J Educ Change
DOI 10.1007/s10833-015-9269-x

Tracking instructional quality across secondary mathematics and English Language Arts classes

Morgaen L. Donaldson[1] · **Kimberly LeChasseur**[1] · **Anysia Mayer**[2]

Abstract Teachers have the largest school-based influence on student learning, yet there is little research on how instructional practice is systematically distributed within tracking systems. We examine whether teaching practice varies significantly across track levels and, if so, which aspects of instructional practice differ systematically. Using multilevel modeling, we find that teachers of low track classrooms provided significantly less emotional support, organizational support, and instructional support to students in their classes than did teachers of high track classrooms. Mathematics classes were also observed to have higher quality instructional support for both content understanding and analysis and problem solving than English classes. We develop cases illustrating how small but significant differences in instructional quality are associated with substantially diverging lived experiences for students in high and low track classes.

Keywords Instructional quality · Teacher quality · Tracking · Secondary schools

Introduction

Research suggests that the quality of instructional practices in secondary schools may vary across instructional tracks. First described as a means to serve different student needs more efficiently (Ansalone 2010), some have posited that tracking allows students who need additional support to receive it without holding back students who are ready for more challenging work (Kulik and Kulik 1982). Many argue, however, that grouping students by perceived ability level actually widens

✉ Morgaen L. Donaldson
morgaen.donaldson@uconn.edu

[1] University of Connecticut, Mansfield, CT, USA

[2] California State University-Stanislaus, Turlock, CA, USA

Published online: 03 February 2016

the learning and achievement gap between groups (Lucas and Berends 2002). Indeed, tracking has been shown to suppress the achievement and attainment of students placed in low tracks (Moller and Stearns 2012; Oakes et al. 1992).

Tracking (and detracking) policies have consequently served as a lever for educational change in the form of improved classroom instruction. In response to growing evidence and political sentiment that tracking reifies inequities, some schools have attempted to detrack classes to create more equitable opportunities to learn (Wells and Serna 1996; LaPrade 2011). At the same time, initiatives such as Gifted and Talented and Advanced Placement, which operate on the notion that students should be separated into groups and provided different resources, persist in US schools.

Little research examines instructional practices and quality at the secondary school level, and even fewer studies have explored whether instructional practices are equitably distributed within schools (Hill et al. 2008; Hill et al. 2012). If tracking fails to distribute instructional practices equitably within schools, then focusing on this organizational arrangement will be unlikely to change—let alone improve—educational opportunities for students. In this paper, we ask whether students in lower track courses receive significantly lower-quality instruction than their peers in upper track courses. Specifically, we examine whether teaching practices in six high schools differ across tracks and whether they differ across schools serving more and less affluent populations.

Literature review

Tracking

The debate over tracking is long-standing (Moller and Stearns 2012; Oakes et al. 1992). While many educators believe that sorting students for instructional purposes and providing students with differentiated curricula is the best way to educate all students, research finds that this is not the case (Diamond et al. 2004). Students' race and family background, rather than more direct indicators of academic need or ability, are often used for sorting (Oakes and Guiton 1995; Riehl et al. 1999).

Qualitative studies of tracking illustrate that low-quality teaching is related to a range of antecedents to student achievement, including lack of engagement (Hand 2010), identity formation (Nunn 2011), and low self-efficacy (Rubin 2003). By being placed in lower tracks, students from families with low levels of material wealth may receive lower-quality academic preparation than their more advantaged peers (Ansalone 2010; Mickelson 2001; Worthy 2010). Several researchers have thus identified tracking as a primary lever driving social reproduction in schools (Lucas 1999; Bourdieu and Passeron 1990).

The manner in which learning opportunities are structured for students differs across tracks. Dreeben and Gamoran (1986) observed that elementary teachers grouped students according to perceived ability and adjusted the pace of instruction accordingly. This study was one of the first to suggest that teachers' instructional practices contribute to differences in achievement along racial and ethnic lines.

Subsequent studies documented that secondary school classrooms also differentiate the content of the curriculum according to track. High-track classes, like college preparatory courses, tend to have better qualified teachers and more engaging curriculum focused on critical thinking, while low-track, remedial, and vocational courses tend to have less qualified teachers and curriculum emphasizing rote memorization of basic facts (Oakes et al. 1992).

These tracks tend to be fairly rigid, truly offering distinct paths of coursework to students. It is difficult for students to move from low-track to high-track classes as students would have to catch up while also moving more quickly (Ayalon and Gamoran 2000). Researchers also find that most courses follow a strict vertical sequence from eighth grade to high school, precluding students from switching tracks after 8th grade (Stevenson et al. 1994). The cumulative result of these differences in experiences is dramatic differences in student outcomes by track (Van Houtte 2004). As such, tracking plays a key role in social reproduction as schools and teachers shy away from recognizing growth in student performance or errors in student placement that would necessitate shifting students across tracks.

Based on an analysis of a national probability sample of elementary and secondary school teachers, Oakes et al. (1990) found that children in "low-ability" classrooms had markedly less access to challenging instruction focused on inquiry and problem-solving in their math and science classes than did their peers in "high-ability" classrooms. In fact, students in "high-ability" classrooms in schools that enrolled a high proportion of low-income students actually had less exposure to such learning opportunities than did "low-ability" students at higher-income schools. This suggests that the value of tracking is not necessarily the same across school contexts.

Students who are placed in lower level courses in high schools tend to remain in them across multiple subjects (Author 2008; Rubin 2003) and throughout their high school career (Archbald et al. 2009; Kelly 2004). We also know that many teachers hold different expectations of students across academic tracks (Harris 2012; Weinstein 1996; Worthy 2010). However, very few studies have employed systematic observations of classrooms in multiple school districts to document whether teachers' instructional practices vary across tracks (Boaler 2000; Dreeben and Gamoran 1986; Watanabe 2008).

Instructional quality and its role in schooling

Researchers have confirmed what parents and educators have long known to be true: teacher quality varies. Hanushek (1992) found that teachers at the 95th percentile of his distribution produced 1.5 years of achievement growth, while those at the 5th percentile produced only .5 years of growth. Studies have illustrated the effects of poor quality teaching on a range of antecedents to academic achievement. The effects of being in a classroom with a low-quality instructor are cumulative (Sanders and Rivers 1996). Moreover, researchers have found more variation in teacher effectiveness within schools than between schools (Rivkin et al. 2005).

The importance of teacher quality in shaping student learning and experiences is widely recognized (Aaronson et al. 2007; Chetty et al. 2013; Rockoff 2004; Rivkin et al. 2005). However, what constitutes "quality" is still debated. A substantial

contingent of policymakers and researchers define teacher quality as the ability to improve student outcomes and, more specifically, standardized test scores (Kupermintz 2003; McCaffrey et al. 2003). Since 2009, 46 states have revised their teacher evaluation policies to increase emphasis on differentiating and improving teacher quality (Steinberg and Donaldson 2015). Many of these new policies weigh student achievement on standardized tests heavily in teachers' evaluation ratings (NCTQ 2013).

Instead of assessing teacher quality through teachers' effects on student outcomes, this study frames teacher quality in terms of practices that have been demonstrated to influence student engagement and learning. The CLASS-S is an observational instrument used to assess secondary school teachers' instruction across three domains: emotional support, organizational support, and instructional support (Hamre and Pianta 2005). Our analyses focus on the practices that teachers employ in the classroom to provide supportive environments and relationships (Roorda et al. 2011; Cornelius-White 2007), manage student behavior (Pace and Hemmings 2007; Flannery et al. 2009), and challenge students to engage in rigorous, scaffolded learning activities (Burris et al. 2008; Lee and Smith 1999). Self-determination theory posits that students will engage in learning when they feel competent, trusted, and autonomous (Connell and Wellborn 1991; Ryan and Deci 2000).

In addition to emotional support, teachers can influence students' social and academic outcomes through organizing the classroom environment (Emmer and Stough 2001). Providing clear expectations for students, coupled with redirection of negative behavior and positive reinforcements for acceptable behavior, is more effective in minimizing disruptive behaviors than strictly punitive systems of control (Flannery et al. 2009). Classrooms with established routines for organizing time engage students in more time on academic tasks (Bohn et al. 2004; Cameron et al. 2005). Providing students with a range of activities that require active participation can prevent disengaged students from disrupting class (Bowman and Stott 1994).

With an emotionally supportive classroom culture and effectively managed behaviors, teachers can engage students in learning activities. This study draws on three lines of research describing effective instruction. First, a number of studies have demonstrated the importance of supporting students' learning of core concepts through strategies such as exploring definitions and locating new concepts across a variety of contexts (Asquith et al. 2007; Woodward and Brown 2006). Another line of research emphasizes the need to teach students how to analyze new information and solve problems (Hiebert and Wearne 2003; Merrienboer and Stoyanov 2008). Finally, research has found that assisting students in developing metacognitive skills can boost the effectiveness of instruction (Veenman et al. 2005; Williams et al. 2002).

Methods

We present findings in this paper from sequential, mixed methods analyses (Creswell 2009) that draw on qualitative observational data to construct quantified ratings of instructional practice. We use multilevel modeling (Luke 2004) to address two research questions:

1. Does teaching practice, in terms of supporting a positive classroom climate, classroom organization, and instruction, vary significantly across track levels? If so, how?
2. Does teaching practice, in terms of supporting a positive classroom climate, classroom organization, and instruction, vary significantly across the socioeconomic status of the population served? If so, how?

A multilevel design is appropriate given the nature of our data, which describe the quality of teachers' instruction through observation of teacher practice at the classroom level over multiple visits. Group-level characteristics—whether the classroom is designated as a low track or high track class and whether the school serves a more or less affluent population, on average—are not only about classrooms, but about larger institutional structures within and across schools. To avoid the atomistic fallacy of attributing conclusions based on individual units (e.g., classroom observations) to larger groups (e.g., tracks and districts), we use multilevel analysis (Hox 2010).

Sample

We use a two-stage, stratified sampling technique. As part of a larger study, six districts in one northeastern state were selected purposively to represent school settings across the state. Three pairs of districts were selected, with one representing a less affluent population (based on percentage of free- or reduced-price lunch) and one representing a more affluent population, based on the state's Department of Education classification system. Each district has a single high school.

Classrooms were selected to reflect the ratio of core English courses offered at various track levels for each high school (see Table 1). This sampling strategy ensures that we did not oversample courses that were not normative experiences for large groups of students. We determined the number of all core English classes offered for each course at each level; for example, the number of Honors English II classes. We determined the ratio of each course to all English Language Arts courses being offered. This set of ratios was then used to determine the sample of classes to observe. We recruited teachers and more than 90 % of those who were approached agreed to participate. Between 20 and 40 % of the core English courses offered were observed at each high school (see Table 1).

Across the 6 schools, 149 classrooms were observed three times each, for a total of 427 observations. We observed 75 English classes (see Table 1). Of the 149 classrooms observed, 58 % (n = 87) were low track and 42 % (n = 62) were high track classes. All classes included in the low track sample were the lowest track available for graduation requirements. Depending on the high school's curriculum, ELA courses include some variation on American Literature, composition, survey courses, and World Literature, also taught at multiple levels. We chose not to include electives in mathematics and English Language Arts curricula on the basis that they also do not represent the core curricula of the schools.

In our sample, 45 % of teachers taught low track classes and 28 % taught high track classes exclusively. The remaining 27 % taught both low and high track

Table 1 Sampling of classrooms observed by subject, track level, and affluence within districts

	Subject						Track level		Affluence
	ELA			Math			Low track	High track	
	Sections offered	Sections observed		Sections offered	Sections observed		Sections observed	Sections observed	Sections observed
District A	81	22	27 %	74	21	28 %	27	16	Not affluent
District B	39	10	26 %	31	12	39 %	11	11	Affluent
District C	37	12	32 %	39	12	31 %	14	10	Not affluent
District D	51	10	20 %	44	12	27 %	17	5	Affluent
District E	44	11	25 %	27	9	33 %	7	12	Not affluent
District F	29	10	34 %	28	8	29 %	11	8	Affluent
Total	281	75	27 %	243	74	30 %	87	62	

classes. Women taught 80 % of the ELA sample. There were no significant differences in the assignment of teachers to tracks across content area [x^2 (2) = .11, $p = 0.948$] or across gender [x^2 (2) = 3.92, $p = 0.141$].[1]

Measures

To measure instructional quality, we use the CLASS-S protocol developed by Hamre and Pianta (2005)[2] (see footnote 1) to assess interactions between teachers and students along ten dimensions, each of which is scored on a Likert scale from 1 (low) to 7 (high). These dimensions are then aggregated into three domains, each representing an element of classroom experience that has been empirically linked to student learning. Scores for these three domains are modeled in our analyses as indicators of instructional quality at observation t in classroom i ($DOMAIN_{ti}$).

Based on research associating higher CLASS ratings with increased positive social behaviors (Mashburn et al. 2008), stronger peer relationships (Pianta and Hamre 2005), greater student engagement (Ponitz et al. 2009), and higher student achievement (Ponitz et al. 2009), the CLASS has been widely used to measure instructional practices. The CLASS was selected as one of only two general classroom observation protocols for the Measures of Effective Teaching Project, which was funded by the Bill and Melinda Gates Foundation to examine the relationships among a range of measures of teacher quality. Findings from this large-scale study indicate that higher CLASS-S ratings are associated with greater

[1] Teachers were reluctant to share information that might be used to identify them within their departments. A small portion of teachers (22 %) agreed to provide us with demographic data. In this subsample, 91 % of teachers were certified and 82 % held a master's degree or higher. Those without a master's degree or certification were clustered in one rural district and were evenly distributed across content area and track assignments.

[2] CLASS protocols are designed to assess similar constructs across the span of developmental stages from early childhood development through secondary school. At the time of data collection, the CLASS-S protocol was being piloted by the developers. However, the instrument has been validated at the K-3 level and domain scores are significantly related to student achievement and other key student outcomes (Ansalone 2010; Ponitz et al. 2009).

student achievement (Allen et al. 2011) and larger teacher effects in value-added scores (Kane and Staiger 2012). Mikami et al. (2011) also demonstrated that interventions providing teachers with feedback using the CLASS-S can lead to increased positive peer interaction in high school classrooms. This growing set of studies suggests that the CLASS-S is a useful tool for examining the quality of instructional practice across classrooms.

Emotional support

Scores for four dimensions of classroom interactions—positive climate, negative climate, teacher sensitivity, and regard for adolescent perspectives—are averaged to create a composite domain score for each cycle. In the CLASS-S protocol, positive climate involves four constructs: relationships, positive affect, positive communications, and respect. Negative climate is enacted through negative affect, punitive control, and disrespect. Teacher sensitivity is demonstrated through awareness, responsiveness, effectiveness in addressing problems, and student comfort. Regard for adolescent perspectives involves support for student autonomy and leadership, connections to current life, student ideas and opinions, meaningful peer interactions, and flexibility.

Organizational support

Scores for three dimensions—productivity, behavioral management, and instructional learning formats—are averaged to create a composite domain score for each cycle. The CLASS-S protocol considers productivity to include maximizing learning time, routines, and transitions. Behavioral management in enacted via clear expectations, proactive, effective redirection of misbehavior, and student behavior. Instructional learning formats include learning targets/organization, variety of materials, modalities, and strategies, active facilitation, and effective engagement.

Instructional support

Scores for three dimensions—concept development, problem solving and analysis, and quality of feedback—are averaged to create a composite domain score for each cycle. In the CLASS-S protocol, concept development assesses depth of understanding, communication of concepts and procedures, background knowledge and misconceptions, transmission of content knowledge and procedures. Analysis and problem solving involves opportunities for higher level thinking, problem solving, and metacognition. Quality of feedback measures feedback loops, prompting thought processes, scaffolding, providing information, and encouragement and affirmation.

We used these measures as predictors to assess patterns across CLASS domain scores:

Low track In three of the schools in our sample, two levels were offered for core courses; the other three schools had a three-tier system. In order to assess any differences in instructional quality across course levels, we created a dummy

variable for the lowest track (LOWTRACK): any classroom in the bottom level for the school was scored as a standard level class and all others were considered "high track."

Mathematics To control for any differences in domain scores across subject area, we use a dummy variable for MATH (1 = mathematics course; 0 = English course).

Affluent A dummy variable for AFFLUENT (1 = affluent school; 0 = less affluent school) indicated whether the classroom was part of a school serving an affluent population.

Observation We collected multiple observations of each classroom (CYCLE).

Data collection

Data were collected by the authors and another faculty member between January and June 2012. Each classroom was observed by two different observers, with observers observing separate lessons.[3] Consistent with recommended procedures for implementing CLASS-S, observations occurred in cycles starting at the beginning of the class, with the observer spending 15 min watching interactions between the teacher(s) and students while taking notes; the observer then stopped, coded observation notes, and scored each CLASS dimension. This process was repeated as many times as the class period permitted. In most cases, all observations of individual teachers were completed within 2–3 weeks.

Observers took extensive field notes on each class. Notes addressed the three domains and indicators of the instrument and provide a more nuanced picture of the interactions in the classes observed. We used these notes to build brief cases that contextualize the quantitative data. Thus, while we draw on both quantitative (CLASS-S) and qualitative (observation) data in this paper, our primary instrument was quantitative and we foreground those data here.

Reliability

Observers were trained and certified by Teachstone, the organization authorized to certify CLASS observers. A one-way ANOVA was performed to explore differences in how observers coded their subsample of classrooms. Although there are some significant patterns of slight differences across observers, they are all within the acceptable reliability parameters.[4]

These data were collected as part of a larger study. With limited resources, we decided to expand the saturation of classes observed to represent at minimum 20 % (and in most cases, more than 25 %) of the core classes per school rather than to halve our sample by double-scoring observations. This has implications for the reliability of our data. However, we were able to investigate observers as a potential

[3] Although double-scoring lessons has been demonstrated to increase the stability of observation ratings, this was not feasible given the scope of the live observations in the study.

[4] According to the official certification training, observers may still be considered reliable when they are one point (on a seven-point scale) away from a master coder's rating in either direction.

source of bias by conducting linear regressions with two blocks to examine whether the observer predicts a significant portion of the variability in domain scores after controlling for the variables of interest. The observer was not a significant predictor for any domain score for any cycle. Adding a second observer to observations would therefore have had little effect on the reliability of our data.

Missing data

In about 20 % of classrooms, we were only able to collect two observation cycles due to scheduling constraints, resulting in 31 cases without a third cycle. Data were missing at random, evenly distributed across track levels and subjects. There were no significant differences in emotional, organizational, or instructional domain scores between the second and third observations for those with complete data [$t(119) = 1.29$, $p = .20$; $t(119) = 1.03$, $p = .30$; $t(119) = -.12$, $p = .90$, respectively], suggesting that domain scores were relatively stable from second to third observations. We imputed missing domain scores in these cases using linear trend at point estimation (Allison 2002). No other variables were missing any data.

Analysis

We use two-level multi-level modeling to assess the relationships between instructional quality (at the observation level) and both track level (at the classroom level) and three-level modeling to assess the relationships between instructional quality (at the observation level) and affluence (at the school level) (Luke 2004). Null models (Model 0) were calculated to examine intraclass correlations.

$$\text{Model 0: } DOMAIN_{ti} = \beta_{00} + r_{0i} + e_{ti}$$

More than 30 % of each domain score can be explained at the level of time, indicating that it is necessary to include the effects of observation cycle at the first level. Models were then built in three stages. In the first stage, we used a random slopes model that allowed the influence of time across observation cycles to vary (Model 1).

$$\text{Model 1: } DOMAIN_{ti} = \beta_{00} + \beta_{10}{*}CYCLE_{ti} + r_{0i} + r_{1i}{*}CYCLE_{ti} + e_{ti}$$

Because the slope was not found to vary significantly for any domain score, we did not allow it to vary in subsequent models. In the second stage, we added classroom level predictors indicating subject and track level (Model 2).

$$\text{Model 2: } DOMAIN_{ti} = \beta_{00} + \beta_{01}{*}MATH_i + \beta_{02}{*}STANDARD_i + \beta_{10}{*}CYCLE_{ti} + r_{0i} + e_{ti}$$

In the final stage, we added a school level predictor indicating affluence (Model 3).

$$\text{Model 3: } DOMAIN_{ti} = \gamma_{000} + \gamma_{001}{*}\text{AFFLUENT}_k + \gamma_{010}{*}\text{MATH}_{jk} + \gamma_{020}{*}\text{STANDARD}_{jk} + \gamma_{100}{*}CYCLE_{ijk} + r_{0jk} + u_{00k} + e_{ijk}$$

In all analyses, final models displayed a better fit to the data than the first model.

Limitations

Like all research, our study has limitations. First, we conducted our school observations using the CLASS-S protocol, which was at that time in development and has since been finalized. While the domains remain the same, one stand-alone dimension (student engagement) that was present in the pilot version of CLASS-S did not remain in the instrument. We do not use this dimension in our analyses for this paper. Moreover, the dimensions and domains examined in this paper were shown to be valid and reliable.

Second, this paper presents findings that are not adjusted for teacher fixed effects. This does not raise questions for our analysis, but it poses issues for interpretation. If teachers with different instructional styles are systematically assigned to different tracks, instructional differences would be due to a teacher's style rather than her response to her students and their particular track. Are teachers with different styles tracked into different student assignments or are teachers adjusting their style based on the students within particular tracks? Our sample size did not permit us to analyze which of these explanations applies in the case of our data, although we discuss some initial analyses below.

Third, the CLASS-S is first and foremost an instrument to collect quantitative data. While we collected qualitative data through field notes, these were less comprehensive than the quantitative data collection. As such, we foreground the quantitative data in our analysis and presentation.

Findings

Low track classes in our sample were observed to have significantly lower emotional, organizational, and instructional support than high track classes, confirming that the low track classrooms in our sample provided systematically lower-quality educational experiences for high school students than those offered by high track classrooms. We also found significant differences according to the affluence of the population served by the school. In describing these findings, we present estimates of differences in classroom quality within each CLASS-S domain. Using qualitative field note data, we include brief profiles of classroom practices at different levels of observed quality to illustrate how substantial even small differences in CLASS ratings can be for students.

Instructional practices across tracks

Emotional support

Across our sample of classrooms, the emotional support provided to students through positive climate, reduced negative climate, teacher sensitivity, and regard for student perspectives in standard level classes was significantly lower than that provided in higher level classes [$t(147) = -3.44$, $p < .001$]. On average, low track

classrooms provided emotional support that fell in the middle of the CLASS quality spectrum, compared to quality in the high end observed in high track classrooms. All four dimensions of emotional support—positive climate, negative climate, teacher sensitivity, and regard for adolescent perspectives—were lower in low track classes than in high track classes (see Table 2). The difference of .46 on a seven-point Likert scale might seem inconsequential, but there are very real advantages for students tracked into these high track classes (see Table 3). The Cohen's effect size for the domain ($d = .56$) suggests a moderate practical significance for being placed in a high track class, and prior research suggests even minor differences between tracks accumulate over a student's career (see Oakes et al. 1990).

To illustrate the difference between being tracked into a class with moderate emotional support and being enrolled in a class with high emotional support, let us consider two exemplar classrooms that we observed. In both classrooms, the teacher created a highly positive climate with evidence of personal relationships with students. The teacher and students alike offered gestures of respect, such as saying "please" and "thank you." Both teachers also displayed fairly high degrees of teacher sensitivity. Students were comfortable volunteering ideas. At one point, the teacher in the standard-level class noticed a pair of students who were not working on the assigned task and went over to them to refocus them on the task; the students subsequently started working on the task. She also monitored student comfort by asking if students were comfortable continuing the roles they had volunteered for earlier in the lesson and by extending the time for a writing task when most students were still actively engaged as they approached the deadline she had originally established. While both classrooms were observed to have positive climates and demonstrated teacher sensitivity, there were striking differences in the two teachers' regard for adolescent perspectives and incidences of negative climate.

Differences in teachers' regard for adolescent perspectives

The CLASS observation tool includes five indicators of teachers' regard for adolescent perspectives: support for student autonomy and leadership, connections to current life, student ideas and opinions, meaningful peer interactions, and flexibility. Both English Language Arts classes were reading *A Raisin in the Sun*, with students reading the play aloud in class and discussing themes and character development. The teacher leading the exemplar low track class did not offer many opportunities for students to lead their own or each other's learning. There were no opportunities for students to set the course for their own learning beyond volunteering to read a character's part in the play. Students occasionally shared their interpretations of scenes, but the teacher did not validate these interpretations by integrating them into her instruction. For example, at one point, the teacher had difficulty following a student's comment analyzing the effects of skin color. After the student talked about how skin color was important to the character, the teacher replaced the student's answer with "racism" and moved on, rather than agreeing with the spirit of the student's answer and allowing the students to build their own analysis from this comment (District A, low track, observed 30/3/12).

In contrast, upper level classrooms were more likely to provide a learning environment that respected and encouraged student perspectives. While the teacher

Table 2 Differences in dimensions of instructional practice across track levels and affluence

Instructional practices	Low track		High track		df	t	Not affluent		Affluent		df	t
	M	SD	M	SD			M	SD	M	SD		
Emotional support												
Positive climate	5.26	1.23	5.85	0.90	147	3.37*	5.31	1.15	5.81	1.05	147	2.73*
Negative climate	1.32	0.45	1.13	0.37	147	−2.86*	1.29	0.47	1.16	0.36	147	−1.88*
Teacher sensitivity	4.81	1.26	5.40	1.02	147	3.19*	4.87	1.18	5.33	1.18	147	2.33*
Regard for adolescent perspectives	3.54	1.37	4.07	1.37	147	2.34*	3.53	1.41	4.09	1.31	147	2.50*
Organizational support												
Behavior management	4.92	1.37	5.98	0.97	147	5.53*	5.11	1.34	5.75	1.21	147	3.01*
Productivity	4.99	1.34	5.65	1.01	147	3.41*	5.17	1.34	5.42	1.10	147	1.20*
Instructional learning formats	4.05	1.25	4.80	1.04	147	3.94*	4.16	1.16	4.68	1.25	147	2.65*
Instructional support												
Content understanding	3.18	1.19	3.85	1.13	147	3.48*	3.24	1.13	3.79	1.26	147	2.81*
Analysis and problem solving	2.82	1.27	3.38	1.41	147	2.53*	2.73	1.24	3.53	1.37	147	3.74*
Quality of feedback	2.99	1.45	3.26	1.42	147	1.11*	3.00	1.49	3.26	1.36	147	1.12*

* $p < .05$

Table 3 Parameter estimates for emotional support

Fixed effects	Model 1			Model 2			Model 3		
	Coef.		*p*	Coef.		*p*	Coef.		*p*
Intercept	5.27	(.09)	<.001	5.57	(.13)	<.001	5.39	(.06)	<.001
Cycle slope	−.03	(.05)	.60	−.04	(.03)	.18	−.04	(.05)	.38
Math				−.09	(.13)	.48	−.08	(.12)	.53
Standard				−.46	(.13)	<.001	−.45	(.06)	<.001
Affluent							.37	(.08)	.008
Random effects	SD		Var. comp.	SD		Var. comp.	SD		Var. comp.
r_{0j}	.77		.60	.73		.53	.72		.52
r_1	.14		.02						
e	.50		.25	.53		.28	.53		.28
u_{00}	.16		.03	.15		.02	.01		<.001
u_{10}	.10		.01						
Model fit									
Deviance	1004			999			993		
Parameters	12			7			8		
AIC	1028			1013			1019		

Standard error in parentheses

in the low track class gave students the choice of which part to read aloud in *A Raisin in the Sun*, the teacher in the upper level class asked students to independently discuss who they could relate to the most in the play. While students shared connections between characters and their own lives, the teacher asked questions to probe students to deepen their analyses, and reinforced students' opinions rather than trying to assert her own. The class then moved into a discussion of evil and the teacher listened while students argued their own perspectives. The group discussion provided opportunities for students to interact with their peers by sharing different ways they thought about characters and themes from the reading (District A, high track, observed 30/3/12 and 4/4/12).

Differences in negative climate

Our two case study classes also differed in the extent to which they featured a negative climate.[5] The standard-level class had a higher score for negative climate than did the upper-level class. In response to student joking, the teacher offered a

[5] The CLASS-S measures positive climate and negative climate as two separate constructs, rather than as opposite ends of a single spectrum. A classroom can exhibit a highly positive climate in terms of relationships, positive teacher and student affect, positive communications, and respect. At the same time, that classroom may display one or more indicators of negative climate, such as an inappropriately sarcastic comment from the teacher, punitive control of students, or other signs of disrespect. According to the CLASS-S, any instance of negative climate is automatically reflected in an imperfect negative climate score.

comment in a cutting tone—"OK, that's enough of that now. We're not in second grade." During another observation cycle, she used the threat of punitive control to get the behavior she wanted from students. Both of these instances led to the classroom receiving a score at the upper end of low for negative climate (District A, low track). Although this is only a single point difference, we did not observe the teacher in the high track class use sarcasm with students or threaten them with punishment. In fact, 40 % of the low track classes we observed had at least one instance of negative climate. In contrast, we only observed instances of negative climate in 14 % of the high track classes in our sample.

This set of contrasts is one of many ways that classrooms could have relatively close scores on the CLASS-S rubric for the Emotional Support Domain, yet offer substantially different experiences to students. They demonstrate how critical even half a point difference in a domain score can be in shaping the instruction provided to students.

Organizational support

The quality of organizational support for students in upper level courses was significantly higher than that provided to students in lower level courses in our sample [$t(147) = -5.00, p < .001$]. Teachers in low track classrooms demonstrated organizational support .78 points lower than those in high track classrooms (see Table 4). All three dimensions of organizational support (behavior management, productivity, and instructional learning formats) were significantly higher in high

Table 4 Parameter estimates for organizational support

Fixed effects	Model 1			Model 2			Model 3		
	Coef.		*p*	Coef.		*p*	Coef.		*p*
Intercept	5.05	(.12)	<.001	5.38	(.16)	<.001	5.15	(.17)	<.001
Cycle slope	−.04	(.05)	.46	−.05	(.04)	.23	−.05	(.04)	.28
Math				.24	(.15)	.11	.26	(.16)	.10
Standard				−.78	(.16)	<.001	−.79	(.11)	<.001
Affluent							.46	(.11)	.01
Random effects	SD		Var. comp.	SD		Var. comp.	SD		Var. comp.
r_{0j}	.95		.91	.84		.71	.83		.69
r_1	.40		.16						
e	.56		.32	.69		.48	.69		.48
u_{00}	.22		.05	.18		.03	.01		<.001
u_{10}	0.04		<.00						
Model fit									
Deviance	1215			1210			1203		
Parameters	9			7			8		
AIC	1233			1224			1219		

Standard error in parentheses

track classes than in low track ones (Table 2). The Cohen's effect size for this domain is high ($d = .82$), suggesting the difference in organizational support for students in high and low track classes was practically significant.

To put this finding into perspective, we examine two classes with mean organizational support domain scores near the predicted norms for low track and high track classrooms. Both classes had a high degree of productivity. Students had routines for tasks, such as passing around handouts, that seemed to help them transition quickly into the work of the class period. The teachers of both classes maximized learning time, a CLASS-S indicator, through efficient pacing (District A, low track, observed 9/4/12; high track, observed 2/4/12 and 9/4/12).

The low track class had a moderate degree of behavioral management. Expectations for student behavior were inconsistent. For example, sometimes the teacher allowed students to talk with one another, and at other times she told them to stop talking. While some students were not working on the specified task some of the time, the students were primarily attending to the assigned work for the duration of class (District A, low track). However, the high track class also had a moderate level of behavioral management. Although the teacher asked students who were talking with each other about non-class related topics to stop talking, she did state why she needed them to be quiet. When students did not stop talking, the teacher threatened them with additional homework. Despite this reactive approach to managing the students, misbehavior did not escalate and the class was generally responsive to the teacher and on-task (District A, high track).

Differences in the instructional learning formats provided

The two classes differed most dramatically in the instructional learning formats used in the lesson. Instructional learning formats represent the instructional "hooks" used to engage students in their learning, such as providing learning targets and offering a variety of modalities, strategies, and materials. The low track class in our case study offered a moderate degree of instructional learning formats. The teacher connected the goals for the class period to a previous lesson and gave some learning targets. The lesson was also actively facilitated by the teacher, with explanations of how to move through the tasks at hand. However, there was limited variety of pedagogies or materials. The class went over questions that the students had already completed; there were no other means of learning the material offered and students did not have any options for working through concepts in a variety of ways (District A low track).

This was in striking contrast to the high track class. The teacher began the first day of observation by giving the students a graphic organizer and connecting the lesson to concepts they had previously covered. The teacher reviewed the concepts with students quickly. Students were then given a new set of tasks and the teacher facilitated student problem-solving in group discussion (District A, high track). On the second day the class was observed, the teacher had three students act out a skit to demonstrate new concepts (District A, high track). These comparisons illustrate significant and substantial differences in the ways classrooms were organized for students in low and high tracks within the same high school.

Instructional support

Instructional support captures aspects of teaching practice perhaps most proximal to student learning. Although the classrooms we observed had, on average, fairly low levels of instructional support, there was significantly more instructional support in high track classrooms than in low track classrooms [$t(147) = -3.99, p < .001$]. The difference in the instructional support provided across track levels is .61 points on a seven-point Likert scale. Support for content understanding and for analysis and problem solving was significantly higher in high track classes than in low track classes; there was no significant difference in the quality of feedback teachers provided across the two track levels (see Table 3). The Cohen's effect size for this domain is moderate ($d = .65$), suggesting a practical significance in the instructional support provided to students in high track classes, as compared to that provided to students in low track classes.

Differences in supporting content understanding

To illustrate the differences across levels of support for content understanding, we compare low track and high track World Literature classes at the high school in District A. Both classes were on the same lesson, which involved learning how to parse persuasion into types based on whether the arguments rely on emotion (pathos), logic (logos), or ethical duty (ethos). Despite working on the same overall lesson, there were meaningful differences in the way the teachers supported content understanding across these two levels of English classes. Content understanding was indicated by promoting depth of understanding, communicating concepts and procedures, attending to background knowledge and misconceptions, and transmitting salient content knowledge. The teacher in the low track class of our case study led the students in reading aloud. She transmitted content knowledge at a fairly low level by spending time explaining new vocabulary words to the students. She also helped students to increase the depth of their understanding by having them brainstorm words to describe the text they were discussing.

However, there was no attention to discovering or correcting misunderstandings, discussing procedures for analyzing the argument in the text, or communicating and discussing the main concepts of the lesson. For example, the teacher had students spend a lot of time describing the products being sold in the ads and the intended audience without explaining why they were doing so or prompting students to connect why they were discussing content and audience to how they could determine the type of persuasion being used (District A, low track, observed 23/3/12 and 30/3/12).

In contrast, students in the high track class doing the same lesson used these main concepts with more familiarity. Students in this class demonstrated some depth of understanding by using the terms "ethos," "pathos," and "logos" to describe advertisements, rather than using other words to talk about content and audience, as in the low track class (District A, high track, observed 3/30/12).

What is perhaps most noteworthy about this illustrative comparison of instructional support across levels of this class is that the differences are based on

supportive interactions, not on the content of the curricula. The findings from our observations about instructional support suggest that even with the same lesson, the quality of teaching practice for content understanding, analysis and problem solving, and quality of feedback play a critical role in different experiences for students in low track classes, compared to their peers placed in high track classes of the same course.

Instructional practices across affluence

In addition to examining the relationship between track levels and teacher practice, we also examined whether there were differences in teachers' practices across schools serving more and less affluent populations. We found that teachers in affluent districts had significantly higher CLASS-S scores in emotional support and organizational support of students, but not in terms of instructional support. Our small sample did not allow us to assess interactions across track level and affluence, though we did find noteworthy trends.

Emotional support

Students enrolled in high school in affluent districts in our sample received higher quality emotional support than those enrolled in less affluent districts (see Table 3). Specifically, students in affluent districts had teachers who provide a more positive climate, displayed better teacher sensitivity, and held a higher regard for adolescent perspectives (see Table 2). There were no differences in the negative climates enabled by teachers.

As with track level, the differences are relatively small in terms of the CLASS-S scale. However, the effect size is large, with Cohen's d = 4.90 and an effect size of .93. Put another way, the effects of track and affluence are approximately the same size and in the opposite direction. A student in a high track classroom in a less affluent district received approximately the same emotional support from teachers as those educated in low track classes in an affluent district. The boost from being in a high track classroom is, in effect, negated by being in a less affluent district.

Organizational support

Students enrolled in high school in affluent districts also received higher quality organizational support in their classrooms than those in less affluent districts (see Table 4). Students in affluent districts received higher quality behavioral management and better instructional learning formats than those in less affluent districts (see Table 2). There were no differences in the productivity supported by teachers across affluent and less affluent districts.

The influence of affluence is again, small in scale with about half a point difference, but large in effect, with Cohen's d = 4.15 and an effect size of .90. The effects of track level and affluence do not entirely cancel each other out for organizational support, as they do for emotional support. The difference across affluence of .46 is about half of the difference across track levels, at −.79. In other

words, being educated in a high track class in a less affluent district makes up for not being in a more affluent district, but not enough to receive the same organizational support as being in a high track class in a more affluent district. As with emotional support, high track classes do not receive the same instructional quality across more and less affluent districts.

Instructional support

T-tests of the individual dimensions that comprise the domain of instructional support on the CLASS-S tool suggested significant differences in teacher practices supporting content understanding and analysis and problem solving (see Table 2). However, when we fit the data to a nested model, affluence was not a significant predictor of the quality of instructional support practices, as a whole (see Table 5). Given the differences in teachers' mean ratings at the classroom level and the significant relationships between instructional support and subject, we are hesitant to interpret our findings as conclusive in this area. We cannot conclude that the quality of instructional support is different across less affluent and more affluent settings; however, we believe our exploratory analyses warrant further investigation using larger samples and more complex analyses to detect more complicated relationships.

Table 5 Parameter estimates for instructional support

Fixed effects	Model 1			Model 2			Model 3		
	Coef.		*p*	Coef.		*p*	Coef.		*p*
Intercept	3.22	(.24)	<.001	3.39	(.27)	<.001	3.12	(.28)	<.001
Cycle slope	.02	(.07)	.77	.01	(.05)	.90	.01	(.08)	.93
Math				.37	(.15)	.01	.38	(.04)	<.001
Standard				−.61	(.15)	<.001	−.60	(.16)	<.001
Affluent							.53	(.44)	.29
Random effects	SD		Var. comp.	SD		Var. comp.	SD		Var. comp.
r_{0j}	.87		.76	.76		.57	.76		.57
r_1	.41		.17						
e	.74		.55	.85		.72	.85		.72
u_{00}	.54		.30	.57		.33	.50		.25
u_{10}	.11		.01						
Model fit									
Deviance	1344			1337			1336		
Parameters	9			7			8		
AIC	1362			1351			1352		

Standard error in parentheses

Discussion

We find that the instruction experienced by students in low tracks is of lower quality than that experienced by their peers in high tracks. This study is one of the first to analyze variation in instruction by secondary school track across multiple districts using an externally validated and reliable observation instrument. The findings confirm and extend the literature (see Oakes et al. 1990, 1992 e.g.) describing mechanisms inside classrooms that provide inequitable experiences to students in low track classes compared to the education provided to their high tracked peers. Although the magnitude of differences between high- and low-tracks were small, Oakes (1990) finds that the effects of tracking are "incremental" and even slight differences between tracks are critically important. Even small differences in instructional quality can accumulate over years, leading students who spend their educational careers placed in lower-tracks to receive a substantially lower-quality education than their peers who are consistently placed in high tracks (Oakes 1990).

Analyses of classroom culture using the emotional support domain of the CLASS-S demonstrate that instruction in the lowest tracks in these schools is more often characterized by negativity, insensitivity towards students, and lack of regard for student perspectives. Given what we know about the importance of care and consideration in the teacher-student relationship (Gay 2002; Gutiérrez and Rogoff 2003), this finding ought to raise concern. Students who are performing less well arguably need *more* positivity and emotional support to boost their self-efficacy and confidence in themselves and schooling (DeckerDecker et al. 2007; Kesner 2000).

Our analysis of organizational support found that teachers provided less classroom structure and less varied instructional activities for students in low track classes than in high track classes. Again, this is cause for concern. Research on instruction suggests that students learn best when they have the opportunity to interact with content through varied media (Costello 2012; Garcia and Guerra 2004; Miller 2010, 2013). The fact that lower-track students had less variety in classroom activities than their higher-track peers meant that they had fewer entry points through which to access and become engaged in the curriculum. In this way, tracking may cause them to fall even further behind their higher-track peers.

Findings related to the instructional support domain reveal systematically less teacher support for content understanding and for analysis and problem solving in low track classes. This is consistent with prior research (Oakes et al. 1990, 1992). As educators, we are committed to the belief that all students—regardless of current skill level—should be challenged to dig deeply into concepts and should be taught how to analyze new information. This does not have to mean that all students must learn the same content at the same pace, but it does involve an expectation that it is not acceptable for students in low track classes to be rarely challenged, as we found in the low track classrooms in our sample. There is ample research demonstrating the importance of rigor in stimulating learning (Lee et al. 1993; Barton and Coley 2009), even when students are achieving at low levels (Lee and Smith 1999; Ford and Moore 2013; Gamoran et al. 1997).

While the unit of analysis in these observations is teacher-student interactions, the CLASS-S dimensions capture these areas of instruction in ways that make it arguable that such differences can be attributed to teachers' classroom practices. For example, the Analysis and Problem Solving dimension of CLASS-S focuses on the extent to which "the teacher helps students to use higher-order thinking skills 'such as reasoning, integration, experimentation (e.g., hypothesis generation and testing), and metacognition (i.e., thinking about one's own thinking)' (Stuhlman et al. n.d., p. 3)". Teachers play an active role in creating opportunities for students to engage in higher-order thinking, designing activities that promote problem solving, and facilitating students' reflection on their own thinking. In this way, teachers design and provide *opportunities* for students to engage in learning experiences; differences across classrooms thus appear to be rooted in variations in whether teachers create these opportunities.

Implications for further research

This study suggests several areas for further research. First, this study should be replicated with additional samples drawn from a variety of contexts. This research should examine whether differences in instructional quality exist and probe any patterns that arise. In such research, scholars could pair the CLASS-S protocol with an equally robust qualitative data collection instrument, to yield a database of quantitative and qualitative data of similar quality. Such methods could produce important, multi-faceted knowledge about the mechanisms through which tracking works. Are teachers systematically changing their instructional techniques as they switch from track to track, as qualitative research suggests (Watanabe 2008)? How do teachers make instructional decisions and how does student track figure into their decision-making? Elsewhere (LeChasseur et al., under review) we explore variations in the instruction of the teachers in our sample who taught multiple tracks of the same course. We find that teachers expressed lower expectations and provided significantly less support, as measured by the CLASS dimensions, to students in low track classes than they did to those in high track classes.

Although we control for differences in the relationships between tracking and quality of instruction across subject area, additional research might extend our understanding of the complexities of instruction by probing these relationships more deeply. Further study of how CLASS-S and other general observation protocols align with subject-specific protocols, such as the PLATO for English and the MQI for mathematics, would provide useful information on how to interpret different ratings across subjects (Steinberg and Donaldson 2015). Future work that places common conceptualizations of instructional quality in the context of subjects (Grossman and Stodolsky 1995; Stodolsky and Grossman 1995) would help explain why ratings might vary across subjects without placing ratings (or teachers) within a competitive, relative hierarchy.

Implications for teacher evaluation policy

A moderate amount of the variation we observed can be explained by the observation cycle, suggesting that there is some instability in the observed quality of

instruction over time. This finding calls into question the extent to which we can consider instructional quality to be a fairly static attribute of teachers, even within the same short window of time. Many teacher evaluation systems presume that it is possible and appropriate to classify teachers' quality of instruction for an entire year with very few assessments. Our study adds empirical evidence to the growing debate over whether such measures offer valid and reliable means of differentiating who should and should not be allowed to lead students' learning.

If our findings are substantiated by further research, they raise serious issues regarding the persistence of tracking in US schools. Addressing the effects of tracking is primarily an issue of local policy: school leaders and teachers must examine the tracking and teaching practices within their own settings. School leaders should consider how they assign teachers to tracks—on what criteria do they make these judgments? At the same time, these findings call for school leaders to identify supports that could help teachers deliver higher-quality instruction to students placed in lower tracks. Our findings also call for teachers to examine more closely their own biases regarding teaching lower tracks and, implicitly, lower-income students.

Beyond in-service teachers, these findings have implications for pre-service teachers. Our findings support previous research on the importance of preparing teachers to address students' needs in all courses and interrogate their own biases about student ability (Abu El-Haj and Rubin 2009). Prompting pre-service teachers to become mindful of their instructional decision making in different settings might begin the conversation about how to mitigate differences in student experiences across tracks before students are directly affected. Our findings suggest we examine more carefully what is expected of students and what is denied them within the context of their high school classrooms—and that discussion should extend to all educators shaping these experiences.

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the chicago public high schools. *Journal of Labor Economics, 25*(1), 95–135.

Abu El-Haj, T. R., & Rubin, B. C. (2009). Realizing the equity-minded aspirations of detracking and inclusion: Toward a capacity-oriented framework for teacher education. *Curriculum Inquiry, 39*(3), 435–463.

Allen, J. P., Pianta, R. C., Gregory, A., Mikami, A. Y., & Lun, J. (2011). An interaction-based approach to enhancing secondary school instruction and student achievement. *Science, 333*, 1034–1037.

Allison, P. D. (2002). *Missing data*. Thousand Oaks, CA: Sage.

Ansalone, G. (2010). Tracking: Educational differentiation or defective strategy. *Educational Research Quarterly, 34*(2), 3–17.

Archbald, D., Glutting, J., & Qian, X. (2009). Getting into honors or not: An analysis of the relative influence of grades, test scores, and race on track placement in a comprehensive high school. *American Secondary Education, 37*, 65–81.

Asquith, P., Stephens, A. C., Knuth, E. J., & Alibali, M. W. (2007). Middle school mathematics teachers' knowledge of students' understanding of core algebraic concepts: Equal sign and variable. *Mathematical Thinking and Learning, 9*(3), 249–272.

Ayalon, H., & Gamoran, A. (2000). Stratification in academic secondary programs and educational inequality in Israel and the United States. *Comparative Education Review, 44*(1), 54–80.

Barton, P. E., & Coley, R. J. (2009). *Parsing the achievement gap II*. Princeton, NJ: Educational Testing Services.

Boaler, J. (2000). Students' experiences of ability grouping-disaffection and polarization. *British Educational Research Journal, 26*(3), 631–648.

Bohn, C. M., Roehrig, A. D., & Pressley, M. (2004). The first days of school in effective and less effective primary-grades classrooms. *Elementary School Journal, 104*, 269–287.

Bourdieu, P., & Passeron, J. P. (1990). *Reproduction in education, society and culture*. London: Sage.

Bowman, B. T., & Stott, F. M. (1994). Understanding development in a culture context: The challenge for teachers. In B. Mallory & R. New (Eds.), *Diversity and developmentally appropriate practices: Challenges for early childhood education* (pp. 19–34). New York: Teachers College Press.

Burris, C. C., Wiley, E., Welner, K., & Murphy, J. (2008). Accountability, rigor, and detracking: Achievement effects of embracing a challenging curriculum as a universal good for all students. *Teachers College Record, 110*(3), 571–608.

Cadima, J., Leal, T., & Burchinal, M. (2010). The quality of teacher–student interactions: Associations with first graders' academic and behavioral outcomes. *Journal of School Psychology, 48*(6), 457–482.

Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology, 43*, 61–85.

Chetty, R., Friedman, J.N., Rockoff, J.E. (2013). *Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates*. National Bureau of Economic Research Working Paper 19423.

Connell, J. P., & Wellborn, J. G. (1991). Competence, authonomy, and relatedness: A motivational analysis of self-system processes. In R. Gunnar & L. A. Sroufe (Eds.), *Minnesota symposia on child psychology* (Vol. 23, pp. 43–77). Hillsdale, NJ: Lawrence Erlbaum.

Cornelius-White, J. (2007). Learner-centered teacher-student relationships are effective: A meta-analysis. *Review of Educational Research, 77*, 113–143.

Costello, A. (2012). Multimodality in an urban, eighth-grade classroom. *Voices from the Middle, 19*(4), 50.

Creswell, J. (2009). *Research design: Qualitative, quantitative, and mixed methods approaches*. Thousand Oaks, CA: SAGE Publications, Incorporated.

Decker, D. M., Dona, D. P., & Christenson, S. L. (2007). Behaviorally at-risk African American students: The importance of student–teacher relationships for student outcomes. *Journal of School Psychology, 45*(1), 83–109.

Diamond, J., Randolph, A., & Spillane, J. (2004). Teachers' expectations and sense of responsibility for student learning: The importance of race, class, and organizational habitus. *Anthropology and Education Quarterly, 35*(1), 75–98.

Dreeben, R., & Gamoran, A. (1986). Race, instruction, and learning. *American Sociological Review, 51*(5), 660–669.

Emmer, E. T., & Stough, L. M. (2001). Classroom management: A critical part of educational psychology, with implications for teacher education. *Educational Psychologist, 36*(2), 103–112.

Flannery, K. B., Sugai, G., & Anderson, C. M. (2009). School-wide positive behavior support in high school: Early lessons learned. *Journal of Positive Behavior Interventions, 11*(3), 177–185.

Ford, D. Y., & Moore, J. L, I. I. I. (2013). Understanding and reversing underachievement, low achievement, and achievement gaps among high-ability African American males in urban school contexts. *The Urban Review, 45*, 399–415.

Gamoran, A., Porter, A. C., Smithson, J., & White, P. A. (1997). Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis, 19*(4), 325–338.

Garcia, S. B., & Guerra, P. L. (2004). Deconstructing deficit thinking: Working with educators to create more equitable learning environments. *Education and Urban Society, 36*(2), 150–168.

Gay, G. (2002). Preparing for culturally responsive teaching. *Journal of Teacher Education, 53*, 110–116.

Grossman, P. L., & Stodolsky, S. S. (1995). Content as context: The role of school subjects in secondary school teaching. *Educational Researcher, 24*(8), 5–23.

Gutiérrez, K. D., & Rogoff, B. (2003). Cultural ways of learning: Individual traits or repertoires of practice. *Educational Researcher, 32*, 19–25.

Hamre, B. K., & Pianta, R. C. (2005). Can instructional and emotional support in the first-grade classroom make a difference for children at risk of school failure? *Child Development, 76*(5), 949–967.

Hand, V. M. (2010). The co-construction of opposition in a low-track mathematics classroom. *American Educational Research Journal, 47*(1), 97–132.

Hanushek, E. A. (1992). The trade-off between child quantity and quality. *The Journal of Political Economy, 100*(1), 84–117.

Harris, D. M. (2012). Varying teacher expectations and standards: Curriculum differentiation in the age of standards-based reform. *Education and Urban Society, 44*(2), 128–150.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Hiebert, J., & Wearne, D. (2003). Instructional task, classroom discourse, and students' learning in second grade. *American Educational Research Journal, 30*, 393–425.

Hill, H. C., Ball, D. L., & Schilling, S. G. (2008). Unpacking pedagogical content knowledge: Conceptualizing and measuring teachers' topic-specific knowledge of students. *Journal for Research in Mathematics Education, 39*(4), 372–400.

Hill, H. C., Umland, K. U., Litke, E., & Kapitula, L. (2012). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education, 118*(4), 489–519.

Hox, J. (2010). *Multilevel analysis: Techniques and applications*. London: Routledge.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. MET project*. Seattle: Bill & Melinda Gates Foundation.

Kelly, Sean. (2004). Are teachers tracked? On what basis and with what consequences. *Social Psychology of Education, 7*, 55–72.

Kesner, J. E. (2000). Teacher characteristics and the quality of child-teacher relationships. *Journal of School Psychology, 38*(2), 133–149.

Kulik, C., & Kulik, J. (1982). Effects of ability grouping on secondary school students: A meta-analysis of evaluation findings. *American Educational Research Journal, 79*, 415–428.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*, 287–298.

Ladson-Billings, G. (1995). But that's just good teaching! The case for culturally relevant pedagogy. *Theory into Practice, 34*, 159–165.

LaPrade, K. (2011). Removing instructional barriers: One track at a time. *Education, 131*(4), 740.

LeChasseur, K., Mayer, A., & Donaldson, M. (under review). The structuring of tracking: Instructional practice of teachers leading low and high track classes.

Lee, V. E., Bryk, A., & Smith, J. B. (1993). The organization of effective secondary schools. *Review of Research in Education, 19*, 171–268.

Lee, V. E., & Smith, J. B. (1999). Social support and achievement for young adolescents in Chicago: The role of school academic press. *American Educational Research Journal, 36*, 907–945.

Lucas, S. R. (1999). *Tracking inequality: Stratification and mobility in American high schools*. NewYork: Teachers College Press.

Lucas, S. R., & Berends, M. (2002). Sociodemographic diversity, correlated achievement, and de facto tracking. *Sociology of Education, 75*, 328–348.

Luke, D. A. (2004). *Multilevel modeling*. Thousand Oaks, CA: Sage.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., & Burchinal, M. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language, and social skills. *Child Development, 79*(3), 732–749.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability*. Santa Monica: RAND Corporation.

Merrienboer, J., & Stoyanov, S. (2008). Learners in a changing learning landscape: Reflection from an instructional design perspective. In J. Visser, M. Visser-Valfrey, D. N. Aspin, & J. D. Chapman (Eds.), *Lifelong learning book series* (Vol. 12, pp. 69–90). Dordrecht, South Holland: Springer.

Mickelson, R. A. (2001). Subverting-Swann: First- and second-generation segregation in the Charlotte-Mecklenburg schools. *American Educational Research Journal, 38*, 215–252.

Mikami, Y., Allen, J. P., Pianta, R. C., & Lun, J. (2011). Effects of a teacher professional development intervention on peer relationships in secondary classrooms. *School Psychology Review, 40*(3), 367–385.

Miller, S. M. (2010). Towards a multimodal literacy pedagogy: Digital video composing as 21st century literacy. In P. Albers & J. Sanders (Eds.), *Literacies, arts, and multimodalities* (pp. 254–281). Urbana-Champaign, IL: National Council of Teachers of English.

Miller, S. M. (2013). A Research metasynthesis on digital video composing in classrooms: an evidence-based framework toward a pedagogy for embodied learning. *Journal Of Literacy Research, 45*(4), 386–430.

Moller, S., & Stearns, E. (2012). Tracking success: High school curricula and labor market outcomes by race and gender. *Urban Education, 47*(6), 1025–1054.

NCTQ. (2013). *State of the States 2013 Connect the Dots: Using evaluations of teacher effectiveness to inform policy and practice*. Washington, DC: National Council on Teacher Quality.

Nieto, S. (1995). *Affirming diversity: The sociopolitical context of multicultural education*. White Plains, NY: Longman.

Nunn, L. M. (2011). Classrooms as racialized spaces: Dynamics of collaboration, tension, and student attitudes in urban and suburban high schools. *Urban Education, 46*(6), 1226–1255.

Oakes, J., Ormseth, T., Bell, R., & Camp, P. (1990). Multiplying inequalities: The effects of race, social class, and tracking on opportunities to learn mathematics and science. Santa Monica: RAND.

Oakes, J. (2005). *Keeping track: How schools structuring inequality*. New Haven, CT: Yale University Press.

Oakes, J., Gamoran, A., & Page, R. N. (1992). Curriculum differentiation: Opportunities, outcomes and meanings. In P. W. Jackson (Ed.), *Handbook of research on curriculum* (pp. 570–608). New York: Macmillan Publishing Company.

Oakes, J., & Guiton, G. (1995). Matchmaking: The dynamics of high school tracking decisions. *American Educational Research Journal, 32*(1), 3–33.

Pace, J. L., & Hemmings, A. (2007). Understanding authority in classrooms: A review of theory, ideology, and research. *Review of Educational Research, 77*(1), 4–27.

Pianta, R. C., & Hamre, B. K. (2005). *Classroom assessment scoring system, secondary manual*. Charlottesville, VA: Teachstone Training.

Ponitz, C. C., Rimm-Kaufman, S. E., Grimm, K. J., & Curby, T. W. (2009). Kindergarten classroom quality, behavioral engagement, and reading achievement. *School Psychology Review, 38*(1), 102–120.

Rice, J. K. (2003). *Teacher quality: Understanding the effectiveness of teacher attributes*. Washington, DC: Economic Policy Institute.

Riehl, C., Pallas, A. M., & Natriello, G. (1999). Rites and wrongs: Institutional explanations for the student course-scheduling process in urban high schools. *American Journal of Education, 107*(2), 116–154.

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica, 73*(2), 417–458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement. *American Economic Review Papers & Proceedings, 94*(2), 247–252.

Roorda, D. L., Koomen, H. M. Y., Spilt, J. L., & Oort, F. J. (2011). The influence of affective teacher–student relationships on students' school engagement and achievement: A meta-analytic approach. *Review of Educational Researcher, 81*(4), 493–529.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*, 175–214.

Rowan, B., Correnti, R., & Miller, R. J. (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.

Rubin, B. C. (2003). Unpacking de-tracking: When progressive pedagogy meets students' social worlds. *American Educational Research Journal, 40*(2), 539–573.

Ryan, R., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*, 68–78.

Sanders, W. L., Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future students academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research and Assessment Center. Retrieved December 6, 2007, from http://www.mccsc.edu/~curriculum/cumulative%20and%20residual%20effects%20of%20teachers.pdf.

Sawchuk, S. (2013). *Teachers' ratings still high despite new measures: Changes to evaluation systems yield only subtle differences*. Education Week, February 6, pp. 1–19.

Slavin, R. (1990). Achievement effects of ability grouping in secondary schools: A best-evidence synthesis. *Review of Educational Research, 60*(3), 471–499.

Steinberg, M., & Donaldson, M. (2015). The new educational accountability: Understanding the landscape of teacher evaluation in the post-NCLB era. *Education Finance and Policy*, 1–40.

Stevenson, D. L., Schiller, K. S., & Schneider, B. (1994). Sequences of opportunities for learning. *Sociology of Education, 67*, 184–198.

Stodolsky, S. S., & Grossman, P. L. (1995). The impact of subject matter on curricular activity: An analysis of five academic subjects. *American Educational Research Journal, 32*, 227–249.

Stuhlman, M., Hamre, B., Downer, J., & Pianta, R. (n.d.). *What should classroom observation measure?* Charlottesville: University of Virgina.

Thapa, A., Cohen, J., Guffney, S., & Higgins-D'Alessandro, A. (2013). A review of school climate research. *Review of Educational Research, 83*(3), 357–385.

Van Houtte, M. (2004). Tracking effects on school achievement: A quantitative explanation in terms of the academic culture of school staff. *American Journal of Education, 110*, 354–388.

Veenman, M. V. J., Kok, R., & Blote, A. W. (2005). The relation between intellectual and metacognitive skills at the onset of metacognitive skill development. *Instructional Science, 33*, 193–211.

Watanabe, M. (2008). Tracking in the era of high stakes state accountability reform: Case studies of classroom instruction in North Carolina. *Teachers College Record, 110*(3), 489–534.

Weinstein, R. S. (1996). High standards in a tracked system of schooling: For which students and with what educational supports. *Educational Researcher, 25*(8), 16–19.

Wells, A. S., & Serna, I. (1996). The politics of culture: Understanding local political resistance to detracking in racially mixed schools. *Harvard Educational Review, 66*(1), 93–119.

Williams, W. M., Bluthe, T., White, N., Li, J., Gardner, H., & Sternberg, R. J. (2002). Practical intelligence for school: Developing metacognitive sources of achievement in adolescence. *Developmental Review, 22*, 162–210.

Woodward, J., & Brown, C. (2006). Meeting the curricular needs of academically low-achieving students in middle grade mathematics. *Journal of Special Education, 40*(3), 151–159.

Worthy, J. (2010). Only the names have been changed: Ability grouping revisited. *Urban review: Issues and ideas in public education, 42*(4), 271–295.