For many years, the bulk of research on mathematics teacher learning has consisted of small-scale qualitative studies. While this work has made significant advances in identifying and explicating instructional practices that have the potential for affording all students equitable access to high-quality classroom learning opportunities, what teacher learning and improvement looks like across large numbers of teachers over time is still insufficiently understood. This report contributes to meeting that need with a longitudinal examination of how changes in more than 200 middle grades mathematics teachers’ instructional practice related to their (a) mathematical knowledge for teaching (MKT) and (b) instructional vision. Results of this multi-level regression analysis suggest that MKT and instructional vision are related to instruction in different ways. On average, MKT scores were positively related to the current year’s quality of instruction, but not growth, while instructional vision scores were positively related to growth in instructional quality. Additionally, the analyses revealed different patterns of change, depending on teachers’ instructional vision and practice at the outset of the study.
Examining Relations Between Mathematics Teachers’ Instructional Vision, Knowledge, and Change in Practice

In 1997, Elizabeth Fennema and Barbara Scott Nelson edited a collection of writings on mathematics teachers’ change. Entitled *Mathematics Teachers in Transition*, the book described theoretical perspectives on the development of instructional practice responding to significant epistemic and pedagogical shifts and new approaches to studying forms of practice and phenomena of change. This burst of new ideas was largely motivated by the first *Standards* published by the National Council of Teachers of Mathematics at the close of the previous decade (NCTM 1989), which radically reconceived the nature of the mathematics that students should learn in school and, consequently, the nature of teachers’ work (Nelson 1997).

But the methodological advances during this time were made primarily within case studies of one or a few teachers. While this approach was invaluable in generating new theory and beginning to establish a research agenda for years to come, it was unable to answer many of the very questions it generated. Several years later, among the shortages identified by Adler et al. (2005) in a review of research on mathematics teacher learning, were a lack of 1) longitudinal studies and 2) large-scale studies – both of which they argued are necessary to understand changes in teaching due to professional learning.

The call for large-scale studies has persisted for nearly two decades. In their *Mathematics Teachers in Transition* chapter, Goldsmith and Schifter (1997) described the “need to supplement the current case-study approach with longitudinal studies of larger cohorts of teachers engaged in changing their practice” (48). Yet, in a follow-up to Adler et al.’s (2005) study, Gellert et al. (2013) concluded that small-scale qualitative
research still predominates. They found that while the number of studies with large numbers of participating mathematics teachers has increased, that 89% of the studies still involve less than 100 teachers (331). Thus, studies of teaching with large numbers of teachers remain quite rare.

As teachers attempt to enact the Common Core State Standards for Mathematics (CCSSM), including the eight Standards for Mathematical Practice, the need to understand how teaching practice changes over time remains. In particular, if researchers could identify factors that predict changes in teaching, then professional developers could construct interventions to better support teaching development. But such learning and improvement is multi-faceted (e.g., knowledge, beliefs, practice, etc.) and variable (Fennema et al. 1996), and, as the abovementioned reviews suggest, what it looks like at scale over time is still insufficiently understood.

In this paper we provide a longitudinal look at more than 200 middle grades mathematics teachers’ instructional practice with respect to some of those multiple facets of teachers’ learning. Specifically, we investigated how changes in instructional practice are related to mathematical knowledge for teaching (Hill et al. 2004) and instructional vision (Author 2014).

Theoretical Perspective

Over time, research in mathematics education has clarified the need for providing certain kinds of learning opportunities to students. These include opportunities to struggle to make sense of important mathematical ideas (Hiebert and Grouws 2007); build on current understanding (Carpenter and Lehrer 1999); mathematize situations (Putnam et al. 1990); communicate their ideas (Sfard 2003); critique the ideas of others (CCSSM);
pursue their own conjectures (Carpenter and Lehrer 1999); and reflect on their experiences and current understanding (Hiebert et al. 1997)—all in ways and contexts that afford opportunities to connect to and build from cultural practices and life experiences (Nasir 2002) and pursue agendas of mathematical empowerment and social change (Gutstein 2003).

More recently, many in the field have converged on specifications of particular instructional practices that afford such opportunities (cf. Ball et al. 2009; Boaler and Staples 2008; Greer et al. 2009; Horn, 2011; Jackson et al. 2013; NCTM, 2014; Stein et al. 2008). Consistent across this work are certain pedagogical commitments that are consistent with dialogic mathematics instruction (Author 2015b). These include regularly inviting students to invent solution strategies for open-ended, cognitively demanding problems for which no solutions have been demonstrated; facilitating discussions in which the classroom community compares and contrasts their ideas and builds shared understandings of mathematical ideas; sharing authority with students in posing and solving problems, and in driving discussion; and, although less consistently defined, ensuring that opportunities are equitable and responsive with respect to students’ interests, identities, and current understandings.

We view such forms of instructional practice as representative of high-quality mathematics instruction, and take the enactment of such practices to be a primary goal of (supporting) teacher learning. However, clarity about desired instructional practices and their purpose does not necessarily help in defining pathways of teaching development toward those goals (Staples 2007). Furthermore, the multi-faceted nature of teaching goals complicates the study of teaching because teachers could productively develop their
practice in a number of different ways. In recent years, a number of strands of research have attempted to investigate changes in teaching practice due to teachers’ beliefs (Philipp 2007); organizational change (Gamoran et al. 2003); and, to a lesser but increasing extent, cultural and racial competence (Milner 2003). The project that collected the data we analyze in this manuscript afforded opportunities to make use of two other strands of research as predictors for changes in teaching: the knowledge base for teaching (Shulman 1986), and the appropriation of pedagogical tools (Grossman et al. 1999). We briefly review each in turn.

**Mathematical Knowledge for Teaching**

Since Shulman’s (1986) argument that the “skills of teaching” comprise only half of the profession’s knowledge base, considerable progress has been made in conceptualizing and assessing forms of knowledge for teaching mathematics. As Depaepe et al.’s (2013) review concluded, some of this work has been conducted from a cognitive perspective, investigating categories of “declarative knowledge” and their correlations to aspects of instruction and student outcomes (Thompson 2015), while others have adopted a situated perspective, investigating “knowledge in action” (Seymour and Lehrer 2006). Both of these approaches have strengths and shortcomings (Depaepe et al 2013). In general, an achievement of the cognitive perspective for teacher education has been confirming the association between PCK and students’ learning outcomes (e.g., Hill et al., 2005; Tirosh et al, 2013)—due in part to the relative ease with which it can be investigated with large numbers of teachers and students.

One of the most prominent efforts in the cognitive vain has been the assessment of teachers’ “mathematical knowledge for teaching” (MKT; Hill et al. 2004). In this
work, researchers have argued that, in addition to common knowledge of content (e.g.,
being able to use a procedure appropriately to produce a correct solution), mathematics
teachers need specialized knowledge of content (e.g., being able to explain where a
procedure comes from and why it works), enabling them to, among other things, interpret
a student’s invented strategy mathematically and determine whether it will work in
general; employ and connect multiple representations of a particular concept; and
develop questions and examples around particular mathematical ideas to support
students’ understanding (Hill and Ball 2004). Hill and colleagues have found teachers’
MKT to be related to both the mathematical quality of their instruction (Hill et al. 2008)
and their students’ achievement (Hill et al. 2005). Similarly, in a one-year study of grade
10 students and teachers in Germany, Baumert et al. (2010) found relations between their
own assessment of teachers’ pedagogical content knowledge and both instructional
practice and student learning gains.

However, to date, all of these studies have been cross-sectional. While half of the
studies reviewed by Depaepe et al. (2013) examined the development of PCK over time,
and a small number of studies have examined the association between PCK and
instruction, no studies were listed examining how measures of PCK relate to the
development of teaching practice over time. Furthermore, examinations of the relation
between teachers’ PCK in mathematics and instruction have typically been limited to the
mathematical quality of teachers’ instruction, regardless of pedagogy. Such an “agnostic”
stance may allow for differences in pedagogical approaches even when teachers have the
same measured level of PCK. These differences in pedagogical approaches are likely
based on different commitments and assumptions about teaching and learning
mathematics (e.g., whether supporting students in communicating their reasoning and participating in rigorous classroom discussions is emphasized; Author 2015b), and may therefore be differentially related to teachers’ content knowledge. Because we are interested in understanding teachers’ development of particular instructional practices—those that afford the types of opportunities listed previously—our goal was to examine the relation of teachers’ MKT to the quality of their mathematics instruction (and not merely the mathematical quality of their instruction).

**Appropriating Pedagogical Tools**

A second body of work on which we have drawn concerns the processes by which teachers learn to effectively enact many of the practices referenced above. Theories of teaching development have been most prevalent in studies of pre- as opposed to in-service teachers’ practice. While there may be aspects of teaching development unique to either domain, we were interested in ideas that might apply to in-service teachers, regardless of the settings in which those ideas originated. Although it will be difficult to test mechanistic explanations of theories built from prior qualitative studies (e.g., Grossman et al. 1999; Thompson et al. 2013; etc.), our empirical examination is guided by the general principles outlined by these researchers.

Taking a sociocultural view of teachers’ learning, Grossman et al. (1999) framed the adoption of particular teaching practices as a matter of appropriating pedagogical tools. These tools, they argued, consist of two types: conceptual and practical. Conceptual tools include the theories, principles, and frameworks that teachers use for thinking about teaching, both in general (e.g., planning, instruction, and assessment) and in reflecting on and critiquing their own practice. Practical tools include the instructional
practices and material resources that teachers employ in interaction with students.

Appropriation, as they defined it, is “the process through which a person adopts the pedagogical tools available for use in particular social environments (e.g., schools, preservice programs) and through this process internalizes ways of thinking endemic to specific cultural practices” (15).

An implication of such a theory of teaching development, particularly from a research standpoint, is that teachers’ appropriations of tools likely differ by type in their “observability.” While appropriations of practical tools may result in (observable) changes in instructional practice, teachers’ appropriations of conceptual tools may be less visible. For example, one kind of conceptual tool may be what Thompson et al. (2013) described as critical pedagogical discourse: “personal theories about ‘what counts’ as productive teaching and learning… threads of internalized dialogue that constitute teachers’ narratives about their current and future teaching selves” (578-579).

Understanding this aspect of teachers’ learning would, therefore, require “externalizing” those narratives (more on this in the methods section).

Regardless of how observable it is, appropriation can take many forms; it is not an “all-or-nothing” phenomenon. Grossman et al. (1999) described “degrees” to which teachers appropriate conceptual and practical pedagogical tools. These degrees range from using a tool in name only or perhaps employing some of its surface features, to articulating the conceptual rationale underlying a tool without yet making effective use of the tool, and, eventually, achieving “mastery” (1) (Herrenkohl and Wertsch 1999).

Although not argued explicitly by Grossman et al. (1999) (and not tested directly in the analysis reported here), we assume that the degree of appropriation of each type of
pedagogical tool can potentially influence that of the other. For example, Thompson et al. (2013) argued that, to the extent that critical pedagogical discourses are well-defined, they “filter” the historical, social, and institutional influences that teachers encounter, and determine the degree to which new “visions of practice” are developed and enacted (578). This suggests that, through an iterative process of enactment and reflection, teachers might leverage the appropriation of one kind of pedagogical tool to appropriate the other to a greater degree. To the extent that teachers can explicate an instructional vision, they are more or less primed to take advantage of pedagogical resources in their environment (e.g., tools or resources for teaching practice provided by new curricula, coaches, etc.). These visions, then, form the basis for how teachers filter resources, the degree to which they appropriate these resources in their practice, and the extent to which they allow their visions to change upon reflection of their classroom experiences.

In summary, limited as it may be due to the difficulties of examining changes in practice with large numbers of teachers (described further below), we seek empirical evidence that it is important to measure and account for teachers’ personal theories. Specifically, we examine whether these personal theories, or visions of high-quality instruction, are associated with the appropriation of practical pedagogical tools aligned with the forms of practice described at the beginning of this section.

**Studying changes in teaching with large numbers of participating teachers**

As the field continues to develop theories of mathematics teaching development, there also continues the need for longitudinal studies of both small (e.g., Franke et al. 2001) and large (e.g., Desimone et al. 2002) numbers of teachers in order to iteratively test and refine those theories. Prior small-scale studies of teaching development suggest
that studies of change with large numbers of teachers will be complex. The study of teaching is infused with both conceptual and methodological challenges (Author 2012), which are magnified when examining teaching longitudinally. Below we highlight a few of the challenges particular to studying changes in teaching.

First is the seeming conundrum of needing changes in practice to be close enough to what an intervention intends in order to be able to (dis)confirm hypotheses, but not so close that you are unable to contribute to broader theories of teacher development. For example, researchers examining the program Cognitively Guided Instruction demonstrated how a focus on increasing teachers’ understanding of and attention to students’ mathematical thinking produced aligned changes in their instruction (Fennema et al. 1996) and how the initial professional development (PD) led to generative growth for some teachers as they came to investigate student thinking as sites for their own learning (Franke et al. 2001). What is clear from that body of work is the power of attending to student thinking; what is less clear is the extent to which the generativity of that practice is a unique property of that particular practice, or whether it’s simply one case of what might be a more general explanation of how teachers “continue to build their repertoire of ambitious practices” (Thompson et al. 2013, 604).

In our view, studying teacher change in large numbers across a number of sites can, to some extent, help to resolve this dilemma. Examining associations with large numbers of teachers is, of course, complicated by certain inevitabilities. For example, PD is often conducted locally, likely with different foci and purposes in mind; fidelity to the design of the PD likely varies between (and even within) local contexts; and the more ambitious the goals, the more likely there will be variability in the extent to which new...
practices are taken up by teachers (Author 2015a). However, we argue that to the extent that local PD initiatives and policy contexts are aligned with respect to underlying rationales and goals for instructional improvement, such complications can actually be leveraged to generate new insights in the middle ground between the two constraints of the conundrum noted above: the variation between different local contexts ensures some distance between interventions and outcomes, while cross-site conceptual and structural commonalities afford opportunities to generate more general explanations for changes in teaching—applicable at least to settings in which policies and PD are similarly oriented.

However, we risk squandering this promise with certain methodological assumptions we make when examining large numbers of teachers and by using linear growth models. First of all, when initial measures of instructional quality are high, whether as a result of teaching expertise among a subset of teachers or by measurement error, there may be limited opportunities for assessing growth for some teachers. Measuring these teachers against others starting lower on the index of instructional quality may mask association between the primary independent variable(s) and measured changes in teaching. We might consider examining sub-groups of teachers that start more or less alike in order to compare and contrast the progression of a measure of teaching for teachers that vary initially on our primary independent variable. This would align with prior small-scale, longitudinal studies that have shown that the ways and rates in which teachers develop new practices are not uniform (Fennema et al. 1996; Franke et al. 2001).

A second methodological concern is that assumptions that all teachers would be expected to grow in a linear fashion may be unfounded. Take, for example, findings from Grossman et al.’s (2000) study of novice literacy teachers’ first three years of teaching,
during which researchers observed a lag between university training and aligned practices, which were not present in the first year of instruction but were observed in the second year. Thus, we should also test for non-linear models in order to better understand if there are some teaching practices for which linear assumptions do and do not hold—one more way in which studying change in large numbers of teachers is beneficial.

In the next section, we describe our methods for leveraging the similarities and differences among a large number of teachers across multiple sites in order to engage in preliminary empirical investigations of teaching growth as it relates to the constructs described above, and the variability with which growth occurs.

Methods

Study Setting

The data that we analyzed were collected in the first four years of the XXXX XXXXX XXXXXXXXX XX XX XXXXXXXX XXXX XX XXXXX (XXXX) project, a longitudinal study of how mathematics instructional reform can be achieved at scale (Co-PI and Co-PI 2011). Members of the project collaborated with the public school districts of four large U.S. cities (A-D), whose leaders had recently begun implementing comprehensive initiatives aimed at ensuring that students in every middle grades mathematics classroom had regular opportunities to collaboratively make sense of and solve challenging mathematical problems and engage in rigorous discussions about their thinking.

The districts’ reforms efforts were initiated independently from XXXX researchers, and therefore varied to some degree in focus and structure. However, central to the support efforts in each district was the provision of conceptual and practical
pedagogical tools aimed at promoting mathematics instruction aligned with what was described previously. Across the different districts, conceptual tools included guiding principles for supporting students’ learning; tools for reflecting on lesson planning (e.g., Thinking Through a Lesson Protocol; Smith et al. 2008); vision statements; and observation protocols. Practical tools included curriculum materials (which, in three districts, were based the second edition of the Connected Mathematics Project series; Lappan et al. 1998); diagnostic assessments; and Accountable Talk® prompts that teachers could use in pressing students to explain their reasoning or in facilitating a discussion. Thus, the districts provided rich opportunities to study teacher learning and change in environments in which a particular vision of instruction was being promoted and supported.

Participants

The analytic models we describe below included data from 248 middle-grades mathematics teachers from the four XXXX districts. Across the schools in those districts (10 in A; 7 in B; 6 in C; and 7 in D), about 74% of all students, on average, were receiving free or reduced price lunch. Teachers’ classroom experience ranged from 1 to 38 years (mean=8.9). About 35 percent of the teachers had a masters degree.

Data Collection and Measures

In the following paragraphs, we describe our data collection and analysis with respect to three measures: instructional quality, mathematics knowledge for teaching, and instructional vision. Over four years, 506 observations (classroom video-recording, pencil-and-paper assessment, and interview) were collected from the teachers. In addition to these measures, we collected demographic information through annual surveys,
including levels of certification; number of advanced mathematics methods and content courses completed; number of bachelor and masters degrees; and years of teaching experience.

**Instructional Quality.** One classroom of each teacher was observed and video-recorded annually on two consecutive days of instruction, mid-school year. Observations were scheduled ahead of time, and teachers were asked to conduct lessons in which students solved a problem and had opportunity for discussing solutions. (Therefore, we do not interpret the observations to reflect *typical* practice per se, but rather teachers’ current “best attempt” at a particular kind of lesson).

XXXX researchers assessed the quality of teachers’ instruction by using an adapted version of the Instructional Quality Assessment (IQA) Mathematics Toolkit (Boston 2012). The IQA’s two primary sections include eight rubrics for assessing the cognitive demand of classroom activity over the course of a lesson (Academic Rigor) and specific aspects of discourse during the whole-class discussion after students have had a chance to work on solving the task (Accountable Talk®). A lesson is assigned a score from 0-4 on each rubric, with 0 indicating an absence of the construct. In general, “a demarcation line exists between score levels 2 and 3 that separates high- and low-level cognitive demands, talk moves, or expectations in each dimension” (Boston 2012, 85-86). Figure 1 provides a representation of the orientation of the instrument’s rubrics and scoring.

Scores from the two Academic Rigor rubrics were averaged to create the Task sub-score. Scores from the other six rubrics, all of which pertain to the concluding whole class discussion, were averaged to create the Discussion sub-score. Finally, the Task and
Discussion sub-scores were averaged to create an overall IQA score, which could range from 0 to 4—though only once was an overall score below 1. In our analyses, for each year, we used teachers’ best of the two lessons’ scores as the outcome variable.

**Mathematical knowledge for teaching.** In March of each year, participating teachers completed the MKT pencil-and-paper assessment developed by the Learning Mathematics for Teaching project (Hill et al. 2004). For each of the two subtests (number concepts and operations; patterns, functions and algebra), raw scores were translated into developer-provided scale scores, which we averaged to form a single MKT score for each participant annually.

**Instructional vision.** As explained in detail elsewhere (Author 2014), teachers’ instructional visions were assessed with the interview-based Vision of High-Quality Mathematics Instruction (VHQMI) instrument in January of each year. The instrument includes a series of interview prompts and rubrics for assessing the sophistication of participants’ articulations of key dimensions of classroom practice: role of the teacher, mathematical tasks, student engagement in classroom activity, and classroom discourse, the last of which includes five sub-dimensions (patterns and structure of talk, the nature of talk, teachers’ questions, students’ explanations, and students’ questions). Figure 2 provides brief characterizations for a sample of the instrument’s dimensions and sub-dimensions in terms of their ranges and scoring.

The VHQMI assessment builds on Hammerness’s (2001) notion of personal vision in that it captures “a set of images of ideal classroom practice for which teachers strive” (p. 143), but differs in that it privileges a particular instructional vision—one aligned with the instructional vision being promoted by leaders of the participating
districts. Especially important in the development of the coding rubrics were the kinds of
degrees of appropriation previously described, which helped distinguish responses
limited to naming labels or surface features from those that also included articulations of
conceptual underpinnings (Grossman et al. 1999).

Along with three sub-dimensions of classroom discourse (student questions,
teacher questions, and student explanation), one dimension that is not represented in
Figure 2 is student engagement in classroom activity. Its rubric has only two levels, one
for capturing an emphasis on students being “on task” with respect to whatever the
teacher assigns (level 1), and another for capturing the use of reform-oriented labels (e.g.,
“hands-on” activity) without articulating an underlying function (level 2). We expanded
these scores to four levels by combining them with nature of talk scores: for any
interview that was scored on the classroom activity rubric, if that interview was also
scored on the nature of talk rubric, the classroom activity score was replaced with the
nature of talk score. This combination is based on an inference that for participants who
said that “student engagement” is something to look for, their description of students’ talk
is an elaboration on the nature of students’ classroom activity.

For our analyses, we averaged scores across all dimensions and sub-dimensions
for which scores were assigned (using the nature-of-talk-expanded version of classroom
activity scores) to create a single VHQMI mean score (ranging from 0 to 4) for each
teacher for each year, which can be viewed as mapping roughly onto Grossman et al.’s
(1999) five degrees of appropriation and as being aligned with the IQA range described
above.

**Research Questions**
Given the novelty of our longitudinal data with a large number of teachers we examined research questions to understand whether we could identify predictors related to teaching practice and, in particular, growth in teaching practice. Our first research question was: *To what extent is the quality of teachers’ instructional practice and the change in teachers’ instructional practice predicted by (a) the sophistication of their visions of high-quality mathematics instruction and (b) their mathematics knowledge for teaching?*

To answer this question, scores from the annual measures described above were entered into a hierarchical linear growth model to examine growth over time, where “growth” (a typical term for analyzing change with statistical models) refers to observed changes in the quality of teachers’ mathematics instruction (or the degree to which teachers have appropriated pedagogical tools). This model answers several basic questions for us. First, we can understand whether the quality of teachers’ instruction is, on average, growing. Second, we generate estimates of the reliability of the average growth estimate (a ratio of true variance to error variance in growth estimates) as well as the variance around that underlying growth. Third, we can estimate whether covariates can predict the variance we observe in teacher growth rates.

However, some of the model’s underlying assumptions may be tenuous. For example, it is likely that different teachers actually appropriate pedagogical tools at different times and at different rates. To the extent that this is true, it could influence the “measurement error” when we assume a linear growth pattern. This would have the undesirable outcome of effectively reducing the reliability of our growth estimates.
Because we did have low reliability for our growth estimates we decided to probe beyond our linear growth model to examine different intervals (pieces) of change in teachers’ practice. We examined these models to understand whether they provided a better fit to the data, but also to explore additional hypotheses about the degree of teachers’ appropriation of new pedagogical tools in relation to their forms of practice and perspective at the outset. Informed by findings from previous (qualitative) research (cf. Fennema et al. 1996; Thompson et al. 2013), we sought evidence to suggest that different teachers’ appropriations of pedagogical tools occurred at different rates and stages over this four-year interval.

Our second research question, then, was: Do teachers’ instructional vision and practice at the outset predict whether they improve their instructional practice at different intervals over the course of a four-year study? Specifically, do piecewise models have a better fit to the data, and to what extent are changes in instructional practice predicted by categories of initial perspective and practice?

To test our hypothesis, we created groups of teachers based on the degrees of their appropriation of conceptual or practical pedagogical tools (as assessed by VHQM1 and IQA, respectively) at the outset, which, we conjectured, may have differentially influenced how they experienced the district reform efforts and the forms of support they required. We illustrate with the graphic in Figure 3, which helps describe the theoretical distinction we are making with our groupings.

For example, if we were to identify “high” and “low” scores for both VHQM1 and IQA we would define four distinct groups. Teachers below a cut-point on both measures fall in the group indicating a low degree of appropriation of both conceptual
(VHQMI) and pedagogical (IQA) tools at baseline (“LoLo” in Figure 3). On the opposite diagonal are teachers whose initial instructional visions and practice are both above the cut-points (“HiHi”). The off-diagonal categories are defined by a greater degree of appropriation of one kind of pedagogical tool but not the other (relative to the sample population). In the upper left group (“LoHi”) are teachers with high IQA scores at baseline, but less sophisticated instructional visions, with the opposite true of teachers in the lower right group (“HiLo”).

In our analyses we sought to understand how different groups of teachers changed over the four years of the study over two different study intervals – the first piece measuring the first year of growth (years 1-2; capturing an immediate response to reform implementation), the second measuring the growth interval for the final two study years (year 2-4; capturing a delayed response to continued reform implementation). Our hypothesis (informed by the prior research referenced above) was that the instructional practice of teachers in each group may have, on average, changed at different rates over different intervals. For example, one group of teachers might be primed to respond immediately to district reform efforts (i.e., “HiHi”), while in a different group we may expect a lag between professional development and other support opportunities and change in practice (i.e., “HiLo”).

To test this, we operationalized the potential differences as represented in Figure 3. Teachers scoring lower than “2” on the IQA, on average, were considered to be low, while teachers scoring “2.5” or higher were considered to be high. Approximately 66% of the teachers were between 2 and 2.5 on the IQA, meaning approximately 33% were either low or high. The cut points for the VHQMI were 1.75 and 2.5, respectively. Again, about
33% of the teachers were located in either extreme. Four of the categories depicted in Figure 3 are those defined above; a fifth category includes teachers whose initial VHQM1 and IQA scores were in the middle for both ("Mid"). In all, 168 teachers in the first two years of the study were placed into these five groups. We used these groups to understand not only whether teachers improved on our measure of instructional practice but, if so, during which intervals we were likely to see those improvements.²

Analyses

We utilized our longitudinal data to explore changes in teaching practice over time. In order to examine the influence of MKT and instructional visions (VHQMI) on variance in teachers’ practice at a particular point in time (i.e., status) and variance in their rates of change (i.e., growth) in practice, we analyzed a series of hierarchal linear models using HLM7.1 (Raudenbush and Bryk 2002). These models nested teachers’ IQA scores over the four study years within teachers and also accounted for the nesting of teachers in schools and districts. We centered our conditional models at different stages of the study to explore the effects of our independent covariates on the variance in teachers’ practice at given points in time (i.e., the beginning and end of the study). The conditional model can be summarized as follows:

\[
\text{IQA}_{tij} = \pi_{0ij} + \pi_{1ij}(\text{year}_{tij}) + \epsilon_{tij} \tag{1.1}
\]

\[
\pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{VHQMI}) + \beta_{02j}(\text{MKT}) + \Sigma \beta_{0qj}(X_q) + \tau_{0ij} \tag{1.2}
\]

\[
\pi_{1ij} = \beta_{10j} + \beta_{11j}(\text{VHQMI}) + \beta_{12j}(\text{MKT}) + \Sigma \beta_{1qj}(X_q) + \tau_{1ij}
\]

\[
\beta_{00j} = \gamma_{000} + \gamma_{001}(\text{DistrictB}) + \gamma_{002}(\text{DistrictC}) + \gamma_{003}(\text{DistrictD}) + \epsilon_{00j}
\]

\[
\beta_{01j} = \gamma_{010}
\]
\beta_{0qj} = \gamma_{0q0}; \text{ for } q = 3 \text{ to } Q \quad \text{[equation 1.3]}
\beta_{10j} = \gamma_{100} + \gamma_{101}(\text{DistrictB}) + \gamma_{102}(\text{DistrictC}) + \gamma_{103}(\text{DistrictD}) + u_{10j}
\beta_{11j} = \gamma_{110}
\beta_{12j} = \gamma_{120}
\beta_{1qj} = \gamma_{1q0}; \text{ for } q = 3 \text{ to } Q

Our primary foci in these models were the relations between (a) teachers’ VHQMI and both the status of (\gamma_{010}) and growth in (\gamma_{110}) teachers’ IQA scores over the four-year interval, and between (b) teachers’ MKT and both the status of (\gamma_{020}) and growth in (\gamma_{120}) teachers’ IQA scores.

To answer our second research question, a second set of analyses sought to relax assumptions about the linear growth models across teachers. In order to understand whether the instructional practice of different teachers appeared to change at different time intervals we ran a piecewise analysis, where the first piece measured change over the entire interval (a base growth rate) and the second piece measured an increment or decrement to change from the second to the fourth year. Using the base-increment model it is possible to combine coefficients to determine point estimates at each study year (Raudenbush and Bryk 2002). The combined pieces allow for each teacher to have two different growth rate estimates, one from the first to second year and another from the second to the fourth year.

For this analysis we retained teachers whose first time point was the first or second year of the study\(^3\) because there could still theoretically be enough time points to measure their change longitudinally. We then centered time in our analysis so that the
intercept was the estimated IQA score for teachers at the end of the study. Thus we were able to reconstruct growth trajectories for each group separately and understand how their growth trajectories differed, but also whether there were significant differences between groups at the end of the study. Our conditional model can be summarized as follows:

$$\text{IQA}_{tij} = \pi_{0ij} + \pi_{1ij}(\text{BaseRate}_{tij}) + \pi_{2ij}(\text{IncrementalRate}_{tij}) + \epsilon_{tij}$$ \hspace{1cm} \text{[equation 2.1]}

$$\pi_{0ij} = \beta_{00j} + \beta_{01j}(\text{LoHi}) + \beta_{02j}(\text{HiLo}) + \beta_{03j}(\text{Mid}) + \beta_{04j}(\text{HiHi}) + \sum_{q=5}^{Q} \beta_{0qj}(X_q) + \epsilon_{0ij}$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j}(\text{LoHi}) + \beta_{12j}(\text{HiLo}) + \beta_{13j}(\text{Mid}) + \beta_{14j}(\text{HiHi}) + \sum_{q=5}^{Q} \beta_{1qj}(X_q) + \epsilon_{1ij}$$ \hspace{1cm} \text{[equation 2.2]}

$$\pi_{2ij} = \beta_{20j} + \beta_{21j}(\text{LoHi}) + \beta_{22j}(\text{HiLo}) + \beta_{23j}(\text{Mid}) + \beta_{24j}(\text{HiHi}) + \sum_{q=5}^{Q} \beta_{2qj}(X_q) + \epsilon_{2ij}$$

$$\beta_{00j} = \gamma_{000} + \gamma_{001}(\text{DistrictB}) + \gamma_{002}(\text{DistrictC}) + \gamma_{003}(\text{DistrictD}) + \epsilon_{00j}$$

$$\beta_{01j} = \gamma_{010}$$

$$\beta_{02j} = \gamma_{020}$$

$$\beta_{03j} = \gamma_{030}$$

$$\beta_{04j} = \gamma_{040}$$

$$\beta_{0qj} = \gamma_{0q0}; \text{ for } q=5\text{ to } Q$$

$$\beta_{10j} = \gamma_{100} + \gamma_{101}(\text{DistrictB}) + \gamma_{102}(\text{DistrictC}) + \gamma_{103}(\text{DistrictD}) + \epsilon_{10j}$$

$$\beta_{11j} = \gamma_{110}$$

$$\beta_{12j} = \gamma_{120}$$

$$\beta_{13j} = \gamma_{130}$$

$$\beta_{14j} = \gamma_{140}$$

$$\beta_{1qj} = \gamma_{1q0}; \text{ for } q=5\text{ to } Q$$

$$\beta_{20j} = \gamma_{200} + \gamma_{201}(\text{DistrictB}) + \gamma_{202}(\text{DistrictC}) + \gamma_{203}(\text{DistrictD}) + \epsilon_{20j}$$

$$\beta_{21j} = \gamma_{210}$$
\[
\beta_{22j} = \gamma_{220} \\
\beta_{23j} = \gamma_{230} \\
\beta_{24j} = \gamma_{240} \\
\beta_{2qj} = \gamma_{2q0}; \text{ for } q=5 \text{ to } Q
\]

Our primary interest in these models was to understand whether instructional practice of different groups of teachers varied and, if so, in what ways. Thus, after adjusting for all of the teacher characteristics and the district in which the teachers worked, we were primarily interested in the coefficients for \(\gamma_{110}\) through \(\gamma_{140}\) to understand the difference in growth rates for teachers categorized as LoLo, LoHi, HiLo, and HiHi compared to the reference group (Mid) over the entire growth interval—the base growth rate. We were also interested in the coefficients for \(\gamma_{210}\) through \(\gamma_{240}\) to understand increments or decrements to the base growth rates for LoLo, LoHi, HiLo, and HiHi teachers compared to the reference group (Mid) during the second period of time (from study year 2 to study year 4).

**Results**

In Table 1 we first report IQA descriptives for all teachers at each time point. The means reveal an overall pattern for teachers over time: what began at 2.20 in the first year dropped to 2.09 in year 2 and then rose over the next two years to end close to where it began at 2.21. Standard deviations imply a healthy distribution of scores around the mean at each time point, with the greatest in year one. The means also show that teachers’ instruction was typically scored closer to a “2” on the IQA than a score of “3.” It is important to note that these kinds of group level statistics do not account for the influence of attrition on these means (e.g., better teachers being recruited into coaching positions
leading to lower IQA means; districts actively recruiting better teachers, leading to higher IQA means). For that reason we sought to understand IQA growth for individual teachers nested in schools, and to try to predict variation in growth using MKT and VHQMI.

Table 2 provides the results of our linear growth analyses in response to our first question. We present the fully unconditional model, as well as the results with a model centered at the beginning of the study and the same model centered at the end of the study. The purpose of this is to demonstrate the difference in the relations between our measured independent variables with status and growth measures of the IQA.

We found that MKT at the outset was related to teachers’ IQA when status is measured at the first timepoint (coeff.=.11, t=2.15) or the second timepoint (coeff.=.083, t=2.09, not displayed in Table 2), but not beyond that. Thus, adjusting for all of the other parameters in the model, MKT was associated with the status of teachers’ IQA scores at the first two timepoints, but their MKT score at time 1 did not predict growth in IQA.

The sophistication of teachers’ instructional visions at the beginning of the study had a different set of associations. In contrast to MKT, teachers’ VHQMI was negatively associated with model estimates of their IQA score at the beginning of the study (coeff.=-.10, t=2.29), but by the end of the study there was a significant positive association with teachers’ VHQMI at outset and their status on the IQA (coeff.=.12, t=2.74). This is because there was also a significant association between teachers’ VHQMI at entry and teachers’ growth on the IQA (coeff.=.08, t=3.46). The magnitude of this effect is surprising, especially when compared to other covariates in the model. A one standard deviation difference in VHQMI initially netted an average gain on the IQA of about .08 sd per year over the course of the study (or about .24 sd total over the three intervals
during the course of the study). No other teacher covariates in the model predicted growth on the IQA over this interval. Thus, a higher degree of appropriation of conceptual pedagogical tools (as measured by the VHQMI) seems to be predictive of future appropriation of practical pedagogical tools (as measured by the IQA).

Regarding our second research question (regarding whether and how the practice of five different groups of teachers changed differently), Figure 4 provides illustrations of the profiles of individuals in two of these groups: HiLo (Sandra) and LoLo (Pamela). The profiles include an excerpt from their Year 1 interview, representing their characterizations of high-quality mathematics instruction, and a progression of scores on two IQA indices: effort to press students for evidence/explanation (“teacher asking”) and the nature of discussion after students have worked on a task (“rigor of discussion”). In the first year of data collection, the teachers’ instruction was rated at the same (“Lo”) level with respect to their press for explanation and facilitation of rigorous discussion. By Year 4, in Pamela, whose instructional vision at the outset was assessed as being less sophisticated (“Lo”), little change was observed. However, observations of Sandra, whose instructional vision in Year 1 was more sophisticated (“Hi”), suggested that she was taking up new forms of practice with respect to pressing for reasoning and facilitating classroom discourse. There was, of course, variability in teachers’ profiles within each group, but these illustrations are intended to provide a concrete sense of what such categories (HiLo, LoLo, etc.) might mean in terms of appropriating pedagogical tools over time.

Table 3 provides a summary of the change in IQA ($\Delta_{IQA}$) for each group of teachers over the two study intervals (initial year growth, followed by the interval for the
last two study years). These results are also displayed graphically in Figure 5.

Interestingly, the results show that for some groups of teachers there appears to have been regression towards the mean. For example, those initially scoring high (above 2.5) or even moderately high (above 2) on the IQA retraced during the first interval. In the second interval, scores of those teachers initially high on the IQA continued to go down. However, there was a large difference between the rate of decline for the group of teachers scoring high initially on the VHQMI versus the group of teachers scoring low initially on the VHQMI ($\Delta_{IQA}=-.53$ vs $\Delta_{IQA}=-.86$, respectively). For those teachers higher on the VHQMI, the decline was far less. Although HiHi teachers declined over the course of the study, they still ended with significantly higher status on the IQA than those of the LoHi and LoLo groups (e.g., see Figure 5, at time 4 where the average estimate for the HiHi group remains significantly higher than two of the other four groups).

The groups below the mid-line in Table 3 are also interesting to consider. The pattern of the group scoring moderately high on both the VHQMI and initial IQA is very similar to the pattern of the overall mean across the sample (see Table 1), an immediate decline in the 1st interval, followed by steady growth over the (longer) second interval. The final two groups began the study with low initial scores on the IQA. While both groups gained over the entire study interval ($\Delta_{IQA}=.73$ for HiLo vs $\Delta_{IQA}=.24$ for LoLo), the gain of the High VHQMI group was much greater. This is especially apparent during the second interval where a gain of .53 points on the IQA was observed for the HiLo group versus a decline of .12 points on the IQA for the LoLo group. Thus, while some of these gains may be the result of low initial estimates and regression to the mean, it is
especially telling that during the latter interval the HiLo group began to accelerate their growth, while the LoLo group declined over the same interval.

Combined, these findings provide insight into how VHQMI is associated with growth on the IQA in our first set of analyses. All teachers with initially high IQA scores declined in those scores over the course of the study, but teachers who articulated more sophisticated instructional visions declined less. Likewise, when comparing teachers scoring low initially on the IQA, teachers with more sophisticated instructional visions grew more. More importantly, this descriptive analysis shows that, in settings in which instructional reform is being promoted, teachers at different initial degrees of appropriation of various pedagogical tools are likely to demonstrate different future patterns of appropriation.

**Discussion**

The field continues to have limited studies of how teaching practice develops over time (Adler et al. 2005; Gellert et al. 2013). Our work is one attempt to both provide new insights into teacher development and to begin to unpack methodological and conceptual issues that will need to be confronted as researchers collect better data and create better analytical tools to explore changes in teaching over time. Below we elaborate on both types of contributions, as well as acknowledge limitations of our study.

**Conceptual and Theoretical Insights**

The data we analyzed afforded opportunities to examine how teachers’ current and future appropriation of practical pedagogical tools (as measured with the IQA) is related to both their disciplinary knowledge for teaching (MKT) and their appropriation
of conceptual pedagogical tools (VHQMI). In the following paragraphs, we discuss our findings with respect to each.

**Mathematical knowledge for teaching.** First, that teachers’ MKT at the outset was predictive of their instructional practice in that year and the next does not only replicate previous findings that declarative knowledge can be associated with aspects of teaching practice (e.g., Baumert et al. 2010; Hill et al. 2008); examining relations between scores on the MKT assessment and scores on instructional measures that privilege a particular pedagogy are not common, and our results suggest that there is some relation between teachers’ pedagogical content knowledge and how effectively they enact practices associated with inquiry-based or dialogic instructional approaches. (Likewise, it would be interesting to examine whether MKT is related to the quality of teaching that adheres more to a model of direct instruction.) But that the relation between initial MKT and annual IQA scores did not extend beyond year 2 is also interesting (in fact, same-year correlations between MKT and IQA decreased each of the four years). This finding seems to suggest that while declarative knowledge has demonstrated cross-sectional relations to measures of aspects of teaching (notably, the IQA used here and the Mathematical Quality of Instruction, or MQI, used in prior studies), as teachers were introduced to new supports for instructional improvements, other factors predicted the rate at which their instruction improved.

This, of course, is related to the finding that there was a lack of association between initial MKT and instructional growth. While it needs to be replicated, this result has several potential implications. It could indicate that although higher levels of mathematical knowledge for teaching may be beneficial in the classroom, such
knowledge may not, by itself, be generative with respect to learning to employ the kind of practical pedagogical tools assessed by the IQA. Or, it may be that the type of mathematical knowledge necessary for instructional growth is not declarative (as is assessed by the MKT), but rather a kind of “knowing-to-act,” such as that studied from more situated approaches (Depaepe et al. 2013, p. 23), or “usable knowledge” (Kersting et al., 2015) or “mathematical meanings for teaching” (Thompson, 2015). In either case, as the developers of the MKT assessment have acknowledged, teachers’ knowledge (or knowing) likely interacts with other resources in their environments (Charalambos and Hill 2012): new pedagogical tools, whether conceptual (e.g., frameworks introduced in PD) or practical (e.g., new curriculum materials) “can to some degree scaffold low-MKT teachers successfully into [new] practices” (Hill et al. 2012, p. 506).

**Instructional vision.** Our finding that teachers’ visions of high-quality mathematics instruction at the outset are predictive of their growth in instructional quality both validates the notion of “vision” as a construct and, when coupled with our analysis of different patterns of change, provides insight into relations between different forms of pedagogical tools as they relate to teachers’ learning. First, the relation between instructional vision and growth in instructional practice suggests that higher degrees of appropriation of conceptual tools increase the likelihood that the appropriation of practical tools (of a similar pedagogical orientation) will follow suit. This lends credence to Thompson et al.’s (2013) notion of teachers’ critical pedagogical discourses acting as “filters” to determine the degree to which new perspectives or practices are taken up.

This is further corroborated by the results of our descriptive analysis related to our second research question regarding different patterns of change. For example, of the
teachers whose IQA scores were initially in the bottom two quartiles, those who articulated more sophisticated instructional visions at the outset demonstrated greater growth in instructional quality over the course of the study. One interpretation is that, for the HiLo (high vision, low IQA) group, who had appropriated particular conceptual tools (e.g., guiding principles for their role in the classroom, frameworks for defining high-quality mathematical tasks) to a higher degree, the practical tools that were introduced by the districts’ reform initiatives (e.g., curriculum materials, talk moves) were sufficiently congruent with their instructional visions. However, their rate of growth was the greatest in the second interval, suggesting a certain “lag” between their introduction to new practices and their taking them up (Desimone et al. 2002; Grossman et al. 2000).

One potential contributing factor to such a lag is that the appropriation of new practical tools is not simply a matter of “fleshing out” a previously sparse repertoire; the new practices that are introduced must interact with and, in some cases, supplant old practices—which may be deeply entrenched ways of interacting with students. While teachers increase the degree to which they appropriate new tools, they likely must decrease the degree to which they make use of others. In other words, teachers “in transition” are not only transitioning into new forms of practice, but also out of old.

Also contributing to the perceived lag may be a methodological shortcoming. To the extent that instructional visions act not only as “filters” but also as reflective tools, the iterative process of teachers assessing and refining enactments of new practical tools against an instructional vision may entail “smaller” changes in practice than what the IQA was designed to assess. We return to this idea in the next section, a discussion of methodological challenges.
Methodological Challenges and Limitations

While researchers continue to develop measures of teaching, the psychometric properties have, to date, supported conclusions about inter-rater reliability of the measures (and to a lesser extent the validity of the inferences drawn from the measures) at a single point in time (cf. Shavelson et al. 1986; Author 2012; Bell et al. 2012). Because of the dearth of longitudinal studies, less attention has been paid to: (a) specifying nuanced models of teachers’ learning and growth in practice, and (b) metrics for measuring those aspects that would be sensitive to small changes in practice.

In our case, the instructional measure was the IQA, an instrument with a number of strengths: alignment with both findings from research in mathematics classrooms and the professional development initiatives occurring within the XXXX project’s partnering districts; previous demonstrations of its predictive validity on student learning (Matsumura et al. 2008); and, as shown in our study, acceptable reliability. However, the IQA posed challenges for reliably estimating growth in teaching. Items on the IQA are rated on a 5 point scale (0-4, with 0 representing the absence of a construct) and the distribution of scores for the majority of time points is either a “2” or “3”, with “2”s being modal practice. Thus, the instrument may not sufficiently capture teachers’ attempts at “next-level” instructional practices (Thompson et al. 2013), which could simply be the difference between a “low 2” (bordering on “1”) and a “high 2” (bordering on “3”). Similarly, there is a fairly large jump between a score of “2” and a score of “3.” The key for researchers interested in measuring growth in teaching is to identify signals of practice that demonstrate growth towards a “3” in order to appropriately characterize practice on its way to making such a “leap in practice” (which, of course, may be a leap
only from the researcher’s perspective). This would potentially increase the reliability of our growth estimates.

Second, more work needs to be conducted to understand the number of IQA observations needed to capture reliable (e.g., Wilhelm and Kim 2015; Matsumura et al. 2008) and valid estimates of, in this case, the effectiveness with which teachers can enact certain instructional practices on an annual basis. This is especially true given the discussion thus far suggesting the need to expand the goal from simply demonstrating predictive validity at a single point in time to accurately measuring growth in teaching over time.

One limitation of our work is simply the reliability of our growth estimates in our analyses answering research question 1. Moreover, in answering the second question, we also demonstrated that different patterns of change were observed in different groups of teachers. Thus, we find the underlying assumptions that all teachers begin growing at the onset of new professional learning opportunities and that they grow at a constant rate to be problematic. Figure 5 demonstrates descriptively the problem with these assumptions, the implications of which extend to both those studying and those supporting teacher learning. In our view, researchers would be well served to clarify which groups they expect to grow and at what rate it should be expected, which may require taking account—with some theoretical basis—of teachers’ perspectives and practices at the outset of a study. And, as we discuss below, we think these analyses also provide some foundation for thinking about how initial teacher profiles could influence decisions about how an intervention is structured and focused to effectively support teachers’ continued development.
Conclusion

We wish to conclude by suggesting that the findings in this report also have implications for how we conceive of and provide support for improving teaching. In particular, our analysis of different patterns of change suggests that professional development support could be informed by considerations of the degrees to which teachers have appropriated certain types of pedagogical tools. For example, “vision-forward” (Thompson et al. 2013) teachers (the HiLo group in Figure 3) might benefit from an emphasis on refining their enactment of particular practices, whereas “practice-forward” teachers (the LoHi group) could be supported in solidifying conceptual underpinnings of practices through reflective investigations of their own (and others’) practice.

Such vision- and practice-forward trajectories map roughly onto the distinction between “pedagogies of reflection and investigation” and “pedagogies of enactment,” respectively, in recent conversations about supporting pre-service teacher learning (Ball & Forzani, 2009; Grossman & McDonald, 2008; Lampert, 2010). The emphasis in this literature has been primarily on the latter, with the argument that preservice teachers have few if any experiences on which to reflect, and that we should therefore focus more on training them to enact a manageable number of “high-leverage” practices. In-service teachers, however, do have experiences on which they can reflect and build, and may therefore benefit by support in increasing the sophistication of their reflective (or conceptual) tools.

We do not mean to dichotomize forms of professional support, or oversimplify the complexities of teachers’ ongoing learning. But, it is unlikely that the provision of
practical tools (e.g., new curricular programs) without conceptual tools (Grossman et al. 1999)—or, without regard to the nature of teachers’ critical pedagogical discourses or instructional visions (Thompson et al. 2013)—will lead to lasting improvement. Where failure of reform initiatives is often attributed to insufficient professional development support (e.g., failing to provide teachers with sufficient support in using a new curriculum), our analyses suggest that a focus on amount or even quality of support may not fully explain why initiatives do not take hold. Viewing professional development supports as forms of tools that teachers, to some degree, appropriate (or not) orients us to examining the congruence between new tools and current instructional vision and practice.

It is our hope that future research can support such endeavors. We think it will be important to continue to conduct large-scale, longitudinal work to document changes in teaching and to narrow in on specific populations ripe for demonstrating growth in predictable ways. As noted above, this will require more nuanced measures of change, which, perhaps ironically, will likely also require close qualitative work with a particular focus. As argued by Grossman et al. (1999), achieving appropriation and mastery may take years and, therefore, “argues for a longitudinal look at teachers' development, since they may only be able to master some of the pedagogical tools after several years of classroom practice” (18).
References


Depaepe, Fien, Lieven Verschaffel, and Geert Kelchtermans. 2013. “Pedagogical Content Knowledge: A Systematic Review of the Way in Which the Concept has...


### IQA descriptives

<table>
<thead>
<tr>
<th>Year</th>
<th>Mean</th>
<th>N</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.20</td>
<td>116</td>
<td>0.59</td>
</tr>
<tr>
<td>2</td>
<td>2.09</td>
<td>120</td>
<td>0.53</td>
</tr>
<tr>
<td>3</td>
<td>2.12</td>
<td>121</td>
<td>0.54</td>
</tr>
<tr>
<td>4</td>
<td>2.21</td>
<td>114</td>
<td>0.52</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2.16</strong></td>
<td><strong>471</strong></td>
<td><strong>0.55</strong></td>
</tr>
</tbody>
</table>
Table 2

Effects of MKT and VHQMI on IQA Status and Growth

<table>
<thead>
<tr>
<th></th>
<th>Coefficient (se)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Null Model</td>
<td>Full Model Centered</td>
<td>Full Model Centered</td>
</tr>
<tr>
<td></td>
<td></td>
<td>At Start</td>
<td>At End</td>
</tr>
<tr>
<td>IQA Status (Year 2) – School Fixed Effects&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (γ₀₀₀)</td>
<td>2.16 (.05)</td>
<td>2.07 (.05)</td>
<td>2.16 (.04)</td>
</tr>
<tr>
<td>IQA Status (Year 2) – Teacher Fixed Effects&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VHQMI (γ₀₁₀)</td>
<td>-.10** (.04)</td>
<td>.12 ** (.05)</td>
<td></td>
</tr>
<tr>
<td>MKT (γ₀₂₀)</td>
<td>.11 * (.05)</td>
<td>.03 (.05)</td>
<td></td>
</tr>
<tr>
<td>IQA Growth (Linear) – School Fixed Effects&lt;sup&gt;a&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Intercept (γ₁₀₀)</td>
<td>.01 (.02)</td>
<td>.03 (.03)</td>
<td>.03 (.03)</td>
</tr>
<tr>
<td>IQA Growth (Linear) – Teacher Fixed Effects&lt;sup&gt;b&lt;/sup&gt;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VHQMI (γ₁₁₀)</td>
<td>.08*** (.02)</td>
<td>.08 *** (.02)</td>
<td></td>
</tr>
<tr>
<td>MKT (γ₁₂₀)</td>
<td>-.03 (.02)</td>
<td>-.03 (.02)</td>
<td></td>
</tr>
</tbody>
</table>

* p<.05, **p<.01

<sup>a</sup> Adjusting for district membership, teachers in Districts C and D had significantly lower IQA status (Year 2) compared to teachers in District A; teachers in Districts C and D had significantly faster growth on IQA compared to teachers in District A.

<sup>b</sup> Adjusting for non-significant teacher controls: Certification (full-time, part-time, elementary, middle school), Number of Advanced Mathematics Methods and Content Courses, Number of Bachelor Degrees (0-2), Number of Masters Degrees (0-2), Years Experience, Study Entry Year.
Table 3
*Piecewise Model Results for different Teacher Groupings on Growth*

<table>
<thead>
<tr>
<th>Group</th>
<th>ΔIQ A 1st Interval</th>
<th>ΔIQ A 2nd Interval</th>
<th>Growth Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High VHQMI + High IQA&lt;sub&gt;base&lt;/sub&gt; (n=23)</td>
<td>-.20</td>
<td>-.16</td>
<td>-.36</td>
</tr>
<tr>
<td>Low VHQMI + High IQA&lt;sub&gt;base&lt;/sub&gt; (n=43)</td>
<td>-.36</td>
<td>-.39</td>
<td>-.75</td>
</tr>
<tr>
<td>Mid VHQMI + Mid IQA&lt;sub&gt;base&lt;/sub&gt; (n=56)</td>
<td>-.25</td>
<td>.17</td>
<td>-.08</td>
</tr>
<tr>
<td>High VHQMI + Low IQA&lt;sub&gt;base&lt;/sub&gt; (n=28)</td>
<td>.11</td>
<td>.43</td>
<td>.54</td>
</tr>
<tr>
<td>Low VHQMI + Low IQA&lt;sub&gt;base&lt;/sub&gt; (n=37)</td>
<td>.20</td>
<td>.13</td>
<td>.33</td>
</tr>
</tbody>
</table>

*The first interval is from the 1st to the 2nd year of teachers’ participation.*

*The second interval is from the 2nd (up) to the 4th year of teachers’ participation.*
<table>
<thead>
<tr>
<th>IQA Rubric</th>
<th>Score Range</th>
<th>Score Range</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>TASK SUB-SCORE</strong></td>
<td>Absent (0)</td>
<td>Low (1-2)</td>
</tr>
<tr>
<td>• Task Potential</td>
<td>Students did not work on a mathematical task.</td>
<td>Emphasis on memorization and procedures without conceptual connections.</td>
</tr>
<tr>
<td>• Task Implementation</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>DISCUSSION SUB-SCORE</strong></td>
<td>No whole-class discussion occurred.</td>
<td>Weak and/or very infrequent attempts in teacher’s and students’ talk to press for reasoning, provide explanations and justifications for claims, or link ideas, likely with few participants.</td>
</tr>
<tr>
<td>• Participation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Teacher Linking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Student Linking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Teaching Asking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Student Providing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>• Rigor of Discussion</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Figure 1. Summary of IQA rubrics and scoring.*
<table>
<thead>
<tr>
<th>Level</th>
<th>Role of the Teacher</th>
<th>Classroom Discourse (2 of 5 sub-dimensions)</th>
<th>Nature of talk</th>
<th>Mathematical Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Proactive role in providing / fostering learning opportunities; shared authority</td>
<td>Whole-class, student-initiated conversations</td>
<td>Conceptually oriented talk, including conjectures and arguments—for the purpose of supporting students in ‘doing mathematics’ and/or spawning new investigations.</td>
<td>Genuine inquiry and/or connecting procedures to conceptual ideas, which leads to rich discussion of connections.</td>
</tr>
<tr>
<td>3</td>
<td>Students explore / discover at least part of lesson on own with little interference from teacher</td>
<td>Whole-class conversations of ideas with teacher at the center of talk</td>
<td>Calculational orientation or emphasizes questions / explanations without describing their nature</td>
<td>Problem solving through multiple solution paths, but reasons are not tied to rich discussion or making connections between ideas, or tasks described lack complexity</td>
</tr>
<tr>
<td>2</td>
<td>Students involved in mediating / replicating what the teacher said / did</td>
<td>Student-student in small group/partner work</td>
<td>Insists that talk be about math (and accurate), but does not specify questions, explanations, etc.</td>
<td>“Reform”-oriented aspects (e.g., “hands-on,” “higher order”) without elaborating on the nature of learning opportunities</td>
</tr>
<tr>
<td>1</td>
<td>Traditional role of explaining / demonstrating</td>
<td>Teacher-to-student talk, IRE (Mehan 1979)</td>
<td>n/a</td>
<td>Skill-oriented tasks, leading to application</td>
</tr>
<tr>
<td>0</td>
<td>Teacher is motivating / is engaging / relates to students</td>
<td>n/a</td>
<td>n/a</td>
<td>Not varying in quality/not manipulable</td>
</tr>
</tbody>
</table>

*Figure 2. Sample of VHQM dimensions and ranges.*
Figure 3. Five categories based on initial VHQMI and IQA scores.
<table>
<thead>
<tr>
<th>Sandra’s Year 1 VHQMII Interview excerpt (score: 3.25)</th>
<th>IQA sub-score</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asking questions that pertain to [the students’] lives around them or connected, yeah, to other things in the building, other things that they’ve done in previous days that, you know, showed some sort of cohesiveness to the, to even the beginning of their lesson and that it would, you know, require the kids to learn a concept not just by being told what it is and how to do it but to actually think about what it is they were doing and then coming up with the why or &quot;Oh look this looked for all these problems, so is this gonna work for all of our problems?&quot; yes, no, maybe, you know, and do some critical thinking about it in some way having the students maybe be the leaders more than the teacher and have the teacher just kind of do the prompting to have the class just kind of flow on it's own is, would be really neat to see.</td>
<td>Teacher asking</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rigor of discussion</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pamela’s Year 1 VHQMII Interview excerpt (score: 1.33)</th>
<th>IQA sub-score</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>[Y]ou look for student engagement. You look to see if the students, if there's clear expectations, if students know what expected of them and are they following the model...when teachers tell them what they want and they model for them how they want them to do it and then we look to see if students are able to follow the model.</td>
<td>Teacher asking</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Rigor of discussion</td>
<td>1</td>
</tr>
</tbody>
</table>

Figure 4. Year 1 VHQMII excerpts and 4-year progression of IQA scores for Sandra, a “HiLo” District A teacher, and Pamela, a “LoLo” District C teacher.

Teacher asking (effort to press students for evidence/explanation): 1—no effort; 2—superficial efforts, or at most one strong effort; 3—at least two strong efforts during the lesson.

Rigor of discussion (nature of discussion after students have worked on a task): 1—students provide only brief or one-word answers; 2—presentation of ideas but no discussion, or discussion of only one strategy; 3—discussion of task’s important ideas and the validity of one strategy but without explicit connection, or multiple strategies are discussed but without sufficient conceptual rationales.
Figure 5. Change in IQA scores for five groups of teachers based on initial VHQM1 and IQA scores
Notes

1 Although Herrenkohl and Wertsch wrote about appropriation more in terms of a disposition to take up and use a tool, and mastery as the “cognitive skills involved in using cultural tools” (418) (and that, in fact, individuals can master but elect not to appropriate a tool), what is important for our purposes is a distinction between coming to understand a pedagogical tool conceptually, and making the tool one’s own and putting it to effective use.

2 An analysis with only year 1 teachers (n=118 teachers; 313 observations) yielded similar results. These findings show that our model estimates were not duly influenced by the sparseness of our data. Indeed, the consistency of our estimates confirms what prior research has indicated about the robustness of parameter estimates to sparse clustered data within hierarchical linear models (e.g., Bell, Ferron, Kromrey, 2008; Clarke, 2008; Raudenbush and Bryk, 2002). Reluctant to throw away data, we present the findings with the larger group of teachers only.

3 We reset the time variable for those beginning in their second year. Thus, for every teacher in this analysis their first time point is coded (-3), second (-2), etc. until the fourth time point (0) – indicating that our intercept is measuring difference between teachers at the final time point. These analyses include year 1 and year 2 teachers (n=168 teachers; 411 observations).

4 Teachers’ names are pseudonyms.

5 These change scores were derived by first calculating the point estimates for each time point from the model estimates, reconstructing a group score at each time point for each
group, and then subtracting the point estimates at the beginning of the interval from those at the end of the interval to get a measure of change in IQA score for that interval.