Article

Measuring Instructional Differentiation in a Large-Scale Experiment

Educational and Psychological Measurement
2014, Vol. 74(2) 263–279

DOI: 10.1177/0013164413507724
epm.sagepub.com
SAGE

Ryan T. Williams[1], Andrew Swanlund[2], Shazia Miller[2], Spyros Konstantopoulos[3], Jared Eno[2], Arie van der Ploeg[2], and Coby Meyers[2]

Abstract

This study operationalizes four measures of instructional differentiation: one for Grade 2 English language arts (ELA), one for Grade 2 mathematics, one for Grade 5 ELA, and one for Grade 5 mathematics. Our study evaluates their measurement properties of each measure in a large field experiment: the Indiana Diagnostic Assessment Tools Study, which included two consecutive cluster randomized trials (CRTs) of the effects of interim assessments on student achievement. Each log was designed to measure instructional practices as they were implemented for eight randomly selected students in the participating teachers' classrooms. A total of 592 teachers from 127 schools took part in this study. Logs were administered 16 times in each experiment. Item responses to the logs were scaled using the Rasch model and reliability estimates for the differentiation measures were evaluated at the log level (observations within teachers), the teacher level, and the school level. Estimated reliability was above .70 for each of the log- and teacher-level measures. At the school level, reliability estimates were lower for Grade 5 ELA and mathematics. The variance between teachers and schools on the scaled differentiation measures was substantially less than within-teacher variation. These results provide preliminary evidence that teacher instructional logs may provide useful measures of instructional differentiation in elementary grades at multiple levels of aggregation.

[1]The University of Memphis, Memphis, TN, USA
[2]American Institutes for Research, Chicago, IL, USA
[3]Michigan State University, East Lansing, MI, USA

Corresponding Author:
Spyros Konstantopoulos, Michigan State University, 460 Erickson Hall, East Lansing, MI 48824, USA.
Email: spyros@msu.edu

Keywords

instructional differentiation, Rasch model, variance decomposition, psychometrics, ecometrics

Introduction

Interim assessment techniques are on the vanguard of education research. Instead of summing up class, school, or district results against a defined set of standards and as part of an accountability system (Perie, Marion, & Gong, 2007, 2009; Shavelson, 2006), interim assessment or assessment for learning (Stiggins, 2002) centers on active feedback loops that assist teachers and students in the learning process (Black & Wiliam, 1998; Heritage, 2010; Sadler, 1989). Perie et al. (2007, 2009) suggest that formative assessment is administered by the teacher for the explicit purpose of diagnosing student learning, identifying student gaps in knowledge and understanding, and redirecting teaching and learning.

Instructional change is the link between interim assessment and student performance. For instance, interim assessments are hypothesized to lead to differentiated instruction (Tomlinson, 2000), which is commonly believed to, in part, mediate the relationship between assessment and academic performance. Tomlinson (2000) describes differentiation in four domains: content, process, product, and learning environment. However, methods with which to optimally measure this complex instructional process are not well-refined nor have they been extensively evaluated from a psychometric perspective thus far.

Our article seeks to fill in that gap in the literature and has three main aims. First, we formally operationalize instructional differentiation and from those operations we derive four measures of instructional differentiation from teacher logs. Second, we present the measurement properties of each of the four measures of instructional differentiation at different levels of aggregation: within teachers across multiple observations, among teachers in schools, and among schools. Third, we describe treatment and control differences in instructional differentiation in the Indiana Diagnostic Assessment Tools Study. The following section provides a theoretical overview of interim assessment methods and instructional differentiation along with a discussion of methods used to measure complex instructional phenomena.

Related Literature

Interim Assessments

Interim assessments are increasingly popular as an educational intervention intended to improve academic achievement. These systems serve multiple purposes including formative assessment, diagnostic assessment, and predictive assessment (e.g., Perie et al., 2007, 2009). Regardless of the purpose of using interim assessments, they are administered multiple times throughout an academic period with the intention of

providing educators a greater opportunity to respond to and remediate specific deficiencies about student learning and to monitor their own effectiveness in responding to those deficiencies.

Meta-analytic evidence has typically found that instructional feedback systems, broadly defined, yield positive mean effects on academic achievement. Black and Wiliam (1998), for example, revealed gains of a half to a full standard deviation in academic achievement, with low-achieving students making the largest improvements. Nyquist (2003), Kluger and DeNisi (1996), and Fuchs and Fuchs (1986) also found positive mean effects on academic achievement. According to Nyquist (2003), the strongest forms of feedback are immediate, frequent, and computerized or written, providing the student with the correct answer, an explanation of why an incorrect answer is incorrect, and a relevant activity designed to improve the student's current knowledge.

Several large empirical studies, however, have recently examined the effectiveness of interim assessment methods for improving academic achievement and have shown mixed results. Henderson, Petrosino, Guckenburg, and Hamilton (2008) examined the effects of an interim assessment system in Massachusetts using a comparative interrupted time series design. The authors compared 22 schools participating in the new assessment system with 44 control schools that had not yet begun using the new system. The authors could not identify a difference between treatment and control schools and noted that their study was underpowered, that control condition may have received comparable treatments, and that treatment effects may not manifest in the first year after implementation. Quint, Sepanik, and Smith (2008) evaluated the impact of Formative Assessment Student Thinking in Reading (FAST-R) on third and fourth grade student achievement in 21 Boston schools using an interrupted time series design. While the authors observed positive effects of the FAST-R program, they were not statistically significant. They mentioned that a large proportion of comparison school teachers were receiving potentially comparable amounts of formative assessment professional development, which could have artificially shrunken the treatment effect estimate. Carlson, Borman, and Robinson (2011) conducted a district-level randomized experiment to investigate the impact of the Center for Data-Driven Reform in Education (CDDRE) assessment on academic achievement across 56 districts in 7 states. This study found positive effects in both reading and mathematics but only the mathematics results were statistically significant. The authors speculated that the positive effects may have been driving by instructional modifications the teachers were making based on the assessment feedback they received. Slavin, Cheung, Holmes, Madden, and Chamberlain (2013) examined longitudinal impacts of CDDRE and observed statistically significant results in reading and mathematics achievement by year four of the experiment. The results from Slavin et al. (2013) indicate that positive effects of interim assessment may not be apparent until teachers and students are fully oriented to the assessment procedures. That is, it may take time for teachers to learn how to use, interpret, and modify

their instruction effectively in the first year(s) of implementing an interim assessment system.

Cordray, Pion, Brandt, Molefe, and Toby (2012) evaluated the impact of the Measures of Academic Progress assessment system on student achievement and instructional practices. This study randomized 32 elementary schools in Illinois to treatment and control conditions. The authors found no difference in academic achievement between treatment and control schools. This study is unique in that in addition to evaluating the effect of interim assessment methods on academic achievement, it also examined the effects of interim assessment methods on instructional differentiation (i.e., the mediating mechanism between assessment and achievement). However, they found no difference in instructional practices, specifically instructional differentiation, between treatment and control schools.

Instructional Differentiation

Differentiation is theorized to mediate the relationship between interim assessment and achievement: increasing teachers' exposure to student progress data will promote instructional differentiation, leading to achievement gains. Differentiation is a complex and dynamic instructional practice that may exist in multiple domains and on multiple levels within each domain (Tomlinson, 2000; Tomlinson & Strickland, 2005). Content differentiation would involve varying the instructional topics, for example, that students within a classroom would receive. Process differentiation might involve teaching different students at different levels of cognitive difficulty (e.g., ''remedial'' vs. ''advanced''). Product differentiation might involve assigning different tasks to different students. Perhaps because of its complexity, methods for measuring instructional differentiation, among other instructional processes, are not well-refined. Some studies have relied on survey methods to measure differentiation. Graham et al. (2008), for example, conducted a national survey of differentiated instruction in spelling. The authors constructed a set of survey items that asked teachers about the quantity and frequency of various instructional practices, including instructional adaptations for elementary school students of varying ability levels. Only about 58% of teachers in their sample reported that they modified their instruction according to the spelling needs of their students. Other studies have relied on classroom observations to examine differentiated instruction. Carolan and Guinn (2007) conducted a series of interviews and classroom observations with five teachers. They amassed more than 35 hours of descriptive and qualitative data about teachers' attitudes and instructional practices in relation to differentiation, and they qualitatively identified four common characteristics of instructional differentiation: offering personalized instruction, using flexible means to reach defined ends, mining subject-area expertise, and creating a caring classroom in which differences are seen as assets. In the context of a 3-year randomized experiment of professional development and evidence-based curriculum, Van Tassel-Baska et al. (2008) used a structured classroom observation instrument to measure change in differentiated

instruction over 3 years and observed teachers in the first and the last month of the academic year. Trained research staff conducted observations twice each year of the experiment. The authors found that teachers in the experimental condition were rated significantly higher than comparison condition teachers in the effectiveness of their instructional differentiation.

The methods outlined above have a number of limitations, according to Rowan, Camburn, and Correnti (2004). On the one hand, surveys may provide a large amount of data with relatively little cost. However, surveys often require teachers to recall instructional behavior over a long period of time (e.g., an entire academic year). This method is susceptible to retrospective self-report bias, especially when reporting on rare or frequent behaviors. Classroom observation systems avoid teacher self-report entirely. But, trained observers may not always be able to observe the entirety of the instructional process (e.g., understanding why questions were assigned to particular students or how the ability levels of each student in the classroom affect instructional decisions). Another method for deriving measures of instructional differentiation is through the use of teacher instructional logs (e.g., Rowan, Correnti, & Miller, 2002). Instructional logs are commonly administered multiple times throughout an academic year in hopes of capturing an instructional profile for that year. Because logs are administered over multiple time points throughout an academic year, they are more representative of an educator's instructional profile than classroom observations and cross-sectional survey methods. Logs are also less burdensome on teacher memory because teachers focus on a specific date and complete the log on or around that date.

Rowan and Correnti's (2009) work with instructional logs has guided the methods we used in this study. From the instructional logs, we conceptualized differentiation as any difference between students on the instructional practices their teacher enacted and that they therefore received. For example, if teachers indicated they covered vocabulary at an ''enriched'' level for some students and at a ''regular'' level for other students, our operations state that differentiation occurred. The following section describes in greater detail the content of the instructional logs and how we transformed responses to those items into meaningful and useful measures of instructional differentiation.

Method

Sample and Setting

The current study of differentiation is part of the Indiana Diagnostic Assessment Tools Study, which conducted two CRTs on the effects of interim assessment systems on student achievement, one in 2009-2010 and a second in 2010-2011. Indiana's statewide rollout of Diagnostic Assessment Tools began in 2008-2009 and included two products, Wireless Generation's mCLASS: Reading 3D and mCLASS: Math were used in Grades K-2, and CTB/McGraw Hill's Acuity was used in Grades 3 to 8. Indiana's intent was to ''encourage the advanced and gifted child, drive progress in

the student who is ready, and accelerate progress for the student whose learning reflects gaps in preparation and readiness'' (Indiana State Board of Education, 2006, pp. 11-12); this would happen because teachers' instructional choices would be better informed, leading to improved student achievement as measured by the state accountability test (ISTEP +).

Both assessment vendors' products were aligned to Indiana academic standards. Both offer teachers (and administrators) near immediate access to student results via networked computers and digital devices. Numerous preformatted reports, most keyed to Indiana's academic standards, as well as ad hoc displays are supported. Indiana defined specific testing windows statewide for each product (either three or four times per academic year). Both vendors supply a variety of instructional resources, packaged exercises to practice or explore skills.

mCLASS provides diagnostic measures in literacy and numeracy to teachers of K-2 students. mCLASS: Reading3D comprises Dynamic Indicators of Basic Early Literacy Skills, which alerts teachers to problems in students' learning of basic literacy and a running records implementation that helps teachers track student's error patterns, reading strategies, and comprehension. Teachers are guided through the assessment process on a portable digital device. Within the interface, they can immediately view results and compare them to prior performance. The digital devices synchronize data to a networked computer in the classroom. These in turn synchronize nightly with vendor servers. The mCLASS: Math assessments are paper and pencil and teachers enter results into a computer. Detailed individual and group reports as well as ad hoc queries are available to the classroom teacher and other authorized personnel via a web interface for all mCLASS assessments.

CTB/McGraw Hill's Acuity provides online assessments in reading and mathematics for Grades 3 to 8. The assessments are 30- to 35-item multiple-choice tests typically completed within a class period, usually in group settings in a computer lab. The Acuity diagnostic assessments identify individual student needs and targets for personalizing instruction. Item content across windows follows state-recommended scope and sequence. Acuity predictive assessments forecast student performance on ISTEP + and are reported in its standard scale. Each participating school selected one of these two types.

Vendors conducted 2-day workshops on their systems to select staff from multiple schools before each school year began. These staff were in turn expected to train their colleagues. Indiana Department of Education and vendor staff were available during the academic year to offer support and resolve technical or financial issues. However, no extended training in data interpretation or instructional practice was associated with each product. Rather, Indiana Department of Education leveraged a variety of supports and partners to increase capacity for data-driven decision making throughout the state.

The Indiana Diagnostic Assessment Tools study included a total of 127 schools, 57 from 2009-2010 and 70 from 2010-2011. These two cohorts of schools were

randomized into treatment and control conditions separately, and are designated CRT1 and CRT2, respectively. Indiana's statewide interim assessment system rollout was voluntary; therefore, schools applied to the Indiana Department of Education for the systems. Study schools were drawn from state lists of applicants for the intervention. The rollout began in the 2008-2009 school year, and the study took place in the 2009-2010 and 2010-2011 school years. Study schools applied for the intervention in the second and third years in which it was available. Therefore, the schools are not the earliest adopters, but CRT1 schools are earlier adopters than CRT2 schools.

Schools could apply for interim assessment systems at different grade levels: K-2 (mCLASS only), Grades 3 to 8 (Acuity only), or K-8 (both). From the school that applied for the interim assessment systems, a stratified simple random sample was drawn and the sampled schools were randomized into treatment and control conditions.

Measures

Our team developed four instructional logs: one for Grade 2 mathematics, one for Grade 2 English language arts (ELA), one for Grade 5 mathematics, and one for Grade 5 ELA. Each measure was analyzed separately, that is, the analysis was repeated four times (see Statistical Procedures section below). The ELA instructional logs were based on Rowan and Correnti's (2009) instructional logs. Topic areas in the ELA logs included comprehension, writing, word analysis, reading fluency, and vocabulary. These topic areas had items related to teacher instruction, concept or skill difficulty, student materials, and student activities. Table 1 provides a blueprint for the Grade 2 and Grade 5 ELA instructional logs.

The mathematics logs were developed by internal content experts, following the ELA model and guided by the Indiana mathematics standards. The Grade 5 mathematics logs had seven topic areas: number sense, computation, algebra and function, geometry, measurement, problem solving, and data analysis and probability. Only the Grade 5 mathematics log had items related to data analysis and probability. Like the ELA logs, each topic area contained items related to teacher instruction, concepts or skill difficulty, student materials, and student activities. Table 2 provides blueprints for the Grade 2 and Grade 5 mathematics logs.

The checklists were constructed to measure the types and level of content teachers delivered to their students during an instructional period and to whom that content was delivered. For example, a teacher could indicate that a particular student was instructed in fractions and that instruction occurred at the ''enriched'' level (as opposed to ''regular'' or ''remedial''). For each checklist item, the teacher was coded as having differentiated on that aspect of instruction if the item was endorsed for at least one student but not others. If the item was endorsed for all students or no students, the teacher was coded as not having differentiated on

Table 1. English Language Arts Instructional Log Blueprints.

	C	W	WA	RF	V	Total number of items
Grade 2 ELA						
Teacher instruction	4	—	—	5	3	12
Concept or skill difficulty	20	10	12	5	6	53
Student materials	6	4	7	3	2	22
Student activities	10	9	3	5	6	33
Total number of items	40	23	22	18	17	120
Grade 5 ELA						
Teacher instruction	4	—	—	5	3	12
Concept or skill difficulty	20	10	12	5	6	53
Student materials	6	4	7	3	2	22
Student activities	10	9	3	5	6	33
Total number of items	40	23	22	18	17	120

Note. C = Comprehension; W = Writing; WA = Word Analysis; RF = Reading Fluency; V = Vocabulary.

Table 2. Mathematics Instructional Log Blueprints.

	NS	C	AF	G	M	PS	DAP	Total number of items
Grade 2 mathematics								
Teacher instruction	6	3	4	5	8	3	—	29
Concept or skill difficulty	12	9	4	5	12	5	—	47
Student materials	5	5	5	5	5	5	—	30
Student activities	14	14	14	14	14	14	—	84
Total number of items	37	31	27	29	39	27	—	190
Grade 5 mathematics								
Teacher instruction	4	5	5	6	7	3	6	36
Concept or skill difficulty	7	7	7	9	7	10	5	52
Student materials	5	5	5	5	5	5	5	35
Student activities	14	14	14	14	14	14	14	98
Total number of items	30	31	31	34	33	32	30	221

Note. NS = Number Sense; C = Computation; AF = Algebra and Functions; G = Geometry; M = Measurement; PS = Problem Solving; DAP = Data Analysis and Probability.

that aspect of instruction. For polytomous items about the cognitive level of instruction a concept or skill was taught (i.e., ''remedial,''''regular,''''enriched''), teachers were coded as having differentiated if the item was endorsed at different levels for different students and not having differentiated otherwise. If a teacher did not respond to an item (e.g., he or she may not have covered a certain topic, skipped the item, or not responded to the checklist on that date), the item was coded as ''not administered'' and did not affect teachers' estimated ability to differentiate.

Table 3. Distribution of Teachers in CRT1 and CRT2.

	Grade 2 ELA		Grade 2 Mathematics		Grade 5 ELA		Grade 5 Mathematics		Total	
	T	C	T	C	T	C	T	C	T	C
CRT1	54	45	53	42	47	41	42	37	196	165
CRT2	42	36	42	32	10	28	30	11	124	107
Total	96	81	95	74	57	69	72	48	320	272

Note. T = Treatment Group; C = Control Group.

Data

This study used data obtained from 592 treatment and control school teachers of ELA and mathematics in Grades 2 and 5 from CRT1 and CRT2. The distribution of these teachers is presented in Table 3.

In each CRT, we asked teachers in control and treatment schools in Grades 2 and 5 to complete 16 instructional logs on scheduled dates throughout the academic year, roughly one every 2 weeks. It was not practical to ask teachers to report on the detailed instructional transactions of each student in their classrooms. As such, we randomly sampled eight students in each teacher's classroom at the beginning of each academic year. Teachers were asked to report on the instruction provided to these eight students during each of the 16 log administrations. These design elements helped ensure a representative sample of teacher instructional practices over the course of the academic years.

The participation rate for the teachers in our sample was high. For instance, consistently more than 80% of teachers completed at least 14 of the 16 logs. Teachers completed the logs online and results were stored on servers. Teachers who did not respond on time (e.g., within a day or two of the scheduled date) were sent reminders to complete the checklist, and they had up to 1 week after the scheduled administration to complete their logs. Table 4 presents the percentages of teachers completing at least 14 of the 16 checklists, 15 of the 16 checklists, and all 16 checklists.

Across both CRTs, we accumulated a total of 1,792 observations for Grade 5 mathematics; 1,834 observations in Grade 5 ELA; 2,393 in Grade 2 mathematics; and 2,462 observations in Grade 2 ELA.

Statistical Procedures

We evaluated the measurement properties of the log-, teacher-, and school-level measures of instructional differentiation using the Rasch model. In particular, to evaluate log- or observation-level reliability, teacher item responses, from each of the 16 instructional logs, were fit to the Rasch model for dichotomous item responses (Rasch, 1980),

Table 4. Participation Rates.

Grade/subject area	At least 14 checklists	At least 15 checklists	All 16 checklists	Sample size (*n*)
Grade 2 ELA	83%	73%	57%	169
Grade 2 mathematics	82%	75%	55%	177
Grade 5 ELA	86%	79%	62%	120
Grade 5 mathematics	90%	82%	66%	126

$$P(X_{nit} = 1|\theta_{nt}) = \frac{exp(\theta_{nt} - \delta_i)}{1 + exp(\theta_{nt} - \delta_i)}, \quad (1)$$

which assumes that the probability that teacher *n* differentiates on item *i* at time *t* depends on teacher *n*'s ability to differentiate their instruction, θ_{nt}, and the difficulty of differentiating instruction with respect to item *i*. This analysis was implemented in Winsteps (v3.75.0; Linacre, 2013). At the log-level, which involved repeated observations within teachers, we estimated item-measure correlations and mean square infit statistics. These two statistics help evaluate the conformity of the items to the measurement model. Infit mean squares are the sum of the squared residuals between observed responses and the predicted responses, weighted by the variance of those residuals. Infit statistics have an expected value of 1.0 with larger values indicating poor fit. Values of 2.0 or greater indicate that the unexpected variance in a given item is at least twice what the model predicted and thus for purposes of this study, we eliminated items with infit values greater than or equal to 2.0. We also conducted a preliminary assessment of each measure's dimensionality (i.e., structural validity) by analyzing the component structure of the residuals from the Rasch model, where the variance explained by the measures was removed. This method allowed us to preliminarily detect and evaluate the strength of any secondary structures in the data; structures independent of the underlying differentiation construct.

We also investigated the reliability of teacher- and school-level aggregate (mean) measures of instructional differentiation. Working from the basic conceptualization of reliability, the ratio of true score variance to observed score variance, we used the following multilevel models to partition the variance of the scaled measures within teachers, the variance of observations of teachers within schools, and the variance among schools. The first-level model for observation *i* for teacher *j* in school *k* is

$$y_{ijk} = \pi_{0jk} + \varepsilon_{ijk}. \quad (2)$$

The second-level model for teacher intercepts is

$$\pi_{0jk} = \beta_{00k} + r_{0jk}. \quad (3)$$

And the third-level model for school intercepts is

$$\beta_{00k} = \gamma_{000} + \nu_{00k}, \tag{4}$$

where y is ith estimated differentiation ability for teacher j in school k, π_{0jk} is the estimated average differentiation ability for teacher j in school k, ε_{ijk} is the residual associated with the ith observation of teacher j in school k, β_{00k} is the average differentiation ability of the teachers in school k, r_{0jk} is the random effect associated with the jth teacher in school k at level two, γ_{000} is the grand mean differentiation ability across schools, and ν_{00k} is the random effect associated with the kth school at level three (i.e., the level three variance component). The variances of the second- and third-level random effects capture the clustering in the data. In a single-level equation the mixed effects model is

$$y_{ijk} = \gamma_{000} + \nu_{00k} + r_{0jk} + \varepsilon_{ijk}. \tag{5}$$

Using the internal consistency estimators discussed in Raudenbush and Bryk (2002); Raudenbush, Rowan, and Kang (1991); Jeon, Lee, Hwang, and Kang (2009); and Raudenbush and Sampson (1999), we estimated teacher-level reliability, $\alpha_{\pi_{jk}}$, with

$$\alpha_{\pi_{jk}} = \frac{\tau_\pi^2}{\tau_\pi^2 + \frac{\sigma^2}{n_{jk}}}, \tag{6}$$

where τ_π^2 is the estimated between-teacher variance component (i.e., the variance of r_{0jk}), σ^2 is the estimated observation-level variance component (i.e., the variance of ε_{ijk}), and n is the number of observations per teacher j and school k. The term $\alpha_{\pi_{jk}}$ provides an estimate of the degree to which aggregate teacher measures of instructional differentiation can be reliably distinguished (i.e., low differentiation teachers from high differentiation teachers). Estimating average teacher-level reliability, α_π, as we did in this study, across a sample of observations nested in teachers is accomplished by replacing n_{jk} with the average number of observations per teacher, n:

$$\alpha_\pi = \frac{\tau_\pi^2}{\tau_\pi^2 + \frac{\sigma^2}{n}}. \tag{7}$$

These equations show that reliability increases as the average number of observations per teacher increases or as the ratio of between-teacher variation to total variation increases.

Similarly, we estimated school-level reliability, α_{β_k}, with

$$\alpha_{\beta_k} = \frac{\tau_\beta^2}{\tau_\beta^2 + \frac{\tau_\pi^2}{m_k} + \frac{\sigma^2}{n_{jk} m_k}}, \tag{8}$$

where τ_β^2 is the estimated between-school variance component (i.e., the variance of ν_{00k}), n_{jk} indicates the number of observations per teacher j in school k, and m_k is the

number of teachers in school k. Note here too that each school may have a reliability estimate. To get general or summary estimate, like the ones provided in this study, of school-level reliability (i.e., average reliability), α_β, one can substitute n_{jk} with the average number of observations per teacher, n and m_k with the average number of teachers per school, m:

$$\alpha_\beta = \frac{\tau_\beta^2}{\tau_\beta^2 + \frac{\tau_\pi^2}{m} + \frac{\sigma^2}{nm}}. \quad (9)$$

School-level reliability is then driven by the number of observations per teacher, the number of teachers per school, and the ratio of between-school variation to total variation. It is easy with these reliability estimators to estimate how many additional observations per teacher or school one may need to achieve a given level of reliability for a given set of variance components; an important feature for designing cluster-randomized experiments. For example, if one wanted to estimate how many additional teachers would be needed per school, on average, to achieve .70 school-level reliability, they could solve Equation 9 for m and use the observed values of τ_β^2, τ_π^2, σ^2, and n, say .15, .25, .50, and 15, respectively. Solving Equation 9 for m

$$.70 = \frac{.15}{.15 + \frac{.25}{m} + \frac{.50}{15m}}$$

gives 4.47, or about 5 teachers per school on average to achieve reliability of .70. The nlme R package (Pinheiro, Bates, DebRoy, Sarkar, & R Core Development Team, 2013) was used to estimate teacher- and school-level reliability.

Results

A summary of measurement properties for each of the instructional differentiation measures is provided in Table 5. Item conformity was examined using item and item-measure correlations. For Grade 5 mathematics, one item under student activities in measurement was dropped because of disjoint subsets (i.e., teachers responded to this item too infrequently to estimate its parameters): ''Use flashcards, games, or computer activities to improve recall or skill.'' One other Grade 5 mathematics item, under student activities in data analysis and probability misfit the measurement model (i.e., mean square infit greater than 2.00): ''Work on an investigation, problem, or project over an extended period of time.'' For Grade 2 mathematics, one item under student activities in geometry was dropped because of disjoint subsets: ''Write extended explanations of mathematical ideas, solutions, or methods.'' Three other items, under student activities in algebra and functions, were eliminated because of low (i.e., $<.40$) point-measure correlations: ''Work on an investigation for an extended period of time,''''Write extended explanations of mathematic ideas, solutions, or methods,'' and ''Analyze similarities and differences among representations, solutions, or methods.'' All the items on the ELA measures sufficiently fit the model

Table 5. Instructional Log-Level Measurement Properties.

Measure	NI	NO	Log-level reliability	Item-measure correlations	Variance explained by measures (eigenvalue)	Residual variance explained by first contrast (eigenvalue)
Grade 2 ELA	120	2,190	.77	.67-.92	46.00% (102.20)	1.90% (4.30)
Grade 2 mathematics	186	2,262	.76	.67-.98	42.0% (133.10)	1.70% (5.30)
Grade 5 ELA	120	1,619	.74	.70-.90	39.00% (71.50)	3.10% (5.70)
Grade 5 mathematics	219	1,641	.76	.67-.99	44.00% (166.70)	1.70% (6.30)

Note. NI = number of estimable items; NO = number of estimable observations.

(i.e., <2.0 mean square infit). The item-measure correlation for each item on each measure was high, ranging from .67 to .99. The log-level reliability estimates for each measure were adequate (i.e., >.70), indicating that each measure consistently separated observations along the construct. After removing the observed variance explained by the Rasch measures, we investigated the strength of any secondary dimensions by examining residual principal components. For each of the logs, the variance explained by the measures was overwhelmingly dominant. And although there was some evidence of possible secondary dimensions (first contrasts), the residual variance explained by these contrasts was small enough not to be of concern.

Table 6 provides teacher- and school-level reliability estimates for each of the four differentiation measures. Teacher-level reliability was high for each of the four measures but school-level reliability was low, primarily for the Grade 5 subject areas. For Grade 5, the between school variance estimates, τ_{β}^{2}, were small in comparison with the between-teacher variance estimates, τ_{π}^{2}, and the within-teacher residual variance estimates, σ^{2}. Because the between-school variance components were comparatively small for each of the Grade 5 subject areas, a greater number of teachers per school would be required to approach the log- and teacher-level reliability estimates.

Discussion

Differentiation is a popular instructional strategy. This study operationalized instructional differentiation based on item response patterns to teacher instructional logs and defined differentiation as instructing any student at all differently than other students. The extent to which these item responses could form a useful measure of instructional differentiation was investigated. Log-level instructional differentiation measures demonstrated strong psychometric properties. The items on each of the instructional logs, with the exception of one item on the Grade 5 mathematics log, demonstrated strong conformity and the measures formed a usefully unidimensional representation of the construct. Log-level reliability estimates were adequately high (i.e., >.70) for each of the four differentiation measures. Teacher-level reliability estimates were also

Table 6. Teacher- and School-Level Reliability Estimates.

Measure	n	m	σ^2	τ_π^2	τ_β^2	Teacher-level reliability	School-level reliability
Grade 2 ELA	13.90	2.95	.65	.19	.18	.80	.69
Grade 2 mathematics	14.16	2.97	.67	.15	.17	.76	.71
Grade 5 ELA	14.56	2.33	.66	.22	.12	.83	.52
Grade 5 mathematics	14.34	2.22	.62	.27	.11	.86	.43

Note. n = average number of logs per teacher; m = average number of teachers per school; σ^2 = log-level residual variance; τ_π^2 = between-teacher variance; τ_β^2 = between-school variance.

adequately high for each measure. However, school-level reliability estimates were low for Grade 5 ELA and Grade 5 mathematics. The proportion of total variance in the instructional differentiation measures that was among schools was small, which decreased the reliability estimates. One negative outcome of low school-level reliability is a decrease in statistical power when estimating the effects of school-level treatments on school-level instructional differentiation or when using differentiation as a covariate. Increasing the average number of teacher respondents per school or the average number of observations per teacher would increase reliability.

While the results of this study are likely most applicable to education researchers studying instructional differentiation, local education agencies (LEAs) may also benefit from the methods described in our study. For example, LEAs investing in professional development on instructional differentiation may use logs such the one described here to assess the effectiveness of the professional development courses. Of course, additional items related to instruction and differentiation could be added to the existing logs or could be used to replace the items on the logs used in this study to meet an LEA's needs.

This study was limited to measuring instructional differentiation in Grades 2 and 5 and in ELA and mathematics. Instructional differentiation in other grades and other subject areas may yield different results. We discussed limitations of existing methods for measuring complex instructional processes. The teacher logs used in our study obviate some of those limitations (e.g., expensive observations, cross-sectional survey results). While the way we operationalized instructional differentiation was broad and useful for measurement purposes, we understand that there may be other ways to operationalize differentiation. Still we contend that our way presented in this article is generally useful. We hope that others use our definition, and we encourage others to develop new definitions to best suit the needs of their research while providing researchers and practitioners new ways of examining these complex instructional processes.

Still our definition and results may be potentially compromised if teachers use grouping in the classroom. For example, in small groups it is possible that different students could be engaged in different tasks and that process may seem like

differentiated instruction. Alternatively, teachers may use ability grouping poorly where high achievers constitute one group and low achievers another. This practice may also seem like differentiated instruction. In both cases however, what is observed is not in line with the spirit of differentiated instruction.

Finally, while we discuss differentiation in the context of interim assessment systems in this article, we understand that differentiation may occur outside such systems. For example, teachers may have their own criteria and ideas about differentiating instruction based on their personal understanding of students' abilities.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305E090005 to Learning Point Associates, a subsidiary of the American Institutes for Research. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

References

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, *80*, 139-144.

Carlson, D., Borman, G. D., & Robinson, M. (2011). A multistate district-level cluster randomized trial of the impact of data-driven reform on reading and mathematics achievement. *Educational Evaluation and Policy Analysis*, *33*, 378-398.

Carolan, J., & Guinn, A. (2007). Differentiation: Lessons from master teachers. *Educational Leadership*, *64*(5), 44-47.

Cordray, D., Pion, G., Brandt, C., Molefe, A., & Toby, M. (2012). *The impact of the Measures of Academic Progress (MAP) program on student reading achievement (Final Report). NCEE 2013-4000*. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED537982

Fuchs, L. S., & Fuchs, D. (1986). Effects of systematic formative evaluation: A meta-analysis. *Exceptional Children*, *53*, 199-208.

Graham, S., Morphy, P., Harris, K. R., Fink-Chorzempa, B., Saddler, B., Moran, S., & Mason, L. (2008). Teaching spelling in the primary grades: A national survey of instructional practices and adaptations. *American Educational Research Journal*, *45*, 796-825.

Henderson, S., Petrosino, A., Guckenburg, S., & Hamilton, S. (2008). *A second follow-up year for measuring how benchmark assessments affect student achievement* (Regional Educational Laboratory Technical Brief No. 2). Retrieved from http://0-ies.ed.gov.opac.acc.msmc.edu/ncee/edlabs/regions/northeast/pdf/techbrief/tr_00208.pdf

Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity* (National Center for Research on Evaluation, Standards, and Student Testing [CRESST] and the Council of Chief State School Officers [CCSSO]). Washington, DC: CCSSO. Retrieved from http://129.33.81.41/documents/mde/formative_assessment_next_generation_heritage_338483_7.pdf

Indiana State Board of Education. (2006). *A long-term assessment plan for Indiana: Driving student learning*. Indianapolis, IN: Author.

Jeon, M.-J., Lee, G., Hwang, J.-W., & Kang, S.-J. (2009). Estimating reliability of school-level scores using multilevel and generalizability theory models. *Asia Pacific Education Review*, *10*, 149-158. doi:10.1007/s12564-009-9014-3

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*, 254-284.

Linacre, J. M. (2013). Winsteps (Version 3.75.0) [Computer Software]. Beaverton, OR: Winsteps.

Nyquist, J. (2003). *Reconceptualizing feedback as formative assessment: A meta-analysis* (Unpublished master's thesis). Vanderbilt University, Nashville, TN.

Perie, M., Marion, S., & Gong, B. (2007). *A framework for considering interim assessments*. Dover, NH: National Center for the Improvement of Educational Assessments. Retrieved from http://www.nciea.org/publications/ConsideringInterimAssess_MAP07.pdf

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5-13.

Pinheiro, J., Bates, D., DebRoy, S., and Sarkar, D., & R Core Development Team. (2013). *nlme: Linear and nonlinear mixed effects models* (R package version 3.1-111).

Quint, J. C., Sepanik, S., & Smith, J. K. (2008). *Using student data to improve teaching and learning: Findings from an evaluation of the Formative Assessments of Students Thinking in Reading (FAST-R) program in Boston elementary schools*. New York, NY: MDRC. Retrieved from http://www.eric.ed.gov/ERICWebPortal/detail?accno=ED503919

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press.

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

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 295-330.

Raudenbush, S. W., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*(1), 1-41.

Rowan, B., Camburn, E., & Correnti, R. (2004). Using teacher logs to measure the enacted curriculum: A study of literacy teaching in third-grade classrooms. *Elementary School Journal*, *105*, 75-101.

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

Rowan, B., Correnti, R., & Miller, R. (2002). What large-scale survey research tells us about teacher effects on student achievement: Insights from the Prospects Study of Elementary Schools. *Teachers College Record*, *104*, 1525-1567.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, *18*, 119-144.

Shavelson, R. J. (2006). *On the integration of formative assessment in teaching and learning with implications for teacher education*. Paper prepared for the Stanford Education Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group.

Slavin, R. E., Cheung, A., Holmes, G. C., Madden, N. A., & Chamberlain, A. (2013). Effects of a data-driven district reform model on state assessment outcomes. *American Educational Research Journal*, *50*, 371-396.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*, 758-765.

Tomlinson, C. A. (2000). *Differentiation of instruction in the elementary grades*. Champaign: ERIC Clearinghouse on Elementary and Early Childhood Education, University of Illinois.

Tomlinson, C. A., & Strickland, C. A. (2005). *Differentiation in practice: a resource guide for differentiating curriculum, grades 9-12*. Alexandria, VA: Association for Supervision and Curriculum Development.

Van Tassel-Baska, J., Feng, A. X., Brown, E., Bracken, B., Stambaugh, T., French, H., & . . . Bai, W. (2008). A study of differentiated instructional change over 3 years. *Gifted Child Quarterly*, *52*, 297-312.