EUROPEAN JOURNAL OF EDUCATIONAL RESEARCH Vol. 3, No. 2, 87-109

Measuring Teaching Best Practice in the Induction Years: Development and Validation of an Item-Level Assessment

Laurie Kingsley[1], William Romine[2*]

[1]*211A Townsend Hall, University of Missouri, Columbia, MO 65211, USA.*
E-mail: kingsleyl@missouri.edu

[2]*Department of Biological Sciences, Wright State University, 3640 Colonel Glenn Hwy, Dayton, OH 45435, USA*

**E-mail: romine.william@gmail.com*

Schools and teacher induction programs around the world routinely assess teaching best practice to inform accreditation, tenure/promotion, and professional development decisions. Routine assessment is also necessary to ensure that teachers entering the profession get the assistance they need to develop and succeed. We introduce the Item-Level Assessment of Teaching practice (I-LAST) as a flexible framework-based approach for quantitative evaluation of teaching best practice in the induction stages. We based the I-LAST on a novel framework for teaching best practice, and used Fuller's scale as a framework for understanding the potential of the I-LAST in providing longitudinal measures for growth. Using the context of a year-long teacher induction program in the Midwestern United States, we collected data through an online survey from 46 teaching supervisors who were asked to evaluate their interns. We used the Rasch partial credit model as a criterion for construct validity, and measured dimensionality and reliability from both Rasch and classical frameworks. The I-LAST was found to be a unidimensional, valid, and reliable measure for teaching best practice. It demonstrated the ability to provide reliable scores for specific sub-dimensions of best practice, including those which manifest at various stages along Fuller's scale. Potential uses of the I-LAST to advance understanding of the role of teacher induction programs in fostering productive growth in new teachers is discussed.

Keywords: Teacher induction, development, assessment, Rasch modeling, teaching internship, teacher attrition

Introduction

National reports in the United States have been calling for improvements in teacher education for several decades (e.g. *A Nation at Risk*, 1983; *A Nation Prepared: Teachers for the 21st Century*, Carnegie Foundation, 1986). Recent similar policy documentation in Europe (ETUCE, 2008) indicates that this goal is cross-culturally shared. In the United States, the No Child Left Behind Act (NCLB, 2001) included a requirement for "highly qualified" teachers, defined as those who are licensed by the state and have competence in the subject matter they will teach. More recently the U.S. Dept. of Education released *A Blueprint for Reform: The Reauthorization of the Elementary and Secondary Education Act* (March, 2010 http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf), calling for

ISSN 2165-8714

http://www.eurojedu.com

states to develop techniques to evaluate and identify highly "effective" teachers. School districts implementing the evaluation techniques prescribed by their state though the reauthorization will be allowed flexibility in how they use federal education funds. National and state accreditation teams such as the National Council for Accreditation of Teacher Education (NCATE) and the Teacher Education Accreditation Council (TEAC) also look for evidence that teacher education programs produce highly qualified teachers, and colleges of education are continuously looking to demonstrate their effectiveness (Finn, 2003).

Given the amount of evidence available to support the notion that the quality of the teacher is the most influential determinant of student achievement (Wright, Horn, & Sanders, 1997), it becomes imperative to identify key characteristics of quality teachers that go beyond the NCLB definition, which suggests that quality teachers are merely those who have graduated and are licensed (Thomas and Schubert, 2001; Goldhaber, 2006). Students in classrooms with "effective" teachers benefit significantly (Gordon, Kane, & Staiger, 2006), a trend largely independent of national boundaries (Akiba, LeTendre, & Scribner, 2007). However, there is continued debate regarding what constitutes highly effective teachers, and how to measure effectiveness (Cochran-Smith & Fries, 2001; Mullen & Farinas, 2003).

While policymakers and educators generally agree on the importance of instilling ideas of teaching best practice in pre-service teachers during the internship experience, options for measuring associated outcomes are limited. Self assessments (DeFina, 1992), student reports (Burnett & Meacham, 2002), and assessments for use by administrators and researchers (Capie et al., 1979; Stulac, 1982), which draw upon various frameworks for best practice, have been developed and utilized. However, current instruments are limited by the specificity of the setting in which they can be effectively applied. Further, there are no current instruments for teaching best practice validated using modern psychometric methods. It is necessary to address these needs through development and validation of a flexible assessment of teaching best practice using the Rasch framework for construct validity and a general, widely-applicable framework for teaching best practice.

Review of Literature

Assessing Best Practice in the Elementary Classroom

Ideas about what makes a quality teacher are many, and thus numerous strategies for assessing teaching best practice have been proposed. Self assessment methods, including teacher-constructed portfolios (DeFina, 1992; Wolf, 1989) are commonly used as a method of assessment for new and pre-service teachers. Portfolios often include examples of lesson plans and student work accompanied by the intern's reflections. Burnett and Meacham (2002) advocate use of student-centered teacher evaluation at the elementary level, proposing creation and implementation of the My Teacher Scale from previous items developed to quantify elementary students' observations of their teacher (Bitner-Kratzner, 1995; Thomas & Montgomery, 1998). Teacher self report assessment methods have been criticized for potential biases introduced by teacher self perceptions rooted outside of the classroom (D'Onofrio, 1989). Use of student-centered evaluation strategies is criticized along a similar line, namely that such reports tend to address the extent to which a student likes a teacher or class as opposed to actual best practice (Ross, McDougall, Hogaboam-Gray, & LeSage, 2003).

In light of these objections, subjective qualitative yearly evaluations by an outside expert such as a supervisor or administrator have commonly been used to facilitate decisions regarding professional development, tenure, and promotion. However, since trustworthy qualitative conclusions derived from observational data require much more than an hour or two of observation per year, quantitative survey methods with definitive construct validity and reliability should be considered. To this end, a variety of instruments have been developed as quantitative measures for elementary teaching best practice in both administrative and research settings.

Burnett and Meacham (2002) describe several notable state-level assessments used by administrators in the United States, including Georgia's Teacher Performance Appraisal Instrument (Capie et al. 1979), Florida's Research Based Observation instrument, and South Carolina's Assessments of Performance in Teaching (Stulac, 1982). Burnett and Meacham cite psychometric concerns with these assessments, including insufficient reliability and lack of generalizability to diverse teaching situations.

More recently, the Ohio State Teacher Efficacy Scale (Tschannen-Moran & Hoy, 2001) was developed to measure teaching efficacy in pre-service and inservice teachers using the Teacher Efficacy Scale (Gibson & Dembo, 1984) as a starting point. The 30-item Teacher Efficacy Scale measures two factors (personal teaching efficacy, and teaching efficacy) using a 6-point Likert scale based on Bandura's Social Cognitive Theory (Bandura, 1986) with reported reliabilities of 0.75 and 0.79, respectively. After three implementation cycles, Tschannen-Moran and Hoy (2001) developed a 12-item short form and 24-item long form which measured three factors (instruction, management, and engagement). Subscale reliabilities for the long form ranged from 0.87 to 0.91. Short form reliabilities fell between 0.81 and 0.86.

Ross et al. (2003) developed and validated a 20-item 6-point Likert instrument for measuring nine dimensions of reform-based mathematics best practice at the elementary level, including ability to develop complex, authentic learning tasks for students, facilitate student-to-student interaction, and implement appropriate formative and summative assessment strategies. Content and face validity were established through review by elementary teachers. Although measuring nine topic dimensions, the instrument was treated as a single unidimensional scale for measuring K-8 teachers, and demonstrated satisfactory reliability above 0.80.

Penuel, Boscardin, Masyn, & Crawford (2007) developed a tool to measure teachers' use of student response system technologies across grades K-12 in a variety of subject areas, including mathematics, science, and language arts. Their instrument included teachers' goals for instruction and pedagogical practices. Two underlying factors related to goals ("improving instruction" and "improving assessment") were extracted, with reported item reliabilities between 0.60 and 0.87. Five factors related to pedagogy were extracted in a similar fashion. Examples include ability to check content understanding, pose diagnostic questions, and use feedback to adjust instruction. Item reliabilities between 0.31 and 0.77 resulted from this five-factor model.

We see that a variety of tools have been developed and implemented to measure best practice in elementary teachers, and that these take a variety of perspectives on what it means to be an effective teacher. However, current quantitative instruments are limited by the specificity of the setting in which they can be effectively applied, and are not validated using modern psychometric methods. These instruments take a Classical Test Theory (CTT) validation framework where reliability of items is measured based on correlation with other items on the instrument. While a small number of highly correlated items can give a reliable score for a specific construct related to teaching best practice, a more comprehensive instrument is needed to give a holistic look at teaching best practice and how it changes as teachers develop. We attempt to address this gap through development and validation of the Item-Level Assessment of Teaching (I-LAST). Item development and validation procedures are implemented with diversity and flexibility in mind, supporting the important notion that a set of items appropriate for evaluating one teacher may not be appropriate for evaluating another teacher in a different instructional setting and/or stage of development.

Development of Best Practice

The NCLB act defines a highly qualified teacher as one who is licensed by the state and has competence in the subject matter taught (NCLB, 2001), implying that reaching a certain level of education will automatically make a person qualified to teach. Wenglinsky et al. (2000) uses external attributes such as education and experience only as a partial measure of teacher quality.

Although a positive correlation between an instructor's level of education and his/her quality of practice exists (Darling-Hammond & Youngs, 2002), educators generally agree that teacher effectiveness goes beyond content knowledge (NCATE, 2006). Elements arguably more important, but difficult to quantify, are found in actual classroom practice. Zemelman, Daniels, and Hyde (2005), suggest that "… all the authoritative voices and documents in every teaching field are calling for schools that are more student-centered, active, experiential, authentic, democratic, collaborative, rigorous, and challenging. That's the short definition of Best Practice teaching" (p. vii). These descriptors must be included in any definition of "effective teacher".

Models of teacher development (Fuller, 1969; Berliner, 1988, Kagan, 1992) suggest there are stages through which beginning teachers progress which influence their practice. During initial stages of development, teachers are concerned with teaching ability, management of the classroom, and popularity with students. Middle stage concerns move from those of survival to those of actual teaching such as planning lessons, presenting information clearly, and organizing the classroom for efficiency. By the late novice stage, the teacher's concentration shifts from him/herself as a teacher to the students as learners (Fuller, 1969). Because it has been shown that students taught by more effective teachers make greater learning gains when compared to students with less effective teachers (Sanders & Rivers, 1996; Rivkin et al. 2002; Clotfelter, Ladd, & Vigdor 2007), and that developed teachers are more concerned with student learning (Fuller, 1969), timely transition into the late novice stage becomes critical. Teacher education programs may have greater impact on student achievement if interns from those programs are more easily able to transition to the higher stages of development where issues of student learning become priority concerns.

Increased opportunities for classroom engagement in an extended field experience could help interns move through stages of development with greater efficiency. Fine-tuning procedural skills (early stages of teaching development) in turn may help pre-service teachers transition to a focus on the issues of student learning (later stages of teaching development) which is seen as a measure of teacher effectiveness. Assessing this effectiveness serves as one accountability measure for colleges of teacher education as they seek to justify funding, stay or become accredited, and recruit students for their programs.

Purpose of the Research

The purpose of this project was to design and validate an instrument that effectively quantifies the teaching best practice of new and preservice teachers. Using best practice ideas as summarized by Zemelman et al. (2005), and themes which closely overlap best practice ideas identified in a multiple case study (Kingsley, 2007) through grounded theory traditions (Glaser & Straus, 1967), the I-LAST instrument was developed. Aspects of teaching quality on the I-LAST are: (1) managing students and the classroom ("Management"), (2) teacher holding students accountable for their learning ("Student Accountability"), (3) student assessment ("Assessment"), (4) teacher holding him/herself accountable ("Teacher Accountability"), (5) individualizing instruction ("Individualizing Instruction"), and (6) literacy content and practice ("Literacy"). Since these aspects of practice are learning-focused, they are likely to improve as interns develop (Kingsley, 2007).

Using the I-LAST, the following research questions were explored:

1. Are there unique underlying dimensions to teacher best practice, or are the various practices a subset of a single underlying dimension?
2. What are key statistically identifiable underlying dimensions of best practice?
3. Which aspects of best practice do elementary interns emphasize?
4. Which aspects of best practice are the most difficult for elementary interns to implement?

Methods

Context of the Study

A growing number of researchers agree that a disconnect between coursework and practice in the field prevails (Zeichner, 2010; Goodlad, 1990; Holmes Group, 1986), and many argue that what is learned from practice may have more impact than what is learned in teacher preparation coursework. This study examines pre-service teaching interns participating in year-long senior year placements in partner schools, a program developed in part to ameliorate the disconnect students experience between coursework and field placements. The year-long teacher internship program grew out of a partnership between a large Midwestern United States university and approximately 20 participating school districts in the state. The partnership was created as part of the Goodlad Group, and supports initiatives for educational renewal (Goodlad, 1994). All elementary education seniors are placed in one of these partner district's schools for the entire year. When the school year begins, the intern is immediately immersed in the culture of the school, classrooms, and students. Coursework is completed throughout the year, both on-site and on campus, and the focus of the content is literacy. Interns are treated as faculty and attend all school activities, meetings, and parent-teacher conferences. This year-long program is grounded in reflection and adaptation, theory is tied directly to classroom practice, and multiple faculty from the district and university work together to support the interns. This extended internship is advantageous in that it offers more classroom experience for interns, helping them become more developed and effective teachers.

Sample

The year-long program was a collaboration of 117 elementary teachers from 20 schools around the Midwestern United States. Using Qualtrics, the assessment was sent to all teachers who were asked to evaluate their previous year's intern. Forty six teachers who had interns the previous year responded.

Instrument Development

Using a four-tiered Likert scale, teachers were asked to rate statements, "Strongly Disagree," "Disagree," "Agree", or "Strongly Agree," which were scored 1, 2, 3, and 4, respectively. Both positive and negative statements (see Appendix) were used to prevent response bias. Seventeen items (Q1-Q17) were developed to assess Management, 12 items (Q18-Q29) to assess Student Accountability, 15 items (Q30-Q44) to assess Assessment, 15 items (Q45-Q59) to assess Teacher Accountability, 14 items (Q60-Q73) to assess Individualizing Instruction, and 21 items (Q74-Q94) to assess content instruction (Literacy). All items were reviewed for content and face validity by three professors of literacy education and three elementary teachers. The items above were approved as content and face valid by at least two out of three professors and two out of three teachers.

Rasch Modeling

With the understanding that the I-LAST may be used to assess teaching best practice under a variety of circumstances, our focus was not on eliminating items. Rather, we sought to identify the unique measurement characteristics of individual items to help future researchers make informed choices regarding the most appropriate items for particular assessment situations. Since item-level analysis was our focus, we took a Rasch approach to instrument validation.

Construct validity of I-LAST items was evaluated with respect to the Rasch partial credit model (Masters, 1982). As with all models in the Rasch family, the partial credit model takes a philosophical stance on instrument validity in its statement that the probability of a participant selecting a particular level of agreement should be proportional only to the difference between the supervisor's approval of a

teacher's practice and the item's agreeability (Wright & Stone, 1979). An item fitting these characteristics is expected to fit well with the Rasch partial credit model. Rasch models are also attractive in that they provide a set of non-crossing curves which serve as a fundamental criterion for measurement (Andrich, 1994; Wright, 1997), thus allowing person and item measures to be compared on the same scale (Wright & Stone, 1979). Consequently, Rasch models are useful in identifying items that provide misleading measurement information; items that do not discriminate well, or those that miscategorize students can be identified though lack of fit with the model (Lincare, 2010).

Through BIGSTEPS (Linacre & Wright, 2006), Joint Maximum Likelihood Estimation (JMLE) was used to calculate person abilities, item difficulties, and goodness of fit of each item with the partial credit model. The absence of distributional assumptions of JMLE makes it an attractive method for analyzing a smaller data set. Goodness of fit was measured with mean squares outfit (outlier sensitive) and infit (information weighted to reduce the effect of outliers) statistics. Mean squares fit statistics have been shown to be sample size invariant in polytomous models such as the partial credit model (Smith et al. 2008), and have expected values of 1.0 (Wright & Masters, 1982). Fit values larger than 1.0 indicate noise in the data that is not modeled; lower values indicate a Guttman pattern (lack of stochasticity), which is evidence of unusually high item discrimination. Items with fit values between 0.5 and 1.5 are generally considered productive for measurement (Wright & Linacre, 1996). Point biserial correlations were used as a diagnostic indicator—a negative correlation suggests that an item favors lower ability teachers (similar to negative item discrimination), which is contrary to what one would expect from a well-written item (Linacre, 2010).

Reliability Analyses

We used both Classical Test Theory (CTT) and Rasch perspectives to explore reliability of the entire instrument. In addition, Rasch item reliability was used to quantify the precision of item measures. From the CTT perspective, we used Cronbach's alpha to indicate loss of precision due to instrument design (Nunnally, 1967, Cronbach, 1947, 1951) for both the entire instrument and item clusters of interest. From the CTT perspective, high reliability results from high correlation between responses on items and/or a large number of items (Schmidt, 1996). However, reliability calculated from the Rasch perspective provides an item-level measure of precision, using the sum of information contributed by all items on the instrument. Since an item provides the most information about participants with ability levels proximal to the item's difficulty level, the greatest reliability is achieved when item difficulty indices match the ability of the sample.

Unidimensionality Tests and PCA on Rasch Residuals

Rasch is a confirmatory model, meaning that estimates are based on an *a priori* assumption of unidimensionality. Since we developed this instrument to measure the underlying dimension of teaching best practice, we took the initial assumption of a unidimensional assessment, which we tested through both CTT and Rasch methodologies. From a CTT perspective, we performed a principal components factor analysis on the items using the scree criterion (Cattell, 1966) for dimensionality selection. However, attention was not given to the item-factor loadings since items which correlate highly and perform well in factor analysis may not be optimal from the Rasch perspective (Wright, 1996).

From the Rasch perspective, departure from unidimensionality was quantified by variance not accounted for by the Rasch model which shows up in the residuals. Thus, principal components analysis (PCA) on Rasch residuals served as a way to detect underlying dimensions not accounted for by the model. An eigenvalue around 2 items of variance for an underlying dimension is generally accepted to indicate random noise in the residuals (Raiche, 2005). In addition to detecting unmeasured variance, this method is useful for identifying items which cluster together. However, unlike traditional

factor analysis, PCA on residuals allows detection of item contrasts, or comparison of items with positive loadings and those with negative loadings onto the underlying residual dimension. We considered item loadings with a magnitude greater than 0.4 to have a practically significant contribution to an underlying residual dimension. Since the focus of our instrument validation procedure is not to eliminate items, but to identify the unique measurement characteristics of individual items, results from PCA on residuals are useful in helping researchers select which items provide the most appropriate measurement information for particular aspects of teaching best practice.

Results and Discussion

Description of the Sample

Respondents were teaching internship supervisors in grades 1-6. In addition to survey ratings, the supervising instructors reported on school environment and literacy program used (Table 1). Twenty-six teachers supervised interns in grades 1-3, 13 in grades 4-5, and 7 in grade 6. School sizes ranged from 42 to 900 students, with a mean of 445, and a standard deviation of 192. Class sizes ranged from 18 to 30, with a mean of 22, and a standard deviation of 3. Reported number of students per grade ranged from 20 to 130, with a mean of 72 and a standard deviation of 29. A variety of literacy instructional programs were represented. Five teachers reported using a fixed basal reading series for literacy instruction (Pgm 1), 22 used elements of balanced literacy (Pgm 2), and 16 used instructional schemes that are based highly on individual needs (Pgm 3).

Table 1. *Descriptive data on instructors' school teaching environments*

	N	Min	Max	M	SD
Sample					
Level					
Grd 1-3	26				
Grd 4-5	13				
Grd 6	7				
SchoolSize	46	42	900	445	192
ClassSize	46	18	30	22	3
GradeSize	46	20	130	72	29
LitPrgm	43	1	3	2	1
Pgm 1	5				
Pgm 2	22				
Pgm 3	16				

Rasch Measures and Reliability

Logit difficulty measures for items ranged from -1.40 to 3.01, with a mean of 0 and a standard deviation of 1.08. A majority of items had Rasch infit and outfit measures within the 0.5 to 1.5 range suggested by Wright and Linacre (1996). However, fourteen items had fit statistics outside of these ranges. Only one item, Q93, had a fit statistic below 0.5, indicating that this item had unusually high discrimination. Hence, it may be measuring an additional underlying dimension that is correlated with teacher quality. Thirteen items: Q5, Q10, Q12, Q22, Q32, Q34, Q45, Q46, Q60, Q64, Q65, Q68, and Q94, had a mean squares fit statistic greater than 1.5. High fit statistics indicate un-modeled random noise, which can result from, misinterpretation of the question, or lack of fit of the item with other items on the assessment.

Table 2. *Rasch statistics for I-LAST items.*

Item	Difficulty	Error	Infit(MNSQ)	Outfit(MNSQ)	PtBisCorr
Q1	0.17	0.33	1.11	1.07	0.56
Q2	0.73	0.32	0.81	0.78	0.67
Q3	-1.08	0.34	0.68	0.57	0.76
Q4	0.50	0.32	1.01	1.01	0.57
Q5[a]	-0.10	0.25	1.83	2.38	0.34
Q6	-1.03	0.34	0.86	0.76	0.62
Q7	-0.28	0.30	0.79	0.82	0.71
Q8	-0.40	0.37	1.11	1.08	0.47
Q9	-1.24	0.34	0.61	0.52	0.78
Q10[a,b]	1.91	0.27	2.42	3.83	-0.05
Q11	-0.23	0.27	0.78	0.77	0.74
Q12[a]	-0.39	0.29	1.32	1.86	0.49
Q13	-0.67	0.29	0.78	0.81	0.74
Q14	-0.29	0.33	0.76	0.69	0.72
Q15	1.41	0.29	0.91	0.90	0.63
Q16	0.90	0.29	0.84	0.81	0.67
Q17	-0.10	0.34	0.96	0.92	0.61
Q18	-0.54	0.36	1.29	1.27	0.38
Q19	0.38	0.30	1.33	1.40	0.39
Q20	0.94	0.31	1.11	1.15	0.51
Q21	0.99	0.41	1.14	1.24	0.44
Q22[a]	1.74	0.29	1.91	2.13	0.17
Q23	0.39	0.27	0.79	0.80	0.72
Q24	-1.35	0.34	0.74	0.61	0.67
Q25	-1.40	0.34	0.85	0.75	0.61
Q26	1.84	0.32	0.75	0.70	0.69
Q27	-0.73	0.31	1.29	1.24	0.53
Q28	-0.61	0.36	0.94	0.82	0.63
Q29	0.19	0.29	0.92	0.92	0.66
Q30	-0.74	0.35	0.64	0.52	0.77
Q31	2.95	0.32	1.18	1.20	0.45
Q32	-0.59	0.32	1.54	1.38	0.43
Q33	-0.73	0.31	0.60	0.56	0.82
Q34[a]	3.01	0.32	1.22	2.41	0.36
Q35	0.19	0.29	1.01	1.03	0.62
Q36	-0.27	0.31	0.83	0.78	0.70
Q37	-0.86	0.34	0.86	0.72	0.68
Q38	0.82	0.32	0.85	0.84	0.64
Q39	1.33	0.37	1.02	1.11	0.50

Q40	0.29	0.38	0.99	0.90	0.50
Q41	0.36	0.31	0.84	0.79	0.67
Q42	-0.53	0.29	0.66	0.64	0.79
Q43	0.94	0.31	0.75	0.75	0.70
Q44	-0.37	0.32	0.63	0.63	0.77
Q45[a]	-0.81	0.27	2.16	4.12	0.19
Q46[a]	-0.62	0.28	1.74	2.38	0.34
Q47	-0.06	0.34	0.75	0.68	0.71
Q48	-0.43	0.30	0.74	0.74	0.75
Q49	-0.47	0.26	0.77	0.77	0.75
Q50	-0.54	0.32	1.08	1.12	0.59
Q51	-1.04	0.29	0.63	0.56	0.80
Q52	-0.60	0.27	1.07	1.28	0.63
Q53	-1.19	0.34	0.74	0.65	0.72
Q54	-0.70	0.27	0.75	0.70	0.77
Q55	-0.74	0.32	0.69	0.59	0.74
Q56	-0.85	0.32	0.71	0.68	0.75
Q57	-1.24	0.34	0.91	0.81	0.60
Q58	-0.93	0.30	0.97	1.32	0.65
Q59	-0.62	0.30	0.79	0.78	0.73
Q60[a]	1.35	0.27	1.72	1.86	0.26
Q61	-0.48	0.31	0.86	0.86	0.69
Q62	-0.28	0.30	0.66	0.63	0.78
Q63	0.75	0.34	0.80	0.77	0.67
Q64[a]	1.04	0.30	1.70	2.29	0.24
Q65[a]	0.16	0.29	1.60	1.70	0.30
Q66	2.94	0.36	1.09	1.19	0.42
Q67	0.61	0.28	1.16	1.27	0.52
Q68[a]	0.28	0.32	1.35	1.65	0.38
Q69	-0.32	0.33	0.73	0.77	0.72
Q70	-0.24	0.29	0.95	0.96	0.65
Q71	0.20	0.37	0.98	0.94	0.58
Q72	-0.03	0.35	1.04	0.99	0.58
Q73	-0.76	0.33	0.59	0.53	0.82
Q74	-0.62	0.30	0.75	0.79	0.75
Q75	1.33	0.30	0.99	0.98	0.59
Q76	-0.96	0.29	0.56	0.50	0.83
Q77	-0.49	0.28	0.82	0.79	0.73
Q78	-0.62	0.34	0.81	0.79	0.70
Q79	-0.85	0.32	0.67	0.66	0.78
Q80	-0.51	0.35	0.71	0.62	0.73
Q81	-0.25	0.33	0.86	0.85	0.63

Q82	0.17	0.33	0.76	0.71	0.72
Q83	-0.64	0.31	0.77	0.75	0.74
Q84	-0.67	0.35	0.88	0.75	0.66
Q85	-0.23	0.33	0.79	0.75	0.70
Q86	0.02	0.25	1.31	1.32	0.56
Q87	-0.14	0.31	0.99	0.89	0.62
Q88	1.45	0.33	0.87	0.81	0.62
Q89	0.50	0.32	0.65	0.62	0.75
Q90	-0.89	0.32	1.24	1.12	0.57
Q91	0.58	0.33	1.08	1.03	0.53
Q92	-0.74	0.32	1.10	1.07	0.53
Q93[a]	1.17	0.36	0.61	0.49	0.72
Q94[a]	0.53	0.26	1.70	1.79	0.34

[a]Misfit with Rasch model

[b]Negative Point Biserial Correlation

Rasch person reliability for the 94 items was measured at 0.980, which is identical to the Cronbach's alpha measure. Despite the fact that some of the items are misfitting with respect to the Rasch model, these reliability indices far exceed the standard of 0.85 that is often used for instruments intended to differentiate between individuals (Tennant & Connaghan, 2007), and reliability could likely be further improved by eliminating the misfitting items. Reliability of item measures was 0.880. This is lower than the person reliability value due to the fact that number of items exceeded the number of participants. However, the value of 0.880 is sufficient to differentiate between individual item measures, indicating that 46 participants were sufficient to provide separable Rasch item estimates to the end of establishing construct validity of the instrument (Linacre, 2012).

Dimensionality and PCA on Rasch Residuals

Based on the scree criterion, we can make a strong argument that the I-LAST is a unidimensional assessment. The first factor had an eigenvalue of 38.4 items of variance, whereas the second and third dimensions had eigenvalues of 4.8 and 4.4 items of variance, respectively. Using the guideline of Raiche (2005), PCA on Rasch residuals indicates a presence of four underlying dimensions not accounted for by the Rasch model (Table 3). Factors 1-4 had eigenvalues of 7.96, 7.51, 5.95, and 5.80 items of variance, respectively. These values are higher than 2, but nonetheless make up a small percentage of the total 94 items of variance. Seven items (Q2, Q4, Q11, Q15, Q16, Q23, and Q31) had negative loadings, and fourteen items (Q37, Q57, Q76, Q77, Q78, Q79, Q80, Q81, Q82, Q83, Q84, Q85, Q90, and Q93) had positive loadings onto Factor 1. The seven items with negative loadings measured ability to foster classroom efficiency and time on task (α_c = 0.906), and the 14 positively loaded items measured ability to encourage higher order thinking (α_c = 0.956). The nine items (Q16, Q23, Q29, Q30, Q35, Q36, Q39, Q48, and Q51) with negative loadings onto Factor 2 assess the ability to use data to monitor and structure student learning (α_c = 0.924). Six items with positive loadings onto Factor 2 (Q20, Q22, Q60, Q64, Q65, and Q68) assess a teacher's ability to use unstructured, constructivist methodologies (α_c = 0.799). Four items (Q11, Q25, Q52, and Q92) negatively loaded onto Factor 3, which measure effective communication in the classroom (α_c = 0.751). Six items had positive loadings (Q10, Q29, Q31, Q38, Q39, and Q40) onto Factor 3, and measured the ability to foster effective individualized communication (α_c = 0.781). Five items (Q3, Q30, Q33, Q44, and Q64)

negatively loaded onto Factor 4, and assessed the ability to facilitate learning through teacher-student interaction ($\alpha_c = 0.849$). Six items (Q5, Q18, Q24, Q41, Q52, and Q57) had positive loadings onto this factor, and assessed ability to communicate with students and colleagues in a clear and professional manner ($\alpha_c = 0.850$).

Table 3. *Item-factor loadings on Rasch residual dimensions.*

Dimension 1		Dimension 2		Dimension 4		Dimension 4	
Item	Loading	Item	Loading	Item	Loading	Item	Loading
Q11	-0.54	Q48	-0.54	Q92	-0.55	Q30	-0.53
Q31	-0.48	Q39	-0.52	Q11	-0.49	Q33	-0.50
Q15	-0.47	Q36	-0.50	Q25	-0.46	Q44	-0.50
Q16	-0.47	Q51	-0.47	Q52	-0.43	Q64	-0.49
Q23	-0.43	Q35	-0.46			Q3	-0.43
Q4	-0.41	Q16	-0.45				
Q2	-0.40	Q29	-0.42				
		Q23	-0.40				
		Q30	-0.40				
Q57	0.41	Q20	0.44	Q29	0.42	Q41	0.40
Q90	0.41	Q68	0.52	Q31	0.42	Q5	0.44
Q84	0.45	Q64	0.56	Q38	0.44	Q24	0.52
Q93	0.46	Q65	0.60	Q10	0.52	Q52	0.53
Q77	0.47	Q22	0.62	Q39	0.58	Q18	0.54
Q85	0.48	Q60	0.66	Q40	0.65	Q57	0.62
Q76	0.49						
Q37	0.50						
Q78	0.52						
Q81	0.52						
Q79	0.53						
Q83	0.55						
Q80	0.60						
Q82	0.63						

Applications of the I-LAST for Assessing Best Practice in Diverse Environments

The I-LAST is intended for use by teaching internship supervisors and administrators to assess aspects of teaching quality identified as important predictors in a multi-case study (Kingsley, 2007) and seen in best practice (Zemelman, et al., 2005). These themes go beyond educational attainment and certification as criteria for quality teachers; the focus is on observed teaching practices. Although unobservable factors such as education and experience may play a part in improving teacher quality, these are not directly addressed by the I-LAST. This makes the I-LAST potentially useful for measuring the improvement of instructors as they progress through their internships and into their careers, and as they increase participation in professional development programs. The I-LAST may be helpful to administrators in evaluating the impact of money spent towards teacher education and development, and to researchers attempting to address a number of important questions, including: What differential effects on teaching best practice do different types of college degree and induction programs have?

What types of pre-service education programs and structures are most effective? To what extent do pre-service teachers progress from novice to expert levels through internship experiences?

Choosing items by topic. Unlike many measurement tools, the I-LAST can be characterized as a unidimensional item battery. This presents a variety of opportunities to professionals needing diverse quantitative measures for teaching best practice in clinical and research settings. While use of all 94 items will provide a score with excellent reliability, in the interest of efficiency, we recommend that subsets of items be used to develop a shorter assessment tailored to specific needs.

Table 4. *Items by topic theme and difficulty level (δ).*

Level	Mgmt	δ	StAcc	δ	Assmt	δ	TchAcc	δ	IndivIns	δ	Literacy	δ
High	Q10[a]	1.91	Q26	1.84	Q34[a]	3.01			Q66	2.94	Q88	1.45
	Q15	1.41	Q22[a]	1.74	Q31	2.95			Q60[a]	1.35	Q75	1.33
					Q39	1.33			Q64[a]	1.04	Q93[a]	1.17
Moderate-High	Q16	0.90	Q21	0.99	Q43	0.94			Q63	0.75	Q91	0.58
	Q2	0.73	Q20	0.94	Q38	0.82			Q67	0.61	Q94[a]	0.53
	Q4	0.50	Q23	0.39	Q41	0.36			Q68[a]	0.28	Q89	0.50
	Q1	0.17	Q19	0.38	Q40	0.29			Q71	0.20	Q82	0.17
			Q29	0.19	Q35	0.19			Q65[a]	0.16	Q86[a]	0.02
Moderate-Low	Q5[a]	-0.10	Q18	-0.54	Q36	-0.27	Q47	-0.06	Q72	-0.03	Q87	-0.14
	Q17	-0.10	Q28	-0.61	Q44	-0.37	Q48	-0.43	Q70	-0.24	Q85	-0.23
	Q11	-0.23	Q27	-0.73	Q42	-0.53	Q49	-0.47	Q62	-0.28	Q81	-0.25
	Q7	-0.28			Q32[a]	-0.59	Q50	-0.54	Q69	-0.32	Q77	-0.49
	Q14	-0.29			Q33	-0.73	Q52	-0.60	Q61	-0.48	Q80	-0.51
	Q12[a]	-0.39			Q30	-0.74	Q46[a]	-0.62	Q73	-0.76	Q74	-0.62
	Q8	-0.40			Q37	-0.86	Q59	-0.62			Q78	-0.62
	Q13	-0.67					Q54	-0.70			Q83	-0.64
							Q55	-0.74			Q84	-0.67
							Q45[a]	-0.81			Q92	-0.74
							Q56	-0.85			Q79	-0.85
							Q58	-0.93			Q90	-0.89
											Q76	-0.96
Low	Q6	-1.03	Q24	-1.35			Q51	-1.04				
	Q3	-1.08	Q25	-1.40			Q53	-1.19				
	Q9	-1.24					Q57	-1.24				

[a]Poor fit with Rasch partial credit model

For example, PCA on Rasch residuals selected eight clusters of items related to more specific aspects of teaching best practice, including abilities to communicate clearly and professionally with students and faculty, to make data-driven decisions in the classroom, and to implement appropriate learning and assessment strategies. These item clusters have acceptable reliabilities between 0.751 and 0.956. Using

the argument that the I-LAST is a unidimensional assessment, items can also be chosen qualitatively based on the best practice category of interest. For example, items clustered by topic subtheme (Table 4) have measurement reliabilities between 0.843 and 0.946 (Table 5), which indicates sufficient precision for individual comparisons.

Table 5. *Cronbach's alpha reliability measures of item clusters by topic theme and difficulty level.*

Difficulty	Mgmt	StAcc	Assmt	TchAcc	IndivIns	Literacy	Total
High	0.042	0.404	0.653	NA	0.666	0.634	0.790
Moderate-High	0.819	0.730	0.788	NA	0.680	0.690	0.922
Moderate-Low	0.834	0.450	0.887	0.898	0.881	0.938	0.974
Low	0.793	0.792	NA	0.810	NA	NA	0.905
Total	0.914	0.843	0.920	0.927	0.878	0.946	0.980

Items as measures for development. Models of teacher development (Fuller, 1969; Berliner, 1988, Kagan, 1992) indicate that teachers progress through stages as they transition through the profession. These stages can be observed through implementation of best practice. The initial stages focus on survival, where teachers struggle to balance the practical need to manage the classroom with the emotional need to be liked by students. Middle stage concerns include optimizing lessons, presenting information clearly, and organizing the classroom for efficiency. Student-centered practice takes precedence in the higher stages of teacher development (Fuller, 1969). In observance of these stages, subsets of items can be chosen based on their difficulty, thus increasing suitability of the assessment to a particular stage of a teacher's development. A Wright map of intern and item measures on the same latent continuum (Figure 1) gives qualitative insight into how the distributions compare. The easiest items on the assessment included Q9 (The intern treats all students in a fair and equitable manner), Q24 (The intern interacts with students in a professional manner), Q25 (The intern interacts with students in a compassionate manner), Q53 (The intern uses objectives to guide lesson planning) and Q57 (The intern regularly attends planning meetings with the host teacher, including grade level meetings, committee meetings, and IEP meetings as appropriate) indicating that building a community of support through positive interactions with students and colleagues is a primary focus of an early teaching career. The Wright map shows that all interns had ability measures greater than the difficulty measure of these items, indicating that all interns were likely to be rated highly for these practices. By contrast, the most difficult items included Q31 (The intern encourages students to give each other feedback on assignments), Q34 (The intern often gives students tests or quizzes), and Q66 (The intern encourages students to develop their own strategies for staying on task), suggesting that routine summative assessment and use of student-centered methodologies are not focus practices for most teachers in the induction stage. Only 8 of the 46 interns had measures at or above the difficulty of these items, indicating that 83% of the interns had not reached this level of best practice.

To the end of partitioning the I-LAST in terms of difficulty (see Tables 4 and 5), we present a four-tiered approach based on the idea that logit measures are normalized to a mean of 0. Items with logit measures above 1 can be considered "high difficulty," between 0 and 1 can be considered "moderate-high," between -1 and 0 can be considered "moderate-low," and below -1 can be considered "low." In this sample of participants, selection of high difficulty items yielded a measurement reliability of 0.790. However, in light of the item information construction of reliability, this value would likely be significantly higher if these items were used to evaluate more experienced teachers. Reliability values for item clusters at the lower levels were above 0.9. These values indicate that shorter assessments tailored to a wide variety of specific practices and developmental levels can be constructed from this item bank without significant attenuation of reliability.

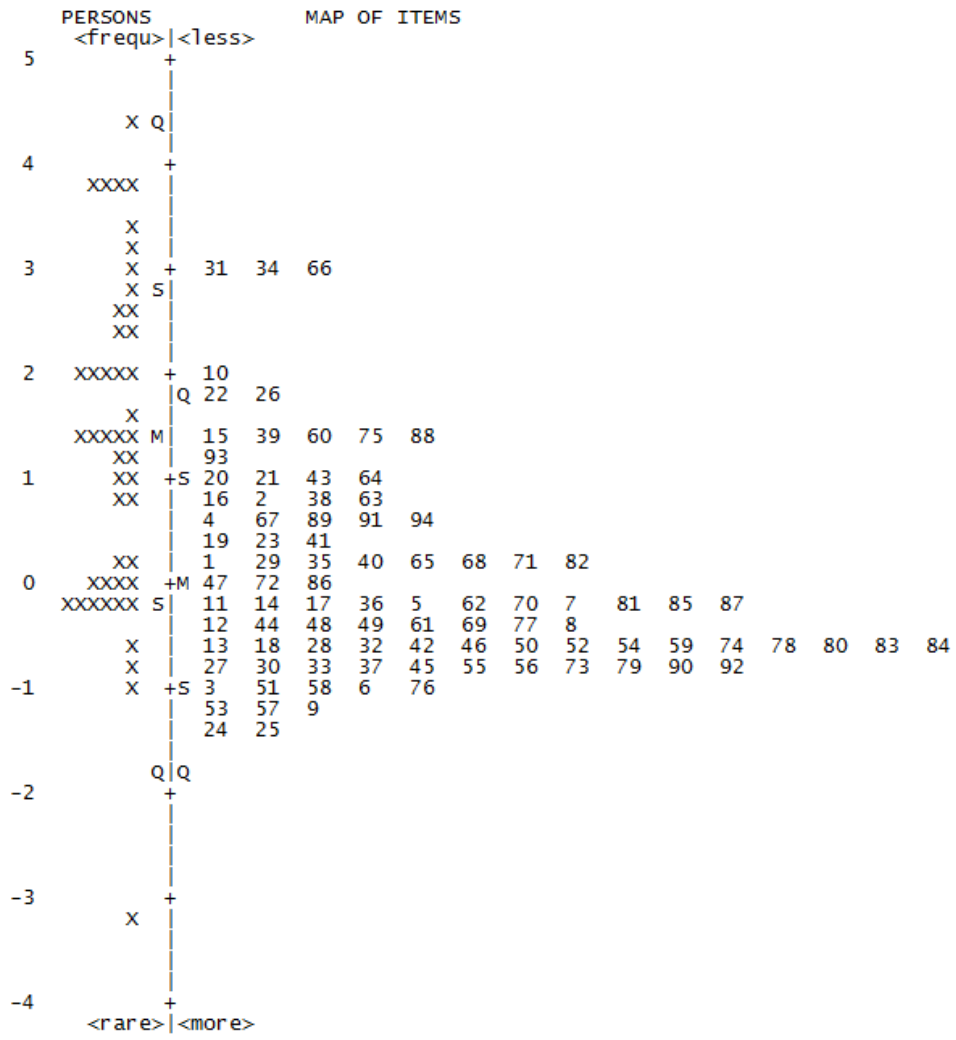

Figure 1. Wright map of person and item measures along the latent best practice continuum

Potential Research and Uses for the I-LAST to Advance the Discipline

Validation of the I-LAST instrument invites a number of research opportunities to advance the field. The practical framework behind the I-LAST, combined with the instrument's diversity and precision, makes it potentially useful for evaluating teachers in many settings where it is necessary to quantify teaching best practice, including teaching internship evaluations and yearly evaluations for practicing teachers. Items from the I-LAST can be used straightforwardly by school principals and internship supervisors during scheduled observations, bringing the opportunity to assess teachers longitudinally as they progress through their internships and into their careers. To this end, use of increasingly difficult items as the teacher develops will not only give more reliable scores, but will also give a score targeted to his/her expected stage of development and best practices expected at that stage. A number of important questions can be answered through successive evaluations. How does best practice change as

teachers become self authored and move along Fuller's scale? Along this line, it may also be useful to compare teachers' self evaluations to external evaluations to quantify how consistency changes as teachers develop. Do teachers become more self aware as they progress along Fuller's scale? Another question that needs to be addressed is how improvement in teaching quality over time differs by school environment and internship structure. What specific aspects of internship experiences and teaching environments help or hinder development? The relationship between teaching quality, school work environment, and teacher attrition can also be explored Do teachers who move through Fuller's stages more quickly tend to stay longer in the profession? Questions such as these can be addressed quantitatively through use of the I-LAST in future research.

In order to make constructive use of the I-LAST for research on teaching best practice across disciplines and grade levels, it is helpful to acknowledge the limitation of a highly specific validation sample, namely elementary literacy instructors in a year-long internship program. Consequently, item and test reliability measures may differ when applied to more experienced teachers, teachers at different grade levels, or teachers of other topic areas such as science or mathematics. Recognizing that best practice can differ between subjects, these differences are far outweighed by best practices common to the teaching profession as a whole. We believe items Q1-Q73 can be used nearly verbatim in many situations since these elements of best practice can be applied to most classrooms. However, the literacy-focused items would need to be reworded to accommodate specific subject areas. For example, item Q74, "The intern incorporates group reading techniques (books clubs, literature circles, partner reading)," could be reworded, "The intern engages students in cooperative inquiry-based activities (demonstrations, discussions, and experiments)" to accommodate a science context. Due to the diversity of the I-LAST, we are confident that it will provide a meaningful system of measurement across a wide range of teaching environments and developmental stages. However, we recommend that researchers using the instrument in a novel context quantify the reliability of items before using scores to draw statistical conclusions.

Helping Teachers Last in the Profession

At this time, the common paradigm for solving our teacher shortage is to put more teachers into the field through reduction of qualification requirements and romanticizing the profession. In the United States, programs such as Teach for America (TFA) take this approach. Not surprisingly, the attrition rate in such programs is very high, up to triple the rate of college-recommended teachers. Only 15-20% of TFA teachers remain in the profession after four years (Helig, 2010). Ingersoll and Smith (2003) suggest that pouring many under-qualified teachers into the system is the wrong approach in that it does not get to the root of the problem, which lies in the labor-intensive, thankless working conditions that many teachers face, the top three of which include low salary, student discipline problems, and poor administrative support. The impact of these problems is especially great during the first years of teaching since many teachers have not had the time to self author themselves and formulate definitions of success independent of outside circumstances and criticism which are largely uncontrollable (Romine, West, & Volkmann, 2010). A quantitative methodology using an instrument such as the I-LAST increases the efficiency and precision of teacher assessment, therefore allowing a busy administration to identify and address new teachers' strengths and deficiencies early to the end of giving appropriate feedback and assistance.

The I-LAST will likely show its greatest utility for measuring new teachers' growth in the context of induction programs. A significant aim in teacher induction programs is to prepare teachers who will last in the profession, especially since certification and experience are significant contributors to teacher quality (Darling-Hammond et al. 2005). That European policy documentation (ETUCE, 2008) also addresses this indicates that this goal is common internationally. To this end, it is important to put teachers in settings which support success, and the I-LAST can be used to quantify the differential impacts of these settings based on comparative analyses of growth. Certain classroom

environments have shown favorability toward the goal of teacher retention. For example, Finn and Achilles (1999) discuss the positive effects of small classes on management. In addition, schools supporting student-centered learning approaches tend to encourage students to manage themselves (Martin, 2004; Savery, 2006), taking pressure off of the instructor. Since focus on student-centered practices is characteristic of Fuller's higher developmental stages, it is a tall order to expect a beginning teacher to master these during the time allotted for an internship even in a school environment that supports these practices. Many schools in the greatest need of qualified teachers, including TFA schools, have the most inhospitable settings for a new teacher learning to effectively implement student-centered instructional approaches. This makes it extremely important to quantify the various impacts that these environments have on the growth of interns and beginning teachers. Data provided by the I-LAST can be used to target specific areas of assistance that new teachers need and to inform the development of ways to help teachers grow despite non-ideal school settings. Efficient routine monitoring and providing of well-targeted, constructive support has the potential to mitigate teacher attrition in these contexts.

Conclusion

The definition of success for a new teacher often relies upon positive feedback from students, mentors, and colleagues. This makes targeted, constructive evaluation and feedback essential in the induction stages. In light of this need, the I-LAST addresses a significant gap by allowing for efficient and precise scoring of teaching best practice based on a general and practical framework. In addition to presenting details on the development of the I-LAST and evidence for its instrument- and item-level validity, we also discussed how the I-LAST can be used to meet a wide variety of needs related to assessing teaching best practice. While using all 94 items on the I-LAST will give an exceptionally reliable score for teaching best practice, the strength of this measurement tool lies in its diversity; it allows a supervisor or researcher to obtain reliable measures for specific dimensions of teaching best practice that are of direct interest. Dimensions of interest could be identified based on research questions, prior observations of teaching, or concordance with the particular mission of a teacher induction program. That items span the range of Fuller's scale will make the I-LAST especially useful for assessing growth of new teachers as they transition from the pre-service stage through the initial years of their teaching careers. In this way, the I-LAST shows potential as an additional method researchers can use to measure the differential effects of various types of induction programs and instructional settings on the growth of new teachers along Fuller's scale. What types of induction programs are most effective in helping new teachers transition along Fuller's scale? What types of school environments are most conducive to the development of new teachers? What interventions are needed to move new teachers along Fuller's scale despite less than ideal school and classroom settings? We hope the I-LAST will be used in future research to find definitive answers to these questions which may lead to data-driven decisions aimed at reducing teacher attrition during the induction years.

References

Akiba, M., LeTendre, G. K., & Scribner, J. P. (2007). Teacher quality, opportunity gap, and national achievement in 46 countries. *Educational Researcher,* 36(7), 369-387.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? Medical Care, 42(1), Supplement: Applications of Rasch Analysis in Health Care, 17–116.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliff s, NJ: PrenticeHall.

Berliner, D.C. (1988). Implications of studies on expertise in pedagogy for teacher education and evaluation. In *New directions for teacher assessment* (Proceedings of the 1988 ETS Invitational Conference, pp. 39-68). Princeton, NJ: Educational Testing Service.

Bitner, T., & Kratzner, R. (1995). A Primer on Building Teacher Evaluation Instruments.

Burnett, P. C., & Meacham, D. (2002). Measuring the quality of teaching in elementary school classrooms. *Asia-Pacific Journal of Teacher Education,30*(2), 141-153.

Capie, W. (1978). Teacher Performance Assessment Instruments: Plans for Practice Rating.

Cattell, R.B. (1966). The meaning and strategic use of factor analysis. In: Handbook of Multivariate Experimental Psychology. Chicago: Rand McNally.

Cochran-Smith, M., & Fries, M. K. (2001). Sticks, stones, and ideology: The discourse of reform in teacher education. *Educational Researcher, 30*(8), 3-15.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd. Ed.) Lawrence Erlbaum and Associates.

Commission on No Child Left Behind. (2007). Beyond NCLB: Fulfilling the promise to our nation's children. Washington, DC: The Aspen Institute.

Darling-Hammond, L. (2003). Keeping good teachers: Why it matters, what leaders can do. *Educational Leadership, 6*(8), 6-13.

Darling-Hammond, L., & Youngs, P. (2002). Defining "Highly Qualified Teachers": What does "Scientifically-Based Research" actually tell us? *The Educational Researcher, 31*(9), 13-25.

Darling-Hammond, L., Holtzman, D.J., Gatlin, S.J. and Heilig, J.V. (2005). Does

teacher preparation matter? Evidence about teacher certification, Teach for America, and teacher effectiveness. *Education Policy Analysis Archives, 13*(42), Retrieved 9/21/2011 from http://epaa.asu.edu/epaa/v13n42.

De Fina, A. A. (1992). *Portfolio Assessment: Getting Started. Teaching Strategies*. Scholastic Inc. Jefferson City, Missouri.

D'Onofrio, C. N. (1989). The use of self-reports on sensitive behaviors in health program evaluation. *New Directions for Program Evaluation, 1989*(43), 59-74.

Duffield, S.K. (2005). Swimming in the water: Immersing teacher candidates in the environment of a school. *Current Issues in Education* [on-line], *8 (11).*

ETUCE. (2008). *Teacher Education in Europe: an ETUCE policy paper*. Brussels, Belgium.

Finn, C. E., Jr. (2003). High Hurdles. *Education Next* 3(2): 62–67.

Fuller, F. (1969). Concerns of teachers: A developmental conceptualization. *American Educational Research Journal, 6*, 207-226.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation.*Journal of educational psychology*, *76*(4), 569.

Glaser, B., & Strauss, A. (1967). *The Discovery of Grounded Theory*. Chicago, IL: Aldine.

Goldhaber, D. (April, 2006). *Everybody's Doing It, But What Does Teacher Testing Tell Us About Teacher Effectiveness?* Center on Reinventing Public Education. Paper presented at the AERA annual meeting.

Goodlad, J. (1994). *Educational renewal*. San Fransisco, CA: Jossey-Bass.

Goodlad, J. (1990). *Teachers for our nation's schools.* San Francisco, CA: Jossey-Bass.

Gordon, R., T.J. Kane, and D.O. Staiger. (2006). Identifying effective teachers using performance on the job. Washington, DC: The Brookings Institution.

Helig, J. (June, 2010). *Teach for America: A review of the evidence.* Great Lakes Center for Educational Research and Practice: Lansing, MI.

Hoffman, J., Roller, C., Maloch, B., Sailors, M., Duffy, G., and Beretvas, S.N., The National Commission on Excellence in Elementary Teacher Preparation for Reading Instruction. (2005). Teachers' preparation to teach reading and their experiences and practices in the first three years of teaching. *The Elementary School Journal*, *105*(3), 267-287.

The Holmes Group. (1986). *Tomorrow's teachers*. East Lansing, MI: Author.

Ingersoll, R. & Smith, T. (2003). The wrong solution to the teacher shortage. *Educational Leadership, 60*(8), 30-33.

Kagan, D. (1992). Professional growth among preservice and beginning teachers. *Review of Educational Research, 62*(2), 129-169.Linacre, J.M. (2010). Winsteps® (Version 3.70.0) [Computer Software]. Available from www.winsteps.com.

Kingsley, L. H. (2007). An examination of how extended field experiences, integrated coursework, and school partnerships influenced the development of four first year teachers' literacy beliefs and practice (Doctoral dissertation, University of Missouri--Columbia).

Linacre, J.M. (2012). Winsteps® Rasch Tutorial 3. Available from http://www.winsteps.com/a/winsteps-tutorial-3.pdf.

Martin, S. (2004). Finding balance: Impact of classroom management conceptions on development of teacher practice. *Teaching and Teacher Education,* 20, 405-422.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.

Mullen, C. & Farinas, J. (2003). What constitutes a highly qualified teacher: A review of teacher education standards and trends. *Teacher Education and Practice*, *16*(4), 318-330.

NCATE. (2006). *What makes a teacher effective? A summary of key research findings on teacher preparation.* Washington DC: National Council for Accreditation of Teacher Education.

No Child Left Behind. (2001). Public Law, 107-110.

Penuel, W. R., Boscardin, C. K., Masyn, K., & Crawford, V. M. (2007). Teaching with student response systems in elementary and secondary education settings: A survey study. *Educational Technology Research and Development*, *55*(4), 315-346.

Raiche, G. (2005). Critical eigenvalue sizes in standardized residual principle components analysis. *Rasch Measurement Transactions*, 19, 1012.

Romine, W., West, A. & Volkmann, M. (March, 2010). *Expectations to Success: The Contrasting Journeys of a Teacher and His Coach*. Paper presentation at the National Association for Research in Science Teaching annual meeting, Philadelphia, PA.

Ross, J. A., McDougall, D., Hogaboam-Gray, A., & LeSage, A. (2003). A survey measuring elementary teachers' implementation of standards-based mathematics teaching. *Journal for Research in Mathematics Education*, 344-363.

Savery, J. (2006). Overview of problem-based learning: definitions and distinctions. *The Interdisciplinary Journal of Problem-Based Learning*, *1*(1), 9-20.

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G., & Sharpe, M. (2008). Rasch fit statistics and sample size considerations for polytomous data. *BMC Medical Research Methodology*, *8*(1), 33.

Smith, T. & Ingersoll, R. (2004). What are the effects of induction and mentoring on beginning teacher turnover? *American Educational Research Journal*, *41*(3), 681-714.

Stulac, J. F. (1982). Procedure and Results of Measurement in the South Carolina Assessments of Performance in Teaching.

Tennant, A., & Connaghan, P.G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care and Research*, 57(8), 1358–1362.

Thomas, J. A., & Montgomery, P. (1998). On becoming a good teacher: reflective practice with regard to children's voices. *Journal of Teacher Education*, *49*(5), 372-80.

Thomas, T. P., & Schubert, W. H. (2001). Reinterpreting teacher certification standards. In J. K. Kincheloe & D. Weil (Eds.), *Standards and schooling in the United States: An encyclopedia, 1* (pp. 229-243). Santa Barbara, CA: ABC-CLIO.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and teacher education*, *17*(7), 783-805.

Wenglinsky, H. (2000). *How teaching matters: bringing the classroom back into discussions of teacher quality.* Princeton, NJ: Educational Testing Service.

Wolf, D. P. (1989). Portfolio assessment: Sampling student work. *Educational leadership*, *46*(7), 35-39.

Wright, B.D., & Stone, M.A. (1979). Best test design. Chicago, IL: MESA Press.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, 3(1), 3-24.

Wright, B.D. (1997). A history of social science measurement. *Educational Measurement Issues and Practice*, 16, 33-45.

Zeichner, K. (2010). Rethinking the connections between campus courses and field experiences in college and university based teacher education. *Journal of Teacher Education*, 61: 89-99.

Zemelman, S., Daniels, H., & Hyde, A. (2005). *Best practice: Today's standards for teaching and learning in America's schools.* Portsmouth, NH: Heinemann.

Appendix: The Item-Level Assessment of Teaching (I-LAST)

Theme	Item	Statement
Management	Q1	The intern identifies and addresses student behavior.
	Q2	The intern keeps instruction moving without allowing off-task interruptions and diversions to interfere with instruction.
	Q3	The intern attempts to engage all students in class.
	Q4	The intern efficiently transitions students from task to task.
	Q5	The intern's instructions to students are unclear and ambiguous.
	Q6	The intern recognizes and reinforces positive behavior.
	Q7	The intern recognizes negative behaviors and uses appropriate interventions.
	Q8	The intern directs disciplinary consequences to individuals, not the whole group.
	Q9	The intern treats all students in a fair and equitable manner.
	Q10	The intern expects all students to achieve at the same level.
	Q11	The intern effectively uses nonverbal cues (including eye contact, moving closer to the student, and quickly returning the class to on-task behavior) in response to inappropriate behavior.
	Q12	The intern effectively uses a problem-centered approach when conferencing with students about their behavior.
	Q13	The intern works with the host teacher to initiate communication with parents concerning student behavior.
	Q14	The intern directly addresses student misbehavior according to classroom guidelines and school expectations.
	Q15	The intern, when asking students to stop the undesired behavior, does not make consequences for continuing the behavior clear.
	Q16	The intern consistently follows through with consequences according to set classroom procedures when inappropriate behavior continues.
	Q17	The intern spends excessive time starting and ending class.
Student Acc.	Q18	The intern sets specific due dates and time limits for work.
	Q19	The intern invokes consequences for students who turn in late work.
	Q20	The intern asks students to articulate how they know what they know.
	Q21	The intern encourages students to hold each other accountable for good behavior.
	Q22	The intern allows students to suggest their own consequences for poor behavior.
	Q23	The intern's expectations for students need to be higher.
	Q24	The intern interacts with students in a professional manner.
	Q25	The intern interacts with students in a compassionate manner.
	Q26	The intern encourages students to check their own work and give themselves oral and/or written feedback.
	Q27	The intern's lessons provide no clear learning objectives.
	Q28	The intern assesses and grades students based on achievement of classroom learning objectives.

	Q29	The intern consistently gives specific written comments on student work.
Assessment	Q30	The intern actively encourages students to ask questions during class.
	Q31	The intern encourages students to give each other feedback on assignments.
	Q32	The intern seldom incorporates student questions into class discussion.
	Q33	The intern observes students working and uses these observations to inform his/her teaching.
	Q34	The intern often gives students tests or quizzes.
	Q35	The intern gives students specific oral and/or written feedback on tests and quizzes.
	Q36	The intern addresses deficiencies he/she sees on tests, quizzes, and assignments to clarify student misunderstandings.
	Q37	The intern considers a variety of measures to assess learning, not just tests and quizzes.
	Q38	The intern puts in effort to make sure tests, quizzes, and assignments are free from gender or cultural bias.
	Q39	The intern asks a variety of types of questions (i.e. true-false, multiple choice, short answer, essay, etc.) on tests and quizzes.
	Q40	The intern makes tests and quizzes appropriate for learners.
	Q41	After the intern collects a grade on a test, quiz or assignment, he/she does not utilize the information for future instruction.
	Q42	The intern uses one-on-one conferences with individual students to assess their learning.
	Q43	The intern conferences with student groups to assess student learning.
	Q44	The intern regularly makes use of formative assessment techniques to inform instruction.
Teacher Acc.	Q45	The intern's lessons and instructional plans are seldom linked to Missouri's Grade Level Expectations (GLEs) or Course Level Expectations (CLEs).
	Q46	The intern regularly records numerical scores on assignments, quizzes, and tests as directed by the cooperating teacher.
	Q47	The intern regularly provides written documentation of students' progress.
	Q48	The intern uses data from previous assignments as a guide for future classroom instruction and/or lessons when given access to them.
	Q49	When students do poorly in class, the intern considers things he/she could do differently.
	Q50	The intern seeks administrative or outside assistance as needed.
	Q51	The intern routinely discusses classes with the cooperating teacher.
	Q52	When given the opportunity, the intern does not collaborate with other teachers.
	Q53	The intern uses objectives to guide lesson planning.
	Q54	The intern seems reflective and is eager to improve his/her teaching as demonstrated by conferences and written documents such as journals.
	Q55	The intern grades student work in a timely manner.
	Q56	The intern engages him/herself in professional development activities.
	Q57	The intern regularly attends planning meetings with the host teacher (including grade level meetings, committee meetings, and IEP meetings as appropriate)

	Q58	The intern develops lesson plans in a timely manner.
	Q59	The intern regularly self assesses his/her strengths and weaknesses and sets goals for improvement.
Indiv. Inst.	Q60	Learners are free to move from one topic to another as needed, without regard for a predetermined sequence.
	Q61	The intern uses organizers such as outlines and KWL charts to help students draw upon prior knowledge.
	Q62	The intern encourages students to reflect upon and evaluate their own learning.
	Q63	The intern requires students to learn at the same pace.
	Q64	The intern encourages students to set their own learning goals for tasks they wish to complete.
	Q65	The intern encourages students to set their own goals related to class behavior.
	Q66	The intern encourages students to develop their own strategies for staying on task.
	Q67	The intern needs to incorporate a greater number of instructional strategies in his/her class.
	Q68	The intern encourages students to develop their own questions and seek their own answers.
	Q69	The intern encourages alternative explanations.
	Q70	The intern takes charge of setting goals for tasks students are to complete in collaboration with the mentor teacher.
	Q71	Students have no say in the course of action they will take in meeting task goals.
	Q72	The intern encourages students to think about their thinking.
	Q73	The intern promotes opportunities for students to engage in learning through a variety of formats/styles.
Literacy	Q74	The intern incorporates group reading techniques (books clubs, literature circles, partner reading)
	Q75	The intern models living a writerly life.
	Q76	The intern utilizes a variety of teaching groups (one-on-one, small group, whole class) to teach literacy
	Q77	The intern models living a readerly life.
	Q78	The intern encourages students to explore a variety of genres.
	Q79	The intern encourages students to make connections between reading and writing.
	Q80	The intern conducts literacy-targeted assignments for instructional purposes.
	Q81	The intern effectively uses cuing strategies for non-fluent readers.
	Q82	The intern meaningfully teaches comprehension strategies.
	Q83	The intern introduces a variety of reading texts and materials.
	Q84	The intern focuses on the meaning of texts and materials being discussed.
	Q85	If the internship supervisor allows, students often get to write their own stories in class.
	Q86	The intern seldom allows students to discuss the stories they read with each other.
	Q87	Students get little or no practice spelling words based on sounds.
	Q88	Students often practice sounding words out following practiced decoding strategies.
	Q89	The intern and students often work together to learn new words.

Q90 The intern encourages students to write about their life experiences.

Q91 The intern gives students adequate time each week to write in journals and/or writer's notebooks in response to what was read in guided or independent reading.

Q92 The intern seldom reads stories aloud to students.

Q93 If the internship supervisor allows, the intern gives students adequate time to read on their own.

Q94 The intern seldom engages students in word games.