7688	7523	7522
AIC	7600	7704	7541	7556
BIC	7632	7746	7589	7647
Number of parameters	6	8	9	17

Note: -2LL refers to negative two times the log-likelihood, AIC refers to the Akaike information criterion and BIC refers to the Bayesian information criterion.

The third research examined the extent to teachers' uses of teacher-directed instruction in grades two and three could be put on a common scale given the targeted strategies. The results contrasting multigroup multilevel item response model with that of a single group model suggested that the fit of the single group model was not improved upon by allowing grade two and three to have different item/strategy parameters (Table 2). In each instance, parameter estimates for strategies resulting from the multigroup model were nearly identical across grades.

Following evidence that teachers' uses of teacher-directed instruction can be placed on a common scale in these grades and that instruction varies both within and across teachers, the fourth and fifth research questions investigated the extent to which this variance was related to systematic differences in lesson topic and teachers' literacy knowledge. Latent regression results demonstrated that teachers' tended to use teacher-directed instruction similarly in phonics, writing, comprehension and vocabulary lessons but tended to use it significantly less than in fluency lessons (Table 3). Finally, the results indicated that teachers' levels of literacy knowledge were not significantly associated with increased levels of teacher-directed instruction.

Table 3

Latent regression coefficients explaining lesson and teacher level variation in teacher-directed instruction

Covariate	Coefficient (SE)
Fluency	0.00 (0.00)
Phonics	1.28* (0.19)
Vocabulary	1.42* (0.20)
Writing	1.40* (0.18)
Comprehension	1.51* (0.20)
Teacher knowledge	0.02 (0.12)

*p<0.05

Relation of Instruction to Achievement

The sixth research question focused on the extent to which teachers' uses of teacher-directed instruction were associated with students’ achievement. After adjusting for baseline differences through the two measures of prior achievement, results suggested that teachers’ uses of teacher-directed instruction were significantly and positively associated with both subtest outcomes (Table 4). For example, students in classrooms with teachers who tended to use teacher-directed instruction one standard deviation above average tended to achieve about 0.06 standard deviations more on the ITBS comprehension subtest.

Table 4

Standardized regression coefficients (and standard errors) for students' achievement

	Comprehension	Vocabulary
Intercept	0.00(0.02)	0.00(0.03)
ITBS-comprehension pretest	0.51*(0.02)	0.54*(0.02)
ITBS-vocabulary pretest	0.28*(0.02)	0.20*(0.02)
TDI	0.05*(0.02)	0.05*(0.02)

*p<0.05

Note: TDI is short for teacher-directed instruction

Comparison to a More Conventional Approach

The complexity added by an item response model and its multilevel formulation raises questions of whether conventional methods which ignore the aforementioned features provide similar conclusions regarding the relationship between teacher-directed instruction and students' achievement. To examine differences, the multilevel item response model was contrasted with the average number of strategies a teacher used in his/her lessons

$$\bar{Y}_k = \frac{1}{J}\sum_{j=1}^{J}\sum_{i=1}^{I} Y_{ijk} \qquad (8)$$

where Y_{ijk} was one if strategy *i* was used in lesson *j* for teacher *k*. Comparative results suggested that in both cases the standardized coefficient and its corresponding *t*-ratio were smaller when using the simple averages. Moreover, for both of the subtests considered, the relationship between the averages and the students' achievement was statistically insignificant whereas indices from the multilevel item response model were significantly correlated to each subtest.

Table 5

Comparison of standardized regression coefficients for students' achievement

	Comprehension		Vocabulary	
	Standardized coefficient (Standard Error)	*t*-ratio	Standardized coefficient (Standard Error)	*t*- ratio
Averages	0.03 (0.02)	1.55	0.04 (0.02)	1.70
Multilevel IRT	0.05* (0.02)	2.39	0.05* (0.02)	2.23

**p*<0.05

Note: *t*-ratio is the ratio of the standardized coefficient to the standard error

Discussion

The current study drew attention to several general issues in the measurement of instruction and put forth a proposition; in order to advance our understanding of effective instruction, we need to align theories of instruction, data collection methods and analytic approaches in ways that allow us to systematically investigate theories of teaching. By aligning theory, data and methods we are able to systematically test our theories of instruction to advance and develop a science of teaching. In line with this proposition, the paper proposed multilevel item response models as a flexible methodological tool to align theory, data collection and methods in the measurement of instruction.

The paper offered an initial investigation into this proposition by studying the measurement of teacher-directed instruction in literacy lessons as it relates to students' achievement. The application of a multilevel item response model provided a more complex, but more precise and theoretically grounded, view of instruction. Specifically, the current investigation set out to examine four commonly accepted theories of literacy instruction for which there is little empirical evidence supporting. These theories suggested that strategies differ in complexity, that patterns of strategies matter, that teachers' uses of these strategies varied by lesson and this variation was associated with lesson and teacher features, and that teachers' engagement in teacher-directed instruction would promote student achievement.

To test the merits of these theories, the paper applied a multilevel item response model. This model provided a more complex, but more precise and theoretically grounded, view of instruction by linking the aforementioned theories to model parameters. The results largely supported these hypotheses. Results suggested that strategies differed in their difficulties and that patterns held information beyond simple counts, and teacher-directed instruction varied considerably from one lesson to the next. The results also suggested that the degree to which teachers' used of teacher-directed instruction was fairly similar in most topics but was less in fluency lessons. Further, contrary to expectation, the results indicated that teachers' knowledge was not significantly associated with teachers' uses of teacher-directed instruction perhaps suggesting that the type of knowledge captured by this measured was not related to this dimension of instruction. Finally, results indicated that teacher-directed instruction was associated with students' achievement gains but that these gains were relatively small.

In what follows, the practical value of this approach in improving theories of instruction, classroom observation instruments and the practice of measuring teaching as it relates to students' achievement is discussed. A first practical utility of applying item response theory to the measurement of instruction is the careful assessment of the properties of measured instructional strategies. Theories of instruction frequently suggest that instructional strategies or features differ in the difficulty and that patterns of instruction may be more meaningful than their sum. A key function of item response models is the ability to carefully assess the difficulty and discrimination of items/strategies. Applied to the measurement of instruction, the practical value of such assessments allow is twofold. First, item parameters allow us to formally test theories of instruction. For instance, although there is general consensus that simple counts of the number of strategies teachers' use are insufficient in describing teaching, empirical analyses have not advanced along with this theory (Hoffman, 1991). Second, and more pragmatically, item parameters provide substantial insight to researchers in refining their classroom observation instruments to improve the measurement of instruction (e.g., measure new strategies). For example, consider the asking questions for evaluation strategy in the current investigation. Both the unexpected easiness of this strategy and its weak relationship with the teacher-directed instruction construct might suggest that this strategy did function as hypothesized and that it provides relatively little information about teachers' uses of teacher-directed instruction. As a result, subsequent investigations might consider replacing this strategy. Without the careful assessment of how the measured strategies relate to the construct of interest (e.g., through averages), this information would be lost.

Second, multilevel item response models add value by allowing researchers to decompose the variation in observed instruction into stable differences across teachers versus variation within a teacher across lessons (Carlisle, Kelcey, & Berebitsky, 2013). Theoretically, this decomposition adds value because it helps us to understand the nature and variability of teachers' instruction and the extent to which teacher effects are potentially due to stable differences among teachers. More practically, this decomposition can be used to test theories linking instruction with key lesson and teacher variables. In the current example, consider the topic of a lesson. It was hypothesized that teachers' uses of teacher-directed instruction would vary by topic. However, our results, suggested that the within teacher variation in instruction was not largely influenced by lesson topic with the exception of fluency lessons. Further theoretical investigation on the potential differences between fluency and other lessons suggests that these differences might be due to the highly prescribed nature of fluency lessons. Based on these results, subsequent studies might examine not only the topic of the lesson but also the extent to which specific lessons are formulaic. The important point is that collapsing across lessons would have obscured this information.

Third, use of item response models can help us understand the extent to which we can place teachers' instruction across multiple grades on a common measurement scale (Kelcey & Carlisle, 2013). Theoretically, such invariance facilitates comparisons among groups and correlations with external variables because it ensures that differences are not an artifact of measurement non-invariance. More practically, such invariance investigations help to understand the stability of instructional strategies as they relate to theoretical dimensions of teaching across grades. Again, an important point is that analyses which draw upon simple averages to describe the latent construct miss this information and potentially draw misleading conclusions

Finally, applying multilevel item response models to the study of instruction has the potential to better differentiate among levels of the latent instructional trait. As a result, we can better differentiate among teachers and are more likely to likely to detect relationships between teachers' instruction and external variables if they exist. In current investigation, for example, analyses were better able to detect relations between teachers' stable levels of teacher-directed instruction and student's achievement as compared to a more conventional approach. Such increased capacity may be particularly important for studies of instruction because, to date, studies of instruction have largely found relatively small effects.

Although measurement of instruction with the current application of a multilevel item response model showed promise, the study and the application were not without drawbacks. Notably, the current study considered only a single dimension of instruction and did so only within the context of early literacy. More generally, there are also some uncertainties as to the extent to which the measurement of teaching is amenable to the theoretical assumptions underlying item response theory. For instance, applications of item response theory assume that levels of the targeted construct are meaningfully associated with monotonic changes in the probability of employing items. This assumption was operationalized in the current application by outlining a theory that the measured strategies gave rise to a certain cognitive demand or involvedness scale whereby use of more cognitive demanding strategies required higher teacher levels of teacher-directed instruction. The amenability of teaching to the assumptions of item response theory will, in general, be dependent on the specific teaching theories and dimensions being studied.

In addition, there are also questions concerning estimation methods. Conventionally, dependable item response models require a substantial number of respondents because maximum likelihood estimates of two-parameter models are known to be downwardly biased in small samples. Most studies of teaching, however, tend to proceed with small sample sizes because of the steep cost associated with observing teachers multiple times across a year. Even in the largest of studies, samples rarely reach 500 teachers. However, a key difference in conventional applications of item response methods (e.g., to student achievement) and the application to the measurement of teaching is that teachers are generally measured multiple times. Because the current implementation (i.e., equation (1)) forces the teacher dimension, θ_k, and lesson dimension, γ_{jk}, to share the same discrimination parameter, a_i, item/strategy parameters are estimated using both lessons and teachers. For example, the current application estimated item/action parameters using over 1500 lessons (but only 87 teachers). In this way, the nesting of items within lessons and lessons within teachers conceptually parallels the nesting of items within students and students within schools. The performance of item response methods under constraints salient to teaching is fundamental to understanding the amenability of item response methods to the measurement of teaching.

Similarly, while the current investigation studied teacher-directed instruction through only four strategies/items, more common classroom observation systems also tend to represent teaching domains through only a few items. For instance, each domain in the Classroom Assessment Scoring System (CLASS) (Pianta & Hamre, 2009) and Framework for Teaching (FFT) (Danielson, 2007) draw on only three to five items. Even collapsing items across the distinct domains within each system would only yield about ten items. The construction of scales using few items also raises the question of the trustworthiness of the information produced by the analyses (van den Berg, Glas & Boomsma, 2007).

Notwithstanding the limitations of the study, the outcomes of this study have important implications for the study of instruction. Principal among them is the value of aligning analytic methods, observation methods and the working theory of instruction (Kelcey, McGinn, & Hill, 2013). By bringing theory, data and methods into line we can make explicit the underpinnings of our theories and provide precise ways to test these theories. In turn, we are better suited to advance a theory of instruction, inform subsequent data collection regimes and uncover relationships between teachers' instruction and their students' achievement.

Multilevel item response models offer one promising way to align line theory, data and methods because they encourage precision in the specification of the theory, data collection, and models that is absent in simpler models. In attending to these features of teaching in more systematic and rigorous ways, we can form more robust theories of instruction, more reliable and valid observation scales and systems and draw more precise inferences. Each of these is critical to the development of a theoretically defined and empirically supported framework for effective instruction.

REFERENCES

1. Bartholomew, D. and Knott, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd ed. London: Arnold.
2. Carlisle, J., Kelcey, B., Berebitsky, D., & Phelps, G. (2011). Embracing the Complexity of Instruction: A Study of the Effects of Teachers' Instruction on Students' Reading Comprehension, *Scientific Studies of Reading,* 15, pp. 409-439.
3. Carlisle, J., Kelcey, B., Rowan, B., & Phelps, G. (2011). Teachers' Knowledge About Early Reading: Effects on Students' Gains in Reading Achievement, *Journal for Research on Educational Effectiveness*, 4, pp. 289-321.
4. Carlisle, J., Kelcey, B., & Berebitsky, D. (2013). Teachers' Support of Students' Vocabulary Learning During Literacy
   a. Instruction in High Poverty Elementary Schools. *American Educational Research Journal*, 50, 1360-1391.
5. Cirino, P. T., Pollard-Durodola, S. D., Foorman, B. R., Carlson. C. D., & Francis, D. J. (2007). Teacher characteristics, classroom instruction and student literacy and language outcomes in bilingual kindergartners. *Elementary School Journal, 107*, 341- 364.
6. Danielson, C. (2007). Enhancing professional practice: A framework for teaching (2nd edition). Alexandria, VA:
   a. Association for Supervision and Curriculum Development.
7. Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational*
   a. *Researcher, 38,* 518-521.
8. Hill, H., Blunk, M., Charalambous, C., Lewis, J., Phelps, G., Sleep, L. & Ball, D. (2008). Mathematical knowledge for
   a. teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction*, 26, pp. 430-511.
9. Hoffman, J. V. (1991). Teacher and school effects in learning to read. In R. Barr, M. L.
10. Hoffman, J. V., Sailors, M., Duffy, G. R., & Beretvas, S. N. (2004). The effective elementary classroom literacy environment: Examining the validity of the TEX-IN3 observation system. *Journal of Literacy Research, 36,* 303-334.
11. Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate Measurement Invariance in Cross-
    a. classified Rater-mediated Assessments. *Frontiers in Quantitative Psychology and*
    b. *Measurement*, 5, 1-13.
12. Kelcey, B. & Carlisle, J. (2013). Learning About Teachers' Literacy Instruction From
    a. Classroom Observations. *Reading Research Quarterly,* 48, 301-317.
13. *Measures of Effective Teaching*. (2012). Gathering feedback for teaching: Combining high- quality observations with student surveys and achievement gains. Retrieved from the Bill and Melinda Gates Foundation website: http://www.gatesfoundation.org/united
14. Moats, L. (2009a). Knowledge foundations for teaching reading and spelling. *Reading and Writing: An Interdisciplinary Journal*, *22*, 379–399.
15. Pacheco, A. (2009). Mapping the terrain of teacher quality. In Gitomer (Ed.), *Measurement issues and assessment for teaching quality (*pp. 160-178). London: Sage Publications.
16. Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement

a. trajectories in elementary schools. *American Educational Research Journal, 45*, 365-397.

17. Pianta. R., & Hamre, B., (2009). Conceptualization, measurement and improvement of classroom processes:

a. Standardized observation can leverage capacity. *Educational Researcher,* 38,2, pp. 109-119.

18. Pressley, M., Wharton, R., Allington, R., Block, C. , Morrow, L., Tracey, D., et al. (2001). A study of effective first-grade

a. literacy instruction. *Scientific Studies of Reading, 5,* 35–58.

19. Pressley, M., Roehrig, A., Raphael, L., Dolezal, S., Bohn, C., Mohan, L., Wharton-McDonald, R., Bogner, K., & Hogan, K.

a. (2003). Teaching processes in elementary and secondary education. In W. Reynolds & G. Miller (Eds.), *Handbook of Psychology: Volume 7, Educational Psychology* (pp. 153-176). Hoboken, NJ: John Wiley & Sons, Inc.

20. Raudenbush, S., Martinez, A., Bloom, H., Zhu, P., Lin, F. (2010). Studying the reliability of group-level measures with

a. implications for statistical power: A six-step paradigm.

21. Rimm-Kaufman, S. E. & Hamre, B. K. (2010). The role of psychological and developmental science in efforts to

a. improve teacher quality. *Teachers College Record, 112*(12), 2988-3023.

22. Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and

a. research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454-499.

23. Soar R. S., & Soar, R. M. (1979).Emotional climate and management. In P.L. Peterson & H. J. Walberg

a. (Eds.),*Research on teaching: Concepts, findings and implications* (pp. 97-119). Berkeley, CA: McCutchen Publishing Co.

24. Snow, C.E., Griffin, P., & Burns, M.S. (2005). *Knowledge to support the teaching of reading: Preparing teachers for a*

a. *changing world.* San Francisco: Jossey-Bass.

25. Taylor, B. M., Pearson, P. D., Peterson, D. S., & Rodriguez, M. C. (2003). Reading growth in high-poverty classrooms:

a. The influence of teacher practices that encourage cognitive engagement in literacy learning, *Elementary School*

b. *Journal, 104*, 3-28.

26. van den Berg, S., Glas, C., & Boomsma, D. (2007). Variance decomposition using an IRT measurement model.

a. *Behavioral Genetics*, 37, pp. 604-616.

27. Wang, W., & Wilson, M. (2005). Exploring Local Item Dependence Using A Random-Effects Facet Model, *Applied*

a. *Psychological Measurement,* 29, 4, pp. 296-318.

28. Yen, W. (1984). Effects of local item dependence on the fit and equating performance of the three parameter logistic

a. model. *Applied Psychological Measurement*, 8, 125-145.

Author' biography

Ben Kelcey is an assistant professor in the College of Education, Criminal Justice, & Human Services at the University of Cincinnati. His research interests include the development and application of measurement and quantitative research methods to understand effective teaching and teachers.