Abstract Title Page
Not included in page count.

Title: A Multilevel Bi-factor Framework for the Measurement of Instruction

Authors and Affiliations:
Ben Kelcey
ben.kelcey@gmail.com
Wayne State University

Joanne Carlisle
jfcarl@umich.edu
University of Michigan

Abstract Body

Background / Context:

Evidence suggests that elementary classroom teachers adapt their instruction based on, for example, the goal, topic and context of instruction as well as students' needs (e.g., Brophy & Good, 1986). In delivering instruction in early elementary reading, lessons conceptually represent a basic unit of teachers' instructional practice. Teachers primarily design and deliver instruction through lesson units. In teaching a lesson on phonics, for example, teachers consider the relation of the instruction and the needs of students participating in the lesson, available materials, time available for the lesson, and so on; separate plans are made to take into account the goal, topic, materials (etc.) for a lesson on reading comprehension or vocabulary. Within lessons, the purpose may dictate the choice of instructional actions (e.g., explaining/questioning), but events within the lesson are likely to affect the manner in which instruction is carried out.

Numerous classroom observation systems have been designed to identify and understand effective teaching practices. Although the scope of these systems varies widely, most make use of repeated observations across blocks of instruction to chronicle the presence, depth, or quality of instructional features or actions believed to be reflective of core dimensions of teaching. In this way, repeated observations of instruction tend to include different types of lessons (e.g., phonics, vocabulary). Despite variation in elementary teachers' use of certain practices in lessons taught for different purposes, researchers have tended to collapse their data across lesson purposes. More specifically, common practice is to make use of simple counts or averages of teacher behaviors across lesson types and this has potentially led to inconsistent evidence regarding various aspects of instructional effectiveness (e.g., Hoffman, 1991).

Purpose / Objective / Research Question / Focus of Study:

Of interest here is the extent to which collapsing observed instruction across lessons ignores salient information introduced by the purpose of lessons. More specifically, our work focused on the extent to which observed instruction is principally informed by a primary overarching dimension that is stable across lessons but also lesson specific dimensions that systematically differentiate instruction across lessons with different purposes (Figure 4). We argue that lessons represent an important conceptual block that binds teachers' instruction, and as a result, the prevalence of specific instructional features varies with lesson purpose (among other conditions). By ignoring the purposes of lessons and collapsing observed instruction across lessons of different purposes, we assume instruction is only informed by a single dimension. Statistically, this introduces the potential for obscuring or missing key differences among teachers on the targeted primary dimension. Substantively, ignoring the relations among features associated with lesson purpose is largely inconsistent with extant instructional theory because it treats instruction as if it proceeded without a goal or purpose. If our theory for instruction suggests that instruction shifts with lesson purpose, then our observation systems and analyses should reflect this.

To this end, we conceptualized a multilevel bi-factor measurement model for the measurement of instruction and explored an application. Specifically, our framework allows observed instruction to be guided by a targeted primary dimension as well as a set of secondary lesson purpose specific dimensions (e.g., Figure 4). Thus, we explored how lesson purposes might drive shifts in instruction that induce excess correlation among features within lessons of the same type. Taking as an example, we drew on a study examining the extent to which teachers' instruction supports students' vocabulary growth across comprehension and vocabulary lessons. Our bi-factor framework first suggests that observed instruction is guided chiefly by an

overarching primary dimension that is common across lesson types. In addition, observed instruction is informed by dimensions specific to each lesson purpose (comprehension and vocabulary). That is, we leveraged lesson purpose specific dimensions to explain systematic changes in the prevalence and depth of supportive vocabulary instruction between comprehension and vocabulary lessons. In this way, we identify how instruction changes with lesson purpose and how instructional indicators might differentially reflect the primary dimension in lessons with different purposes. Below, we briefly describe how features of the multilevel bi-factor model align with salient features of instruction. We then apply this framework to measure the extent to which teachers' instruction is supportive of vocabulary learning in second and third grade reading comprehension and vocabulary lessons.

Setting: This research took place in Reading First schools in Michigan.

Population / Participants / Subjects:
Our sample included 86 second- and third-grade teachers who taught in Reading First schools in Michigan. Table 1 briefly describes characteristics of teachers. The 86 teachers' reading instruction was observed four times across the year; once instruction in these literacy blocks was analyzed, we found that in total, just over 1000 lessons were observed. Lessons lasted on average 15 minutes, and vocabulary and comprehension lesson represented approximately 40% of the total lessons observed throughout the literacy block.

Intervention / Program / Practice:
Application of our framework focused on identifying and coding features of teachers' instruction that reflect their support for students' vocabulary development in grade two and three vocabulary and comprehension lessons. Specifically, we recorded the presence of five instructional features in comprehension and vocabulary lessons taught by each of the teachers. We considered these features as nested within lessons, which are in turn nested in teachers. The features coded during literacy instruction were these: the teacher (1) defines a word or word parts (e.g., a hoe is a garden tool); (2) states or reads a sentence showing the meaning or use of a word in context (e.g., "let's read the sentence to see how the word hoe is used and what it might mean"); (3) asks students to explain a word's meaning (e.g., "Can anyone tell me what a silo is?"); (4) asks students to use a word in a sentence (e.g., "Can you use the word silo in a sentence?"; (5) fosters discussion of word meaning (e.g., "The passage we read mentions hoe, rake and shovel. Let's talk about how these tools have different purposes."). The features are intended to reflect incremental levels of cognitive engagement or challenge the teacher can place on students. For instance, a teachers' simply defining a word for students is thought to engage students cognitively less than fostering a discussion of a word (Stahl, 2010).

Significance / Novelty of study:
Although researchers' attention has been directed toward the content and process of teachers' classroom instruction, analytic methods that synthesize such content have received far less attention. Yet, because the quality of measures derived from such systems is heavily dependent on how the features are synthesized or combined, measurement of instruction is dependent on observation and coding systems and analytic methods suited to relate these codes to underlying dimensions. While measurement is central in many aspects of educational research, measurement seems underemphasized in extant research on instructional practice.

Research capturing observed differences in instruction has predominantly relied on classical test theory (CTT), using simple sums or averages across lessons (e.g., Cirino et al., 2007). Such approaches tend to neglect a number of aspects that are critical parts of instruction. For instance, summing or taking the average use of targeted instructional features assumes features are unidimensional and that features load uniformly on that single dimension. Similarly, use of sums or means inherently assumes that the patterns of instruction offer no useful information in describing instruction. Furthermore, such historical approaches ignore the potential for features to cluster in certain lessons and assume the factor structure is uniform across teacher and lesson levels.

Statistical, Measurement, or Econometric Model:

The dependencies of features brought on by lesson purpose can obscure the signals reflected onto indicators by the primary latent dimension and have important implications for measurement. For example, purpose driven shifts in instruction induce additional correlation among features within lessons of the same type thus violating the assumption of unidimensionality and local independence (e.g., Yen, 1993). Through the multilevel bi-factor model, we leveraged inter-item dependencies to relax the unidimensionality and local feature independence assumptions and to understand how aspects of instruction are guided by primary dimensions but also responsive to lesson purpose.

Within this framework, we situate instructional features within their lesson purpose and then examine them as functions of a primary dimension and lesson purpose specific dimension (e.g., Mcleod, Swygert & Thissen, 2001). We considered the primary instructional dimension as the target and assumed the lesson purpose specific dimensions were orthogonal to the primary dimension. More specifically, we write the multilevel bi-factor model as

$$P(Y_{ijk}=1\,|\,\boldsymbol{\theta})=\frac{1}{1+\exp(a_i^{gw}\theta_{jk}^{g}+a_i^{gb}\theta_{k}^{g}+a_i^{sw}\theta_{jk}^{s}-d_i)} \qquad (1)$$

where $P(Y_{ijk}=1\,|\,\boldsymbol{\theta})$ is the probability of observing instructional feature *i* in lesson *j* for teacher *k* given the item loading parameters for the general dimension within teachers, $a_{\cdot}^{gw}$, and between teachers, $a_{\cdot}^{gb}$ and the lesson purpose specific dimensions, $a_{\cdot}^{sw}$, the item difficulty parameter, d_i, and respective latent traits θ. In the application of the bi-factor model, each instructional feature will have a nonzero loading onto the targeted primary instructional dimension, $a_{\cdot}^{g\cdot}$, as well as one other loading for the lesson purpose to which it belongs (e.g., DeMars, 2006). Remaining loadings (other lesson purposes) to which the feature does not belong are constrained to zero as are the covariances among the traits.

Usefulness / Applicability of Method:

Substantively, consideration of lesson purpose specific dimensions acknowledges the ways in which instruction is coordinated and responsive to the goals of a lesson. Statistically, this affordance helps to address residual dependencies among features to more reliably recover the core primary dimension. In these ways, describing instruction as a multidimensional process presents a more holistic and comprehensive picture of instruction and aligns substantive theory with empirical analysis.

Data Collection and Analysis:

Data were collected at four intervals across the schools year. Observations were captured with

the Automated Classroom Observation System for Reading (ACOS-R). ACOS-R is a system designed to code the presence of features of early elementary reading instruction. Coding was carried out in 5-min intervals; during each interval, observers recorded multiple fields including the purpose of the lesson and use of word meaning discourse actions (features of interest in this study). Observers designated a change in activity to indicate the start of a new lesson within a 5-min interval (Authors, 2011).

To assess the extent to which the multilevel bi-factor model (2L Bi) described instruction appropriately, we applied it to examine how the use of the measured features was informed by a primary vocabulary support dimension and two lesson purpose (vocabulary and comprehension lessons) specific dimensions (Figure 4). To describe its fit compared to alternative models, we also fitted a single level unidimensional model (1L Uni), a single level bi-factor model (1L Bi), and a two level unidimensional model (2L Uni) (Figures 2-4). Similar to traditional item response methods, the single level unidimensional model assumes the factor structure is unidimensional and aggregateable. In other words, it assumes instructional features are solely guided by the overarching vocabulary support dimension and that the intraclass correlation (ICC) for the factor structure is zero. The single level bi-factor model extends this to allow lesson purpose specific dimensions but still assumes that instructional features in the same lesson are otherwise independent (not clustered and ICC is 0). The two level unidimensional model takes into account that features within the same lesson are not independent but assumes that instruction is not informed by lesson purpose.

Findings / Results:

Our results indicated that the two level bi-factor model significantly outperformed alternative models (Table 1). In comparing parameter estimates, we first found that the intraclass correlation (ICC) for the unidimensional model was much lower compared to the ICC for the bi-factor model (Table 2). We also found that each alternative model suggested factor loading on the primary dimension that were higher than their multilevel bi-factor counterparts (Figure 5). Moreover, the results suggested that the extent to which employing an instructional feature was reflective of the targeted primary dimension varied between lesson purposes. More specifically, take for example 'examining a word in context'. While this action was fairly reflective of the primary dimension in comprehension lessons, its use in vocabulary lesson constituted more weight. That is, examining a word in context in a vocabulary lesson was more reflective of high levels or vocabulary support than doing so in a comprehension lesson. Put differently, examining a word in context in a comprehension lessons does not differentiate among teachers as well as examining a word in context in a vocabulary lesson. The results for this specific feature in the alternative models which ignored secondary dimensions and/or the clustering of features in lessons diverged in this regard.

Conclusions:

Our results lend empirical evidence to the importance of studying instruction by treating lesson purpose as the teachers' unit of instruction and examining dimensions or features of instruction within lessons. The results suggest that the multilevel bi-factor model may more appropriately describe instruction and how it changes over lessons. Moreover, the results suggest that assuming unidimensionality across lesson purposes obscures measurement of the primary dimension in ways that both diminish the variance attributed to differences among teachers and overstate the ability of the measured features in discriminating among teachers.

Appendices

Not included in page count.

Appendix A. References

Stodolsky, S.S. (1990). Classroom observation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 175-190). Newbury Park: Sage.

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.

DeMars, C. (2006). Application of the bi-factor multidimensional item response theory model to testlet-based tests. *Journal of Educational Measurement, 43,2, pp. 145-168.*

Hoffman, J. V. (1991). Teacher and school effects in learning to read. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. 2* (pp. 911-950). New York: Longman.

McLeod, L. D., Swygert, K. A., & Thissen, D. (2001). Factor analysis for items scored in two categories. In D. Thissen & H. Wainer (Eds.), Test scoring (pp. 189–216). Mahwah, NJ: Lawrence Erlbaum Associates.

Rosenshine, B. (1983). Teaching functions in instructional programs. *Elementary School Journal, 83*, 335-351.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187–213.

Appendix B. Tables and Figures

Table 1: Brief teacher descriptive statistics

	Mean	SD
New Teacher	0.19	0.39
Female	0.94	0.24
Minority	0.20	0.40
Masters	0.55	0.50
Years Experience	13.88	10.90

Table 2: Model fit indices for one/two level unidimensional/bi-factor models

	1L Uni	1L Bi	2L Uni	2L Bi
-2LL	3852.65	3779.56	3810.16	3712.53
AIC	3892.65	3839.56	3850.16	3772.53
BIC	3987.15	3981.31	3944.45	3913.96
N of parameters	20	30	20	30

Table 3: Variance and intraclass correlation estimates

	Var(General Between)	SE	Var(General Within)	SE	ICC
2L Uni	0.20	0.07	1.00	-	0.17
2L Bi	0.38	0.13	1.00	-	0.28

Table 4: Factor loadings

	1LUni	1L Bi	2L Uni	2L Bi
Examine word in context in vocabulary lesson	2.24	1.21	2.48	2.52
Examine word in context in comprehension lesson	2.67	2.00	2.46	1.38
Students give meaning of word in vocabulary lesson	0.50	0.97	0.27	0.17
Students give meaning of word in comprehension lesson	0.90	4.74	0.77	0.95
Define word in vocabulary lesson	1.57	0.73	1.21	0.62
Define word in comprehension lesson	1.93	1.33	1.50	0.56
Fosters discussion in vocabulary lesson	1.52	1.11	1.14	0.42
Fosters discussion in comprehension lesson	1.23	0.16	1.03	0.15
Students use word in sentence in vocabulary lesson	1.80	1.85	1.04	0.27
Students use word in sentence in comprehension lesson	2.57	1.85	2.01	0.55

Figure 1: Single level unidimensional model

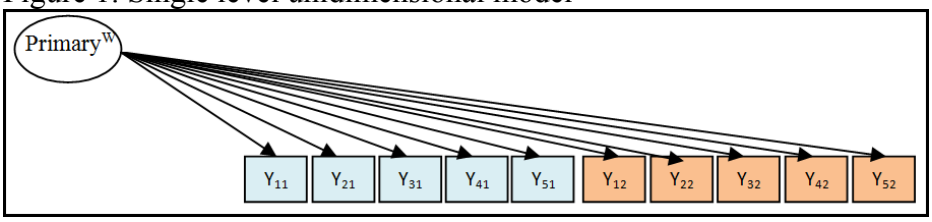

Figure 2: Single level bi-factor model

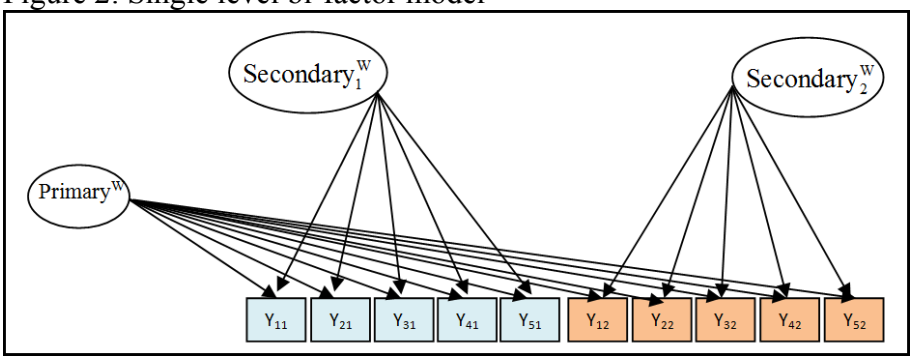

Figure 3: Two level unidimensional model

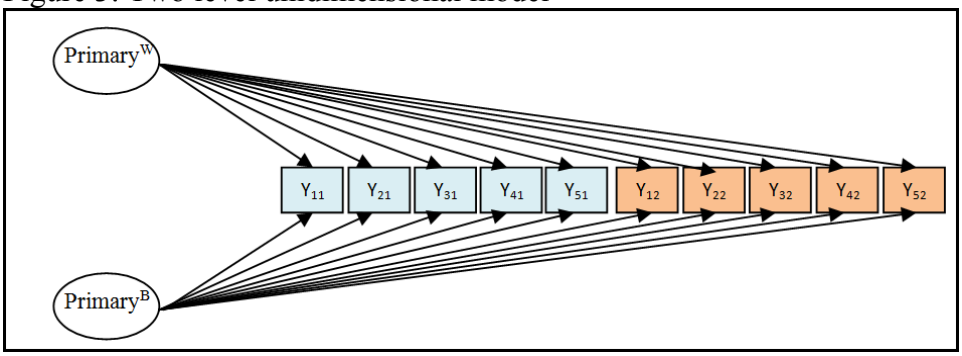

Figure 4: Multilevel level bi-factor model

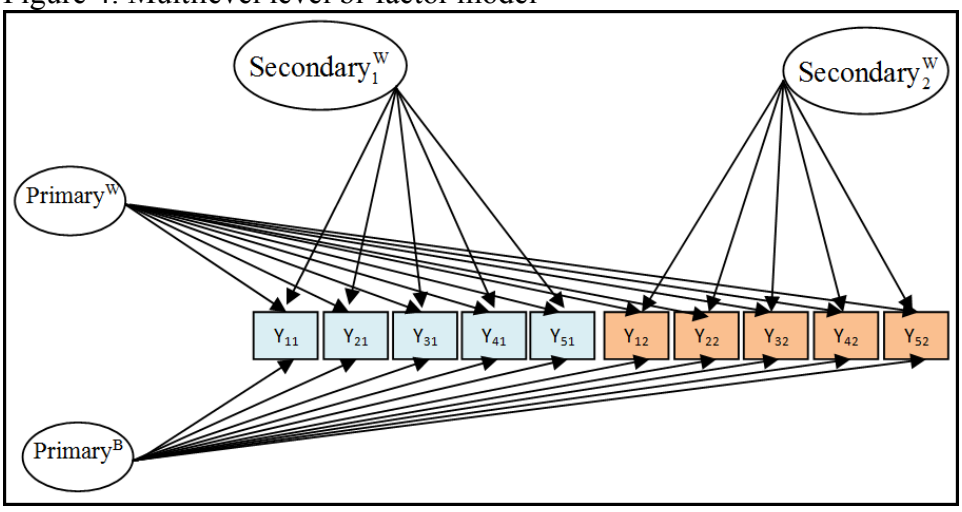

Figure 5: Comparison of teacher level general factor loadings for a two level unidimensional model (2L Uni) versus a two level bi-factor (2L Bi)