Purpose

Recent research has indicated enormous variation across and within teachers in the nature and quality of their instruction (e.g., Rowan & Correnti, 2009). Though there is broad agreement that the quality of this instruction should influence what students learn, literature is sparse in finding and demonstrating what teacher/instructional quality is and how it actually matters (e.g. Nye, Konstantopoulos & Hedges, 2004). To better understand instruction, a number of recent studies have developed systems for classroom observation focusing on small, low inference instructional actions (IAs) in order to effectively chronicle instruction (e.g., Cirino et al., 2007; Foorman et al., 2006; Pianta et al., 2008). Though such systems provide an advanced mechanism to chronicle the observed instruction, their ability to measure teachers' instruction is complicated by how the actions are later combined to represent meaningful constructs of instruction.

Combining IAs to form measures of practice have predominantly relied on conventional approaches akin to summing the observed IAs across lessons (e.g. Cirino et al., 2007; Foorman et al., 2006; Pianta et al., 2008). However, the simplicity of such approaches is frequently coupled with an inability to differentiate among teachers as summative scaling tends to mask both subtle and complex differences. For instance, summative scaling (e.g. simple frequency counts of how many IAs were used) allows routine IAs (e.g. giving directions) to dominate the scale while severely discounting the contribution of rare IAs (e.g. fostering discussion). Similarly, forms of summative scaling that attempt to alleviate scale domination by routine IAs by focusing on only rare IAs can also constrain the data. Although rare IAs may better differentiate some teachers, measuring the existence of rare IAs in isolation may not be indicative of teachers' instruction and likely does not uniquely characterize effective instruction. Rather, recent studies have suggested that effective teachers were those that used a skillful blend of instruction (e.g. McGhie-Richmond, Underwood, and Jordan, 2007). Further, count scales assume that simply doing more is better. While dense instruction may be an effective component of instruction (i.e., through accelerating achievement), dense instruction might also be uncoordinated or inappropriate, quite likely not contributing to students' learning. Similarly, summative scaling additionally ignores the contexts in which the instruction was delivered, thereby potentially masking appropriateness and targeted instruction. Evidence that observed elements of instruction have not been combined to form meaningful characterizations of practice is demonstrated by results of studies that have not found links between observed practice and student achievement (e.g. Foorman et al., 2006; Pianta et al., 2008). This absence, in spite of more precise substantive theories and developed observations systems, raises the question of whether what seems to be a lack of relationship between instruction and achievement is, at least in part, a methodological consideration rather than ineffective observation systems or imprecise substantive theories.

In this study we extended research on instruction by developing a method to effectively measure instruction across multiple dimensions of instruction, multiple lessons and multiple observations. Specifically, to attend to unique features of instructional data, we developed a multivariate, multilevel Rasch (MMR) model to characterize the observed practices. To assess the empirical effectiveness of this approach, we compared the conditional hierarchical associations of the MMR derived measures and conventional measures (summative scaling) of practice with student achievement in third grade reading comprehension.

Background

Recent observation systems differ from previous systems in that they focus on the nature of instruction in lessons and separate them by purpose (e.g. Authors, 2009). A lesson is a conceptual unit of instruction that teachers use to partition their instructional time for specific

purposes. In planning lessons, teachers are likely to consider the content and instructional objectives that have been covered previously, an (in)formal evaluation about their students' relevant skills, the challenges of the new content and so on. To better understand instruction, observation systems have focused inquiry on teachers' IAs (e.g. giving directions, assessing student work) within lessons to capture the content, style and delivery of each lesson. Collectively, the presence/absence of these targeted activities has been used to characterize the nature of instruction. Despite the varying IAs targeted, studies have consistently summarized the collective IAs over multiple lessons using simple summative scaling (e.g. average of the number of IAs used each lesson). However, the limitations of summative scaling in characterizing instruction are apparent in a number of recent studies as they have consistently failed to link reading instruction to students' achievement (e.g. Foorman et al., 2006; Pianta et al., 2008).

Rather than restrict the scope of study to simple sums and solitary aspects of instruction, such as routine/rare IAs, that likely do not capture the fullness and complexity of instruction, we developed a hypothesis that the method of synthesizing IAs is just as important as the quality and focus of the observation system. Further, we hypothesized that the package of instructional practices employed and their effectiveness will depend on a number of contextual, background and temporal factors surrounding the instruction. That is, rather than adopt a view that emphasizes stable differences between teachers in a global propensity to employ certain IAs, we adopted the view that the propensity to employ certain IAs is contextually (e.g. classroom ability) and temporally situated (e.g. lesson duration). As a result, the propensity to employ certain instructional practices will likely depend jointly on factors such as the characteristics of the lesson, teacher and students being taught. Under this view, teachers with similar stable predictors of instruction, such as educational backgrounds, may still vary substantially in their observed practice or propensity to employ an instructional practice. Such residual variation would be a function of contextually and temporally varying factors such as the students' prior achievement trajectories, lesson duration and the time of year the lesson took place. Further, summative scaling often assumes that instructional practice is best characterized using a single dimension. However, recent studies have also suggested the importance of looking at multiple interrelated components of instruction within specific domains (e.g. Pressley et al., 1996; McGhie-Richmond et al., 2007). It follows that a more complex model and view of instruction would consider the IAs a teacher makes use of as a multidimensional system in which contextual and temporal factors impact specific dimensions in diverse ways.

Method

To assess the abilities of summative scaling and MMR in characterizing the nature and quality of instruction, we constructed two sets of empirical indices. The first was the conventional summative approach. For the second set, we developed a multivariate, multilevel Rasch model to measure instruction. Such an approach allowed us to combine IAs across lessons and time to develop an effective and meaningful empirical metric of observed instruction that more closely aligns with salient features of instruction than conventional methods. We briefly describe several salient features and assumptions of our MMR model before comparing indices.

Measurement. The first layer in our approach is a measurement model describing the probabilities of employing each IA and describing how the IAs are related to the latent dimensions. We viewed each observed dimension of instruction as a continuous latent construct and, by making use of observed IAs, used Item Response Theory (IRT) to capture the inferred positions along the constructs of interest. Equivalent to a Rasch model, our implementation specifies the log-odds of using an IA in a given lesson as a function of the complexity of the IA and the propensity of a teacher to employ an IA within its respective dimension. Our Rasch

model estimates IA complexities, ψ_m, and teacher propensities, π_j , to employ such IAs on a logit scale. Further, let $Y_{mj}=1$ if teacher j employs IA m and $Y_{mj}=0$ if not for IAs $m=1,...M$ and teachers $j=1,...J$ and let

$$\mu_{mj} = P(Y_{mj} = 1 | \psi_m, \pi_j) \quad (1)$$

denote the conditional probability that teacher j will employ IA m, and let

$$\eta_{mj} = \log[\mu_{mj} / (1 - \mu_{mj})] \quad (2)$$

indicate the natural log-odds of employing an IA. Given a Rasch model formulation, the log-odds of employing a specific IA is the difference between teacher j's propensity to make use of IAs, π_{pj}, and IA m's complexity, ψ_m: $\eta_{mj} = \pi_j - \psi_m$.

Estimates from a Rasch model yield an interpretable interval scale thereby affording insight as to the complexity of each IA and the locations of teachers along the latent dimensions (Rasch, 1980). Underlying these estimates is a first assumption that each IA discriminates among teachers in a similar manner. Alternative approaches, such as the two-parameter IRT model, may relax this assumption by additionally characterizing the complexity of each IA using a discrimination parameter λ_m such that

$$\eta_{mj} = \lambda_m(\pi_j - \psi_m) \quad (3)$$

The tenability of this assumption is subsequently examined. A second assumption is the (local) independence of IAs. In other words, we assume that given the IA complexity and teacher propensities, the IAs employed by teachers are independent. In studying instruction, this assumption has two implications. First, each IA selected by a teacher is independent of the other IAs selected by that same teacher. Second, local independence implies that the IAs measure a unidimensional construct of instruction. Both of these assumptions are problematic in the study of instruction and give rise to the next two features of our approach.

Multivariate. Our framework suggested that the IAs performed by a teacher in and across lessons are likely informed by several theoretically separate but related dimensions of instruction. To align with this framework, attend to the potential lack of independence among coordinated IAs employed by a teacher and examine the dimensionality of instruction, we extended our measurement model to be multivariate. Our measurement model now maintains multiple (p) latent scores for each teacher j-one for each dimension of instruction. Let the probability that teacher j will employ IA m be principally informed by the specific dimension which the IA belongs to such that

$$\mu_{pmj} = P(Y_{mj} = 1 | \psi_m, \pi_{pj}) \quad (4)$$

and let

$$\eta_{pmj} = \log[\mu_{pmj} / (1 - \mu_{pmj})] \quad (5)$$

indicate the natural log-odds of employing an IA in scale p.

Multilevel. The third layer of our model addresses two more features of studying instruction. To address dependencies among IAs within a lesson and within a teacher, we employ random effects. Corresponding with these random effects, we next give our model a multilevel structure so as to consider how contextual and background factors surrounding the instruction might be associated with IA selection. Together, the multivariate and multilevel structures allow us to additionally estimate different propensities for each dimension and to coarsely assess the dimensionality of instruction by examining the similarity of the associations between multilevel characteristics and each of the dimensions. Our model for instruction was

Level 1-Measurement model:

$$\eta_{ijk} = \sum_{p=1}^{P} D_{pijk}(\pi_{pjk} + \sum_{m=1}^{M_p - 1} \alpha_{pmjk} q_{pmijk}) \quad (6)$$

where

$\eta_{ijk}=log[\mu_{ijk}/(1-\mu_{ijk})]$ is the log-odds that teacher k for lesson j will employ IA i

μ_{ijk} is the probability conditioning on all fixed effects such that $\mu_{ijk} = P(Y_{ijk} = 1|\beta_{pk})$
D is an indicator taking on a value of 1 if the i^{th} item is in the scale that measures practice dimension p, 0 otherwise
π_{pjki} is the log odds of employing an IA in lesson j in teacher k to the reference item within practice trait type p
a_{pmijk}=1 if item i is the m^{th} item within scale p, 0 otherwise
α_{pmjk} is the discrepancy between the log odds of employing an IA for the m^{th} item in scale p for teacher k and the reference item within that scale, holding constant π_{pjk}

$$\pi_{pjk} = \beta_{0pk} + \sum_{s=1}^{S} \beta_{spk} X_{sk} + u_{pjk}$$

Level 2-Between lessons:
$$\alpha_{pmjk} = \beta_{pmk} + \sum_{k=1}^{K} \beta_{pmk} X_{k} + u_{pmk} \quad (7)$$
$$for\ p = 1,\ldots,\ P;\ m = 1,\ldots\ M-1$$

where X_s represent the lesson characteristics. Similarly, we expand the between teacher model using covariates to

$$\beta_{0pk} = \gamma_{00p} + \sum_{n=1}^{N} \gamma_{p0n} W_{nk} + r_{pk}$$

Level 3-Between teachers:
$$\beta_{pmk} = \gamma_{pmk} + \sum_{n=1}^{N} \gamma_{pmnk} W_{nk} + r_{pmnk} \quad (8)$$

where W_N represent teacher and classroom characteristics.

Because we are especially interested in characterizing instruction in an effective manner, we assessed the effectiveness of our empirical indices by relating them to student achievement using a hierarchical linear model adjusted for multiple pretests and other relevant teacher and student variables. Beyond the practices' association with average achievement, we also tested the extent to which each practice modifies the relationship between socio-economic status and achievement and as well as between prior ability and achievement (e.g. Nye et al., 2004).

Setting/Data

Setting. Instructional practice studies are especially useful in domains which have little to no evidence linking practice to achievement. In no area is this relationship potentially more important, and less complete, than early literacy. Our application focuses on grade 3 reading comprehension lessons and assesses the effectiveness our MMR model by evaluating the association of the empirical indices with student achievement in the ITBS reading comprehension subtest. Our review of literature identified three major dimensions of instruction for comprehension: pedagogical structure (PS), support for students' learning (SSL) and teacher directed instruction (TDI) (Figure 1). Rather than be exhaustive, these dimensions and their IAs (Figure 1) were intended to be those most central to early comprehension instruction and those most likely to distinguish effective from less effective instruction (Authors, 2009).

Data. This study focused on reading comprehension lessons (n=287) that were carried out during the literacy block observed four times across the school year in the classroom of 44 third-grade teachers in 19 schools in 6 districts. Observers recorded IAs performed during each lesson; a detailed explanation of the observation system and procedures can be found in Authors (2009). Of third grade teachers, 91% were female, 21% were non-White, 52% had a Masters degree, and the average years teaching experience was 13.

Results

We highlight just a few findings. First, we found that the observed instructional patterns characterized by the MMR were significantly associated with student achievement even after controlling for relevant factors (Table 1). In contrast, every summative scaling measure failed to measure practice in a manner that captures a significant association with student achievement (Table 2). Second, the analyses of the assumptions of the MMR model and the factors associated with instruction produced significant insight as to measuring instruction. First, there is empirical evidence to support our dimensions of instruction and evidence to suggest that the IAs form tractable Rasch scales (Table 3; Figures 2-3). Second, our analyses suggested a substantial and significant portion of the variation in instruction is attributable to the lesson and teacher levels. Such variation tends to be explained by both more stable factors such as teacher knowledge and contextual factors such as a classroom's prior ability (Table 4). Further, though the instructional dimensions largely share similar significant predictors, the direction of such associations is not constant, supporting our multivariate framework.

Discussion

Observing teachers' practice over extended periods of time provides a principal portrait of instruction. This study draws attention to the potential problems in effectively measuring this observed instruction. We believe that, at least in part, stronger methods such as a MMR model attend to specific and the more general complexities surrounding both instruction and its measurement better than simpler approaches. The results suggest that a summary that is too simple may conceal important individual differences in practice and fail to measure instruction as it relates to student achievement. For instance, although routine IAs likely contribute to the effectiveness of teachers, their dominance of summative scales potentially mask subtle differences.

In addition to producing more nuanced empirical indices regarding the nature and quality of instruction for student achievement analyses, the analyses and results of the MMR model are themselves useful. For example, the MMR model is particularly well suited to study the extent to which teachers adjust their practice based on classroom needs. The MMR model expands the scope of studies of instruction from simply summarizing instruction to the detailed study of its structural underpinnings and its relations to contextual factors. Few simpler analytic methods allow an accurate understanding of such underpinnings and contextual associations. A major advantage of using a MMR model is its ability to explore the underlying phenomenon of instruction while testing assumptions. In essence, the MMR model attempts to open the black box by testing assumptions and investigating the complexity of IAs, the ability of IAs to discriminate among teachers, the interrelations of instructional dimensions, the dependencies among IAs and the dimensions relations to contextual factors. Simpler measures, such as the sum of IAs, not only use a black box approach but are frequently composed of untestable assumptions. As a result, a key component and conceptual advance in using the MMR model for instruction is its ability explicitly test assumptions and investigate and revise the scales to understand what the IAs are measuring. For instance, in the current context, our analyses indicated that one IA in the PS scale (providing a summary) discriminated among teachers in a manner inconsistent with the other PS IAs. Simpler analyses may fail to identify such IAs and in turn violate assumptions or poorly measure the intended dimensions. Such discoveries not only improve the quality of our measures and assumptions but further facilitate the development of our reading and instruction theories.

References

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.

Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31*, 1-29.

McGhie, D., Underwood, K., Jordan, A. (2007). Developing effective instructional strategies for teaching in inclusive classrooms. *Exceptionality Education Canada, 17,* 27-52.

Nye, B., Konstantopoulos, S., & Hedges, L. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257.

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008). Classroom effects on children's achievement trajectories in elementary schools. *American Educational Research Journal, 45*, 365-397.

Pressley, M., Rankin, J., & Yokoi, C. (1996). A survey of instructional practices of primary teachers nominated as effective in promoting literacy. *Elementary School Journal, 96*, 363-384.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (Copenhagen, Danish Institute for Educational Research), expanded edition (1980) Chicago: The University of Chicago Press.

Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the Study of Instructional Improvement. *Educational Researcher, 38*, 120-131.

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. *Review of Educational Research, 77,* 454-499.

Table 1: HLM Achievement Model Assessing Conditional Association of MMR Practice Measures (TDI, SSL, PS) with ITBS Reading Comprehension

Covariates	Estimate (SE)		
Teacher Level	Intercept	Free Lunch Status Slope	ITBS 2007 Slope
Intercept	-0.01 (0.05)	-0.09 (0.13)	0.19 (0.05)*
Years teaching experience	0.00 (0.00)	-0.01 (0.01)	-0.00 (0.00)
Masters: Reading/Literacy	0.31 (0.13)*	0.028 (0.30)	0.22 (0.14)
Teacher knowledge	0.02 (0.04)	-0.23 (0.08)*	-0.03 (0.03)
Use of comprehensive reading program	0.13 (0.04)*	0.11 (0.09)	0.10 (0.04)*
Use of differentiated instruction	0.10 (0.03)*	-0.14 (0.08)	-0.03 (0.03)
TDI	0.17 (0.05)*	0.42 (0.12)*	0.15 (0.04)*
SSL	0.08 (0.06)	0.38 (0.13)*	0.15 (0.05)*
PS	0.04 (0.05)	0.20 (0.10)	0.05 (0.04)
Student Level			
Minority	-0.13 (0.09)		
Limited English Proficient	-0.19 (0.08)*		
Free/Reduced lunch status	-0.09 (0.13)		
Special Education	-0.10 (0.08)		
Fall 2007 ORF score	0.39 (0.03)*		
ITBS 2007 RC score	0.19 (0.05)*		
ITBS 2006 RC score	0.07 (0.03)*		

Note: Model is a random intercept and slopes HLM & Free Lunch Status and ITBS 2007 RC slopes varied significantly among classrooms

* Significant at the *p=0.05* level

Note: "*" for practice measures (TDI, SSL & PS) indicate significance at the *p=0.05* level after adjusting *p-values* using a sequential Bonferroni correction

Table 2: Sequential Bonferroni Adjusted *P-values* for Summative and MMR Measures of Instructional Practice

	Intercept		Free/Reduced Lunch		2007 ITBS	
	MMR	Sum	MMR	Sum	MMR	Sum
TDI	0.021*	0.350	0.009*	>0.900	0.008*	>0.900
SSL	0.372	0.128	0.048*	>0.900	0.040*	0.108
PS	0.521	>0.900	0.240	>0.900	0.498	>0.900

Note: For brevity we focus on the difference in inferences resulting from the Summative and MMR measures

* Significant at the *p=0.05* level after correcting for multiple hypotheses using a sequential Bonferroni correction

Table 3: Comparison of Rasch Model vs. Two Parameter IRT Model for Each Dimension of Instruction (*n=287*)

			Rasch		Two Parameter	
	IA	Item Biserial Correlation	Difficulty (SE)	Discrimination (SE)	Difficulty (SE)	Discrimination (SE)
TDI	1	0.64	-1.25 (0.17)	1.25 (0.15)	-1.27 (0.30)	1.21 (0.39)
	2	0.69	-0.48 (0.13)	1.25 (0.15)	-0.42 (0.13)	1.63 (0.52)
	3	0.50	-1.74 (0.21)	1.25 (0.15)	-2.56 (0.92)	0.74 (0.31)
	5	0.66	-0.85 (0.15)	1.25 (0.15)	-0.78 (0.17)	1.44 (0.42)
			BIC=839.43		*BIC=846.65*	
SSL	4	0.70	0.98 (0.18)	1.15 (0.20)	1.47 (0.56)	0.67 (0.29)
	6	0.68	1.37 (0.22)	1.15 (0.20)	1.48 (0.49)	1.03 (0.46)
	10	0.61	2.30 (0.34)	1.15 (0.20)	1.45 (0.40)	3.35 (4.15)
			BIC=839.43		*BIC=846.65*	
PS	7	0.81	0.47 (0.10)	2.14 (0.31)	0.44 (0.12)	2.79 (1.91)
	8	0.59	1.71 (0.18)	2.14 (0.31)	1.72 (0.34)	2.10 (0.93)
	9	0.70	-0.98 (0.12)	2.14 (0.31)	-1.10 (0.24)	1.66 (0.62)
			BIC=822.48		*BIC=833.25*	

Figure 1: Theoretical Dimensions of Third Grade Reading Instruction and IAs representative of these dimensions

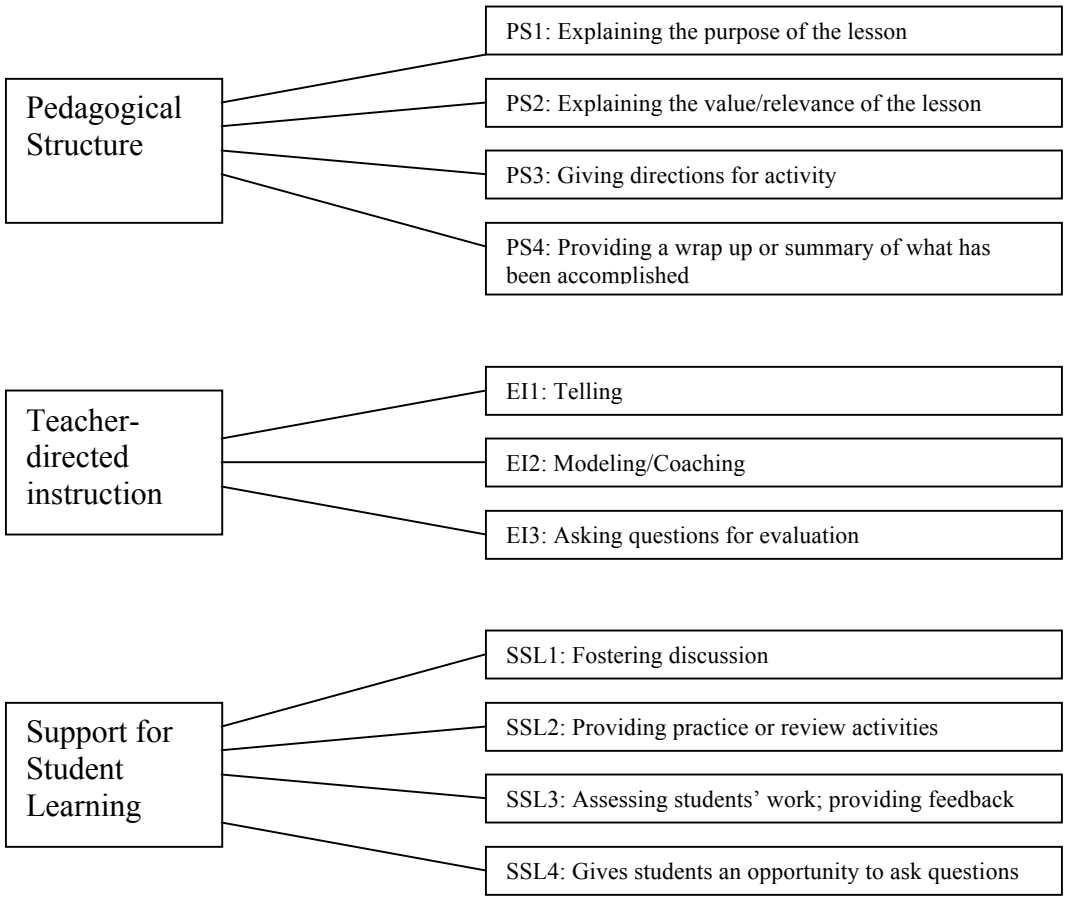

Figure 2: Rasch (a) and Two Parameter (b) ICCs for TDI

(a) (b)

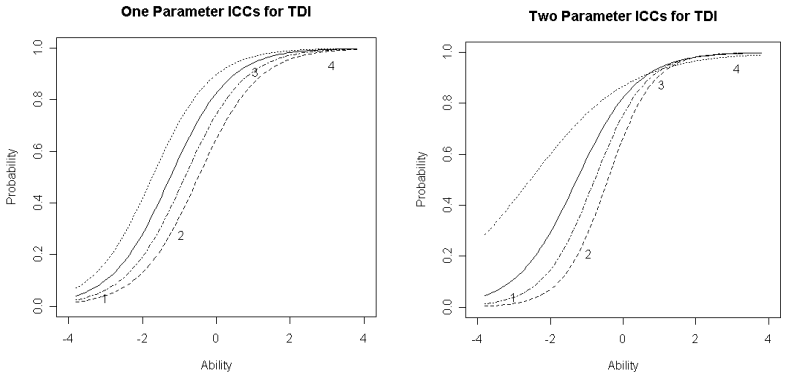

Figure 3: Rasch (a) and Two Parameter (b) ICCs for PS using all IAs

Note: PS4 (provides a summary) violated Rasch assumptions as it appears to discriminate differently than the other IAs and was thus removed to form a Rasch scale.

(a) (b)

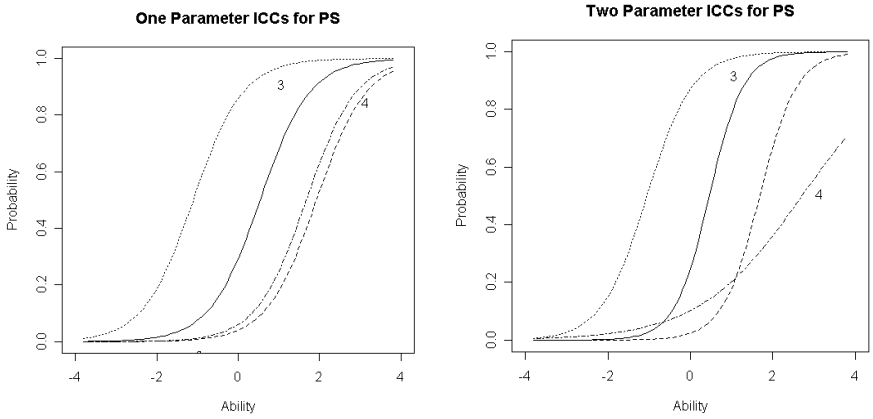

Table 4: Results of Full MMR Model for Instruction

Effect	Estimate (SE)		
	TDI	*SSL*	*PS*
Teacher level			
Intercept	0.13 (1.56)	0.52 (1.71)	6.60 (2.28)*
Masters	0.52 (0.22)*	-0.36 (0.24)	0.19 (0.31)
Teacher knowledge	-0.30 (0.12)*	0.44 (0.14)*	0.41 (0.17)*
Student data usage	0.32 (0.15)*	-0.26 (0.17)	-0.04 (0.22)
Comprehensive reading program usage	-0.14 (0.26)	0.12 (0.27)	-0.98 (0.36)*
Explicit instruction usage	0.98 (0.44)*	-1.20 (0.47)*	-0.12 (0.62)
Percent free/reduced lunch	-3.09 (0.49)*	1.18 (0.50)*	-0.04 (0.64)
Percent minority	1.52 (0.29)*	-1.15 (0.33)*	-1.67 (0.42)*
Average prior ability	-0.03 (0.01)*	0.04 (0.01)*	-0.01 (0.01)
Lesson level			
Lesson duration	0.06 (0.01)*	0.03 (0.01)*	0.07 (0.01)*
Observation period	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

* Significant at the *p=0.05* level

Note: "*" for the TK measure indicates significance at the *p=0.05* level after adjusting *p-values* using a sequential Bonferroni correction