“Will this be on the test?”: Returns to discussion-based mathematics instruction in pre-Common Core assessments

Brooks A. Rosenquist (corresponding author)  
Graduate Research Assistant/Ph.D. Student  
Department of Leadership, Policy, and Organizations  
Peabody College, Vanderbilt University  
brooks.rosenquist@vanderbilt.edu  
PMB 414, 230 Appleton Place  
Nashville, TN 37203 brooks.rosenquist@vanderbilt.edu  
Mobile: (209) 327-5794

Dr. Anne Garrison Wilhelm  
Department of Teaching & Learning  
Simmons School of Education and Human Development,  
Southern Methodist University  
3101 University Blvd  
Suite 345, Box 455  
Dallas, TX 75205  
(214) 768-2347  
awilhelm@smu.edu

Dr. Thomas M. Smith  
Dean and Professor  
Graduate School of Education  
University of California, Riverside  
900 University Avenue  
1207 Sproul Hall  
Riverside, CA  92521  
Ph:  (951) 827-5802  
Fax:  (951) 827-3942  
thomas.smith@ucr.edu

Acknowledgements: The analysis reported in this manuscript was supported by the National Science Foundation under grants ESI-0554535, DRL-0830029, and DRL-1119122. We would like to thanks Paul Cobb, Chris Curran, Charlotte Dunlap, Seth Hunter, Emily Kern, & Morgan Polikoff for their helpful comments.
**Purposes**

The adoption of new, uniform, and more rigorous assessments aligned with new, uniform, and more rigorous standards is part of the policy logic behind the Common Core State Standards (CCSS) initiative. However, the adoption and implementation of the CCSS has outpaced the adoption and implementation of CCSS-aligned assessments; the use of less rigorous, misaligned, or legacy tests may be problematic for measuring teaching and learning which place greater emphasis on higher order thinking skills. The current policy environment of high expectations for rigorous teaching and implementation of potentially less rigorous exams lead us to ask the following research questions: *To what extent does the Instructional Quality Assessment (IQA) – a collection of content-specific rubrics for a high cognitive demand and discussion-based approach to math teaching – predict teacher value-added scores derived from pre-CCSS state tests? Do some aspects of classroom teaching measured by the IQA tend to be more related to these teacher value-added scores than other aspects?*

**Framework**

The standards-based movement in education reflects a number of policy trends with their origins in state and federal initiatives from the 1980s and 1990s. Most standards-based reforms include explicit academic expectations for students’ learning; systemic alignment of teacher training, curriculum, and assessment around these expectations; as well as some degree of accountability rewarding or sanctioning students, teachers, and/or schools on the basis of their performance relative to these explicit expectations (Hamilton, Stecher, & Yuan, 2008). Standards-based reform policies were first implemented at the state level, motivated to a large extent by the theory that making explicit the intended curriculum would equalize students’
opportunities to learn (OTL) within the state (Elmore & Fuhrman, 1995). While the No Child Left Behind (NCLB) Act of 2001 applied standards-based accountability logic to the nation as a whole, individual states were still allowed to dictate and define their own curricular standards. In subsequent years, states were found to vary widely in both the specificity and rigor of their standards (Finn, Julian, & Petrilli, 2006) as well as the level of knowledge students which were required to demonstrate in end-of-the-year tests to qualify as “proficient” (National Center for Education Statistics, 2007). Some analyses suggest that the accountability mechanisms put in place by NCLB provided perverse incentives for states to maintain low standards (Balfanz, Legters, West, & Weber, 2007; Koretz, 2008). The adoption of the Common Core State Standards (CCSS; National Governors Association, 2010) represents an evolution and enlargement of standards-based education policy in the U.S. Just as explicit statewide standards could assist in equalizing opportunities to learn within state, adoption of consistent nationwide standards has the potential to reduce variability in OTL across states. Furthermore, adoption of the CCSS has been interpreted as an effort not only to set the same bar for all states, but also to raise the bar, given that the CCSS are more rigorous and call for more higher order thinking than most of the previous state standards (Porter, McMaken, Hwang, & Yang, 2011).

The alignment of standards, curricula, instruction, and assessment is described as being key to the theory of action behind standards-based education reform (Polikoff, 2012; Roach, Niebling, & Kurz, 2008). The U.S. education system is complex, with policies and activities enacted at different levels: standards and accountability assessments are adopted at the state level, textbooks typically produced by private vendors but adopted at the district- or school level, and instruction is enacted at the classroom level (Martone & Sireci, 2009). The alignment of these elements establishes a consistent and coherent message regarding systemwide goals for
teaching and learning (Martone & Sireci, 2009; Webb, 1997) while providing transparency and allowing for other stakeholder (such as parents and taxpayers) to know what is being taught and tested in schools (Porter, 2002). Yearly results from testing provide a means for measuring systemic progress and guiding the improvement of learning, consistent with an evidence-based and scientific approach to research and development which employs iterative cycles of experimentation, data analysis, and modification (Baker, 2005). The assumption is that teachers and others involved in education policy implementation will respond to data, incentives, and sanctions to align systemic activities and structures in ways that effectively realize the goals and outcomes articulated by the system (Polikoff, 2012; Roach et al., 2008).

However, without sufficient alignment of curriculum, instruction, and assessment, standards-based policies do not function according to this design. Lack of alignment makes it difficult to make reliable and valid inferences about the education system from the data (Baker, 2005; Kurz, Elliot, Wehby, & Smithson, 2010); assessment results will be misleading, and instruction may be ineffective or its effect underestimated (Anderson, 2002; Baker, 2005; Pellegrino, 2009). Without adequate systemic alignment, the logic of accountability breaks down most dramatically at the teacher level. While there are a number of other factors within and outside the school which contribute to learning outcomes, statistical analysis confirms that teachers matter for learning, given that students with different teachers demonstrate different average gains in learning which are both statistically and practically significant (Goldhaber & Hansen, 2010; Nye, Konstantopoulos, & Hedges, 2004; Rivkin, Hanushek, & Kain, 2005). In general, policy schemes which incorporate elements of individual accountability are predicated on an assumption that individuals are responsible for both their actions and the outcomes
which are a predictable result of these actions. More precisely, individual responsibility exists when individuals “possess sufficient operational knowledge to predict the consequences of these various courses of action; when they possess the material, human, and other resources necessary to accomplish the task; and when acting directly influences the result” (Adams & Kirst, 1999, p. 478). Following this argument, if actors within these accountability systems cannot convincingly articulate what teachers need to do to raise test scores in the classroom, or if the form and content of classroom instruction does not seem to be clearly aligned with and linked to test-based measures of student learning, then the mechanisms and theoretical basis for test-based teacher accountability are called into question.

Given the importance of systemic alignment, full realization of the benefits of the CCSS policy initiative requires the adoption of new, more rigorous common tests aligned to these new, more rigorous common standards. In practice, the adoption of the CCSS has outpaced adoption of CCSS-aligned assessments. Some critics suggest that the higher order mathematical thinking skills which the CCSS in mathematics emphasize – such as problem solving, explanation, and justification – are best measured by human-scored open-response questions, which are time-intensive and costly (Bennet & Ward, 1993; Toch, 2006). Concerns regarding the time students would spend taking these tests and the costs of scoring caused designers of the CCSS-aligned assessments to scale back their original and more ambitious designs (Gewertz, 2012). Furthermore, a number of states have delayed plans to implement the new CCSS-aligned assessments and/or withdrawn from the consortia responsible for the design of these new tests. As of June 2014, 46 states had adopted the CCSS or have maintained their commitments to doing so, but only 27 of these states were committed to using one of the CCSS-aligned assessments developed by either of the two national consortia (Gewertz, 2014). A number of
states (e.g. New York and Tennessee) are nominally committed to implementing a consortia assessment but have delayed implementation beyond the 2014-15 school year target. Other states (e.g. Iowa, Kansas, and Kentucky) have chosen to use either existing tests or tests which are developed outside the consortia but are ostensibly aligned to the CCSS. While the rigor of both consortia and non-consortia CCSS-aligned tests has not yet been independently analyzed and confirmed, studies of NCLB era tests describe these assessment as overwhelmingly measuring students’ recall of facts and performance of basic skills (Darling-Hammond & Adamson, 2010; Yuan & Le, 2012).

Furthermore, the test-based measures of teacher effectiveness derived from assessments most recently in use are at best only weakly linked to state-of-the-field measures of what teachers and students do in the classroom, suggesting rather poor alignment. Some of the most recent and comprehensive evidence comes from data gathered by the Measures of Effective Teaching (MET) Project, which collected student test scores and video recordings of classroom lessons from more than 3,000 teachers across the U.S. in order to identify and better operationalize effective teaching (Kane & Staiger, 2012). In order to describe how knowledge and skills were taught and learned in these classrooms, the MET Project utilized five classroom observational instruments selected for their reliability in describing the nature of classroom instruction and interactions. In subsequent analysis, scores from these observation rubrics correlated only weakly with students’ covariate-adjusted gains on state assessments (Kane & Staiger, 2012). Furthermore, the strength of these correlations varied across state tests, with teacher value-added estimates from some of these tests showing no statistically significant correlation with teachers’ scores from classroom observation instruments (Polikoff, 2014). Data was also collected from a subsample of MET Project teachers to determine what knowledge and
skills were taught by asking teachers to report the depth and breadth of content covered during the school year (Polikoff & Porter, 2014). Analysis of these data found that, across the sample, measures of alignment of instruction to state standards at the teacher level was not strongly associated with students learning gains.

Taken together, these analyses suggest that the scores on these state tests are, as a whole, relatively insensitive to differences in the form and content of instruction in any given year. A number of scholars have drawn attention to this troubling phenomenon of instructional insensitivity of widely used tests (D’Agostino, Welsh, & Corson, 2007; Popham, 2007; Polikoff, 2010, 2014; Ruiz-Primo, Shavelson, Hamilton, & Klein, 2002). This issue becomes especially important with the implementation of the CCSS, which requires retaining many standards-based structures and policies while adopting more ambitious goals for teaching and learning and realigning curricula, instruction, and assessment, accordingly.

The need for systemic realignment to more ambitious goals highlights the importance of research that focuses on value-added measures of teacher effectiveness derived from tests or parts of tests which purport to assess higher order thinking skills. This literature suggests that estimates of teacher effectiveness can differ substantially depending on whether tests used to derive these measures are designed to measure more complex learning or to focus instead on more procedural skills (Grossman, Cohen, Ronfeldt, & Brown, 2014; Lockwood et al, 2007; Papay, 2011). For example, the Stanford Achievement Test (SAT) for mathematics generates separate scores for mathematical procedures and for problem solving; multiple analyses of these data suggest that estimates of teacher effectiveness derived from student scores on the procedures section are only moderately correlated with estimates of teacher effectiveness derived from the problem solving section (Lockwood et al, 2007; Papay, 2011). Utilizing MET Project
data, Grossman and colleagues (2014) found that value-added scores from more rigorous, open-response tests were more highly correlated with scores on the PLATO classroom observations instrument than were value-added scores from state assessments. Furthermore, they found that this stronger relationship was driven by the subscale of the observation instrument which measured students’ participation in more intellectually challenging activities and discussion. Grossman and colleagues conclude that, in general, assessments likely vary in the degree to which they are sensitive to and reward more ambitious and cognitively demanding forms of teaching and learning.

**Rigorous, discussion-based instruction in mathematics**

This move towards a more ambitious and cognitively demanding form of teaching and learning is not a new phenomenon in mathematics education. Over the last 25 years, educators, researchers, and policy makers have proposed placing a greater emphasis on developing conceptual understanding of mathematic principles and cultivating higher order thinking skills such as communication, justification, critical thinking, and problem solving. Standards published by the National Council of Teachers of Mathematics (NCTM, 1989, 2000), along with the more recent CCSS, reflect a limited consensus within the research and policy communities for comprehensive reform of what students should know and be able to do at each grade level. For example, the CCSS in mathematics (CCSSM) calls for curriculum and instruction which provides opportunities for students to learn to “construct viable arguments and critique the reasoning of others” and to “justify their conclusions, communicate them to others, and respond to the arguments of others” (p.6).

Acknowledging that the CCSS only specify goals for student learning and do not dictate an instructional approach, NCTM published an update of its guidelines for instruction in
mathematics, with the view that these principles should provide “direction and fill the gap
between the adoption of CCSSM and the enactment of policies and programs required for its
widespread and successful implementation” (Leinwand, Hunker, & Brahier, 2014, p.517). Two
key characteristics of high quality mathematics instruction outlined in these guidelines are (1) the
implementation of tasks that promote reasoning and problem solving, and (2) teacher facilitation
of mathematical discourse among students to analyze and compare problem-solving approaches
and build a shared understanding of mathematical concepts (NCTM, 2014). A common inquiry-
and discussion-based approach to math lessons which incorporate these elements follows a three-
stage format (Lappan, Fey, Fitzgerald, Friel, & Phillips, 1998; Lampert, 2001; Sherin, 2002;
Stein, Engle, Smith & Hughes, 2008). In this framework, the teacher first introduces a problem-
solving scenario and then provides students with time to work on the problem. The lesson
concludes with a whole-class discussion in which the students present and justify their problem-
solving approaches and solutions, with the teacher guiding a discussion to highlight connections
between these approaches and the underlying mathematical concepts.

The benefits of this lesson structure are supported by a series of studies. Qualitative
comparative analysis of education in high-income countries reveals that this approach is similar
to those used in mathematics classes Japan and Shanghai, China – two locales which lie near the
top of the international rankings for student achievement in mathematics (OECD, 2009). In the
U.S., researchers in the Quantitative Understanding: Amplifying Student Achievement and
Reasoning (QUASAR) Project found that students more frequently exposed to cognitively
demanding tasks during instruction demonstrated larger gains in student learning, as measured
by a researcher-developed performance assessment of problem solving, reasoning and
communication (Lane, Liu, Ankenmann, & Stone, 1996; Stein & Lane, 1996). Importantly, the
largest gains were detected in classes in which the teachers supported the use of multiple solution strategies and then facilitated a whole class discussion of these strategies to build students’ conceptual understanding. Whole-class discussion which develops students’ understandings following work on a rigorous task is a key feature of classroom practice for ambitious teaching and learning of mathematics (Cobb, Boufi, McClain, & Whitenack, 1997; Kazemi & Stipek, 2001; Stein et al., 2008). When integrated into lessons featuring cognitively demanding tasks, effective whole-class discussions may allow students to articulate and evaluate their thinking and the thinking of others; reveal and redirect gaps in reasoning; establish the connections between multiple solutions strategies and representations to reveal underlying mathematical concepts; and develop both domain-general and discipline-specific skills of discourse (Stein, et al., 2008).

Used in this way and for these purposes, a concluding whole-class discussion is an important, research-supported element of quality teaching in mathematics. However, it is unclear that students of teachers who engage in these teaching practices are any more likely to perform better on tests which do not directly assess these kinds of skills. Some commentary on education policy maintains that more so than the curricular standards, it is the assessments which drive instructional change, that “what gets tested, gets taught” (Firestone, Mayrowetz, & Fairman, 1998; Herman, 2004). Though the newly written CCSS in mathematics can be interpreted as encouraging this kind of activity in the classroom, it seems less likely to happen in a comprehensive and widespread way if students’ participation in these activities is not clearly linked to their assessments. Furthermore, the weak and inconsistent relationship between measures of the content and form of classroom teaching and learning and students’ scores on existing standardized tests suggests that the assessments currently utilized for accountability are insensitive to instruction and, in that way, are not sufficiently aligned to systemic goals. If the
adoption and implementation of rigorous standards for teaching and learning continues to outpace the adoption and implementation of rigorous assessments of student achievement, the alignment between policy-relevant measures of instruction, student learning, and teacher effectiveness may grow weaker still, eroding the effectiveness of standards-based reform policies in education.

**Data**

Data for this analysis comes from ([project description omitted to protect the integrity of the peer-review process]) ([Author 3, 2008]). Within each of the four large urban school districts, six to ten middle schools were selected purposefully to construct a sample of middle schools which reflected the school-level variation of student demographics and achievement within each district. At each school site, three to five mathematics teachers were chosen randomly from a school roster by our researchers and recruited for participation. When a teacher left the school or study, another teacher from the same school was chosen at random and recruited to maintain the same number of participants.

**Teacher Value-added**

Districts supplied student achievement data linked to both participating and non-participating teachers in our study schools. Teacher value-added (TVA) was estimated separately for each district and each year using a four-level covariate adjusted model, specified as follows:

\[
\begin{align*}
\gamma_{ijkl} &= \beta_{0jkl} + \beta_{1jkl}x_{ijkl} + \epsilon_{ijkl} \\
\beta_{0jkl} &= \gamma_{00kl} + \gamma_{01kl}w_{jkl} + \eta_{0jkl} \\
\eta_{00kl} &= \pi_{000l} + \xi_{00kl} \\
\pi_{000l} &= \theta_{0000} + \zeta_{0000l}
\end{align*}
\]
where $x_{ijkl}$ is a vector of student-level demographics variables\(^2\) for student $i$ in class $j$ taught by teacher $k$ in school $l$. This model includes class-level demographic- and prior achievement variables\(^3\) (vector $w_{ijkl}$) and incorporates class-, teacher-, and school-level random effects ($\eta_{0ijkl}$, $\xi_{00kl}$, and $\zeta_{000l}$, respectively). In these models, the estimated teacher-level random effect is recovered and used as an estimate of teacher value-added.

Because these models control for student- and class-level characteristics, the resulting estimates are described in the literature as teacher value-added and are often used as noisy and imprecise measures of teacher effectiveness (Corcoran & Goldhaber, 2013). When school-level random effects are included in the model, as above, the resulting score can be interpreted as the average performance of a given teacher’s students on the end-of-the-year test relative to similar students in similar classes in the same school. There is some disagreement over when school-level random effects should or should not be included in these models if they are to estimate teacher effectiveness (McCaffrey, Lockwood, Koretz, & Hamilton, 2003), considering that students and teachers are not randomly assigned to schools and because school-level factors such as climate and working conditions can also influence student learning (Uline & Tschannen-Moran, 2008). For these reasons, we also estimate value-added scores which omit school-level random effects and include them in our analysis for comparison and as a robustness check. In models where school-level random effects are omitted, the resulting estimated teacher value-added measure is more representative of each teacher's effectiveness relative to other teachers across the district as a whole, but may also “give credit” or “penalize” teachers for school-level contributions to student learning for which the teacher might not be fully responsible. The teacher value-added estimate calculated by using the model described in equations [1] through [4] is referred to in this paper as within-school teacher value-added TVA, while an alternate
measure, calculated after omitting school level random effects ($\zeta_{000}$) from the model, is referred to as within-district TVA. In our data, these two methods generate measures which, for the group as a whole, are very similar, with a Pearson correlation coefficient of 0.893 and a Spearman's rank-order correlation coefficient of 0.886. The standard deviation for the within-school and within-district TVA measures are 0.10 and 0.12 respectively (see Table 1), which is comparable to teacher-effect sizes estimated in previous literature (Goldhaber & Hansen, 2010; Nye, et al., 2004; Rivkin, et al., 2005). We fit both models separately for each year in each district, such that teacher random effect estimates are mean-centered for each combination of district and study year.

**Teacher Observational Measure**

Building upon earlier work articulating frameworks for assessing the rigor of mathematical tasks and the ensuing cognitive demand during the implementation of these tasks in classroom (Stein, Grover, & Henningsen, 1996; Stein & Lane, 1996), a team of researchers at the University of Pittsburgh developed the Instructional Quality Assessment (IQA, Boston, 2012; Boston & Wolf, 2006; Matsumura et al, 2006). This series of rubrics for classroom observation align with the goals and three-part “launch-explore-summarize” structure of the inquiry- and discussion-based approaches to math teaching and learning adopted by our districts. While the entire IQA consists of 20 rubrics which are intended to categorize the rigor of students’ learning opportunities (Boston, 2012), for this study, we restrict our analysis to the three rubrics which most broadly characterize the academic rigor of classroom instruction and which correspond to the lesson format which teachers in our study were attempting to implement. The three rubrics utilized in our study are: *task potential, academic rigor of task implementation, and academic rigor of discussion* (See Appendix A).
In this framework, the task potential is described as the complexity or rigor of student thinking required to successfully complete the task as it appears in print form in curricular or instructional materials. In contrast, the task implementation rubric intends to characterize what actually occurs in the classroom, asking the question: *At what level did the teacher guide students to engage with the task in implementation?* In practice, students may not consistently realize the high levels of thinking called for by a rigorous task as it appears in the intended curriculum. For example, teachers have been observed lowering the cognitive demand of tasks by telling students to complete only part of the written task (Author 2, 2013, 2014). Alternately, teachers also lower the cognitive demand of the task when they provide multiple, step-by-step examples illustrating the solution to a similar problem scenario (Author 2, 2013, 2014; Henningsen & Stein, 1997). When this occurs, students are no longer required to apply or develop genuine problem-solving skills but are instead asked to apply a mathematical procedure which has been explicitly specified by the teacher. Additionally, students might not be able to realize the levels of rigorous thinking called for in a task when expectations are unclear, when the classroom environment is distracting or chaotic, or when tasks are not appropriate given students’ current knowledge and skills (Henningsen & Stein, 1997; Stein & Lane, 1996).

Under the task potential and implementation rubrics, scores range from 1-4, with a score of 1 representing student thinking that requires only the recall of memorized terms, definitions, or formulae. Scores of 2 are assigned when students apply prescribed mathematical procedures to calculate answers for problems. Level 3 tasks require students to cultivate meaning around the application of a mathematical procedure and to make connections to underlying mathematical ideas in the task through identifying patterns, making conjectures, or using multiple problem-solving strategies or representations. Finally, scores of 4 are reserved for tasks and activities
which have all of the qualities required for a score of 3 but which also explicitly require students to explain and justify their solution and method.

The discussion rigor rubric is scored on a 0 to 4 scale and guided by the key question: *During a whole-class discussion following work on the mathematical task, to what extent did students show their work and explain their thinking about the important mathematical content?* A score of 0 indicates there was no concluding whole-class discussion. A score of 1 indicates that students provide brief or one-word answers in a whole-class discussion. A score of 2 indicates that, in a whole-class format, students describe their written work for solving the task but do not engage in a discussion of their strategies, procedures, or mathematical ideas. A score of 3 indicates that students show or describe written work for solving a task and/or engage in a discussion of the important mathematical ideas in the task. During a level 3 discussion, students provide explanations of why their strategy, idea, or procedure is valid and/or students begin to make connections between procedures and mathematical concepts, but the explanations and connections are not complete and thorough. A score of 4 indicates that, during the discussion, students provide thorough explanations of why particular strategies are valid and make connections between strategies and the underlying mathematical ideas.

In the early spring of each of the four years of the research project, we video-recorded two (ideally consecutive) mathematics lessons conducted by each of the approximately 120 teachers participating in the study. Trained coders later scored these videos using the IQA rubrics. Each year, initial interrater reliability was established and monitored on an ongoing basis; across the three academic rigor rubrics and across the four years of data collection, percent agreement averaged 70.5%, with kappa scores averaging 0.50. These reliability statistics are comparable to those from other classroom observation instruments used in the MET Project (see...
Table 1). Well-cited rules of thumb would characterize these reliabilities as ranging from “fair” to “substantial” (Landis & Koch, 1977), although Hartman, Barrios and Wood (2004) suggest that lower agreement rates (in the range of 70%) are to be expected of more complex instruments and can, in some circumstances, be considered sufficient.

We created one set of scores for each teacher in each study year by choosing the highest value observed in a teacher’s classroom on a given rubric across the two days of coded instruction. Because teachers knew they would be video-recorded and had been asked to engage students in a problem-solving lesson, we do not assume that our sample of their instruction is necessarily representative of their typical classroom practice. Instead, we consider the best score on each rubric to represent the upper end of the teacher’s capacity for implementing cognitively demanding tasks with subsequent whole-class discussion in their current school context.

**Different approaches to modeling the independent variable.** Maximum likelihood factor analysis (Preacher & MacCullum, 2003) of teachers’ best score on the task potential, implementation, and discussion rubrics in a given year revealed that these three variables loaded onto a single factor. We use these factor loadings to calculate values for a composite IQA score, which we then use as an independent variable. However, given the ordinal nature of the measures of the IQA rubrics, we find it theoretically justified to model this relationship in ways which might reveal possible non-linear relationships between movements along these scales and changes in teacher value-added measures. In these analyses, we use indicator variables to model each level of the four- or five-value ordinal scales articulated in the IQA’s academic rigor rubrics. Finally, we also compare these results with a more parsimonious model which operationalizes task potential, implementation, and discussion as either “high” or “low,” with scores of 1 or 2 categorized as “low” and scores of 3 or 4 categorized as “high.” Recent analyses
by IQA developers utilized this high-low distinction in these rubrics (Boston, 2012), mirroring earlier work using a high-low distinction in analysis of classroom tasks in mathematics (Henningsen & Stein, 1997; Stein & Lane, 1996). Given this precedent, we are comfortable dichotomizing these four-point scales, contrasting “low” levels of task potential and implementation (1 and 2) with “high” levels (3 and 4).

**Methods**

In investigating the relationship between teachers’ scores on the IQA rubrics and teacher value-added estimates, we fit a three-level hierarchical linear model (HLM):

\[
TVA_{ykl} = \alpha_{0kl} + \alpha_{1kl}x_{ykl} + \delta_{ykl}
\]  

(5)

\[
\alpha_{0kl} = \kappa_{00l} + \lambda_{0kl}
\]  

(6)

\[
\kappa_{00l} = \omega_{000} + \omega_{00k}District_k + \varphi_{00l}
\]  

(7)

where we have repeated estimates of teacher value-added (\(TVA_{yk}\)), one for each study year \(y\), for teacher \(k\) in school \(l\). Teacher value-added is modeled as a function of a variable (or number of variables, represented by vector \(x_{ykl}\)) derived from measures of classroom observation from the IQA instrument.\(^6\) We include in this model teacher- and school level random effects (\(\lambda_{0kl}\) and \(\varphi_{00l}\)) and a district fixed effect (\(District_k\)). As a robustness check, we include a teacher fixed-effect model of the following specification:

\[
TVA_{yk} = \nu_{0k} + \nu_{1k}x_{yk} + \psi_{yk}
\]  

(8)

again with repeated estimates of teacher value-added (\(TVA\)) for each teacher \(k\) in year \(y\) as the predicted variable, modeled as a function of \(x_{ykl}\), a vector of scores from observations using the IQA rubrics, with \(\nu_{0k}\) as the teacher-level fixed effect and \(\psi_{yk}\) the year-level residual.
Results

Describing the distribution of these measures

Means and standard deviations of key variables in this study are presented in Table 2. Since teacher value-added is mean-centered within district each year, we would expect averages of these variables to be very close to zero. In District C, the averages for both within-district and within-school teacher value-added estimates are negative, indicating that participants in our study had, on average, lower teacher value-added estimates than their non-participating counterparts in the same schools, although this difference is not significant at the 0.05 level. We also see systematic differences between districts in IQA composite scores, with teachers in District A having on average the highest scores and teachers in District C having scores that are, on the whole, lower than the average within our study. This systematic variation between districts confirms that our models should include district fixed effects to control for district-level confounding variables.

Table 3 gives the distribution of scores within each of the three rubrics. We see that even in our sample size of 461 teacher-year observations (nested within 221 teachers over 4 years), some values within these rubrics occur relatively infrequently in this sample. For example, even with our decision rule of taking a teacher’s highest score for each rubric across two annual observations, we observed only one instance of a teacher presenting a Level 1 task, representing 0.2% of the sample. However, a classroom observation with a Level 4 discussion was also relatively rare (3.2% of the sample). For our regression analysis, when scores for a level within a subscale occurred at a relatively low frequency (i.e. less than 5% of the total sample), two adjacent levels within an IQA subscale were grouped together, with one indicator variable used
to represent both groups. This collapsing most typically occurred for values at the extremes of the scale (i.e. 1 or 4 on a four point scale).

Cross-tabulations (Table 4) reveal that observing a teacher present a high-level task is very strongly associated with the highest task implementation level and discussion scores observed in the same year. For example, high implementation scores (3 or 4) were 8 times more frequent among teachers observed selecting high rigor tasks than for teacher observed only selecting low rigor tasks. Additionally, discussion scores of 3 or greater were observed in approximately 27 percent of classes where high level tasks were observed, whereas only 7 percent of classrooms with low level tasks reached the same level of discussion. Classrooms where only low rigor tasks were observed were more than twice as likely to have no observable whole class discussion (i.e. a discussion score of zero) then were classrooms where a high rigor task was observed.

Further analysis of the distribution of these scores reveals differences between districts, with some change over time (Figure 2). In districts A, B, and D, the percentage of teachers observed using high-rigor tasks showed no dramatic overall increases over the duration of our study, averaging to 82, 75, and 73 percent across four years for those three districts, respectively. However, compared to other districts, teachers in District C presented high-level tasks to their students much less frequently during the first year of data collection, and this practice became even less common over time. In Year 1 of the study, we observed 57 percent of teachers in District C presenting students with high level tasks; by year four of the study, the percentage of teacher in District C presenting students with a high-level task had decreased to 29 percent.
Results from HLM and teacher fixed effects regressions

In the HLM analysis (Table 5), the IQA composite score is very weakly associated with teacher value-added and the relationship is not significant at standard confidence levels. The models employing indicator variables for different levels in the three IQA rubrics suggest that the relationship between the subscales and teacher value-added estimates are not linear. A large portion of the effect of the IQA composite on teacher value-added seems to be driven by scores on the task potential rubric, with more rigorous tasks (scored at 3 or 4) associated with average increases in student achievement scores ranging from 0.026 to 0.042 standard deviations, with p-values ranging from less than 0.05 to less than 0.001 (Table 5, columns 3 and 4). This effect size on teacher value-added is equivalent to an addition 19 to 30 days of classroom instruction.\textsuperscript{7}

Contrary to expectation, the estimated effect sizes for choosing a level 4 task is slightly smaller than the estimates for choosing a level 3 task, although these differences are not significant at conventional levels of significance. Additionally, after controlling for task potential, increases in task implementation or discussion are not associated with increases in teacher value-added at conventional levels of significance. Because the effect-size difference between level 3 and level 4 task potential is small – and because of the practice in previous research of grouping together these two levels – a model was tested using only a “high/low” indicator of task potential, with levels 3 and 4 collapsed to form a single “high rigor” category. The results from this model estimate that, on average, students in classrooms where high level tasks were observed outperformed similar students by 0.031 (p=0.001) to 0.041 (p<0.001) standard deviations, or the equivalent of 22 to 30 additional days of instruction.

For the fixed-effects models, a Wooldridge test for autocorrelation in panel data (Drucker, 2003; Wooldridge, 2002) estimated the within panel correlation of error terms to be
−0.36 for the full model presented in Table 6, column 3 and −0.31 for the model presented in Table 6, column 4 (both p<0.05). Consequently, standard errors adjusted for clustering of observations on the teacher-level were estimated and reported here (Nichols & Schaffer, 2007).

Overall, results from the teacher fixed effects models (Table 6) are similar to those from the HLM regressions. Because the teacher fixed effects models only seek to explain changes of teacher value-added within individuals over time, the size and statistical significance of the relationship between the selection of high-rigor tasks and teacher value-added suggests that this variable does account for some of the growth or decline in individual teacher value-added estimates over time.

**Significance: Implications for policy and practice**

The most current, state-of-the-field recommendations for judging teacher effectiveness advocate for evaluating teacher effectiveness with data from both classroom observation and student test scores (Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Ideally, data from both of these kinds of measures would not only identify particularly effective or ineffective teachers but also provide teachers with feedback to improve their practice (Kane & Staiger, 2012). Looking at the relationship between our measures of student learning and individual IQA rubrics – and allowing levels within these rubrics to have different and nonlinear relationships with the outcome variable – gives more insight into the relationship between these two kinds of measures and has implications for instruction which analysis of a single IQA composite variable could not reveal.

Results from the teacher fixed effects analysis allows us to infer that, from year to year, teachers who have increased the rigor of the tasks they give their student have generally improved their value-added scores over time, while teachers who have shifted toward using
lower-level tasks have seen reductions in their teacher value-added estimates. This inference is potentially informative to practice, because it not only suggests that schools can improve student learning and learning opportunities by being staffed by teachers who tend to choose higher level tasks, but that schools can improve student learning and learning opportunities by supporting current teachers to utilize high value tasks in their own teaching.

It might be tempting to posit an endogenous explanation for some of these relationships. One might conjecture that teachers in this sample may present more challenging tasks to students who show higher levels of previous ability. The fact that we employ value-added scores which control for prior ability at the student and class levels should account for this possible endogenous relationship. In addition, unadjusted correlations (not shown) indicate that the teacher’s choice of lower rigor tasks is, in these data, not correlated with the aggregate prior achievement level of a teacher’s students. One might also conjecture that certain kinds of teachers might be more likely to take note of the prearranged date of video recording and make an effort to plan classroom activities which strongly conformed to the researchers’ requests for lessons focusing on problem solving and incorporating a whole-class discussion. Specifically, teachers who are more conscientious are more likely to conform to researchers’ requests than less conscientious teachers; other research has linked contentiousness to job performance in a wide variety of settings (Barrick & Mount, 1991), including teaching (Rockoff, Jacob, Kane, & Staiger, 2011). However, teacher fixed effects models (and to some degree, HLM models with teacher random effects) do allow us to control for time-invariant teacher traits such as contentiousness. However, it may be that teachers in our sample experience particularly challenging years in which there may be many demands on their time, and that in these years in particular, teachers could be conceivably be less likely to both adjust instruction to researchers
request and realize above average gains in student learning. Our data and models cannot control for this potentially confounding factor.

As important as the rigor of the mathematical task has been shown to be in this and other analyses (e.g. Stein & Lane, 1996), teachers’ choices of classroom tasks is heavily influenced by the curricular materials with which they are provided. The selection of curricular materials provided to teachers at a given school is, in turn, dictated by organizational choices made at the school district level. The fact that, in most states, school districts determine textbook adoption is well documented (Cobb, McClain, de Silva Lamberg & Dean, 2003; Spillane, 1996). We found this to be true in our four study districts at least for the period of time described in this analysis.

Between 1990 and 2007, the National Science Foundation spent an estimated $93 million developing educational models and materials to support a greater emphasis in problem solving, conceptual understanding, reasoning, and justification in K-12 mathematics (National Research Council, 2004). By Year 1 of our study, Districts A, B, and D had adopted an NSF-supported curriculum mathematics curriculum in pursuit of these kinds of goals. District C had adopted an NSF-supported curriculum approximately 5 years before data collection for this study commenced. However, through our interviews with teachers and administrators in District C, we learned that this adoption of an NSF-supported textbook was ultimately unpopular with teachers, who were reportedly given neither appropriate opportunities for input before the textbook adoption decision was made nor adequate professional development in using the new curricular materials after adoption. As a result, the teachers failed to develop “buy-in” with regards to the more ambitious curricula and the accompanying pedagogical approach. When it was time to adopt a new mathematics curriculum one year prior to the first year of our study (i.e. Year 0), the teachers in District C voted in favor of a more traditional, non-NSF-supported curriculum.
District administrators continued to be committed to a more ambitious approach to mathematics education and began to develop “in house” modules or units for teachers to use to supplement the more traditional curriculum teachers preferred. Whether it be because of the difference in curricula, a lack of buy-in for the inquiry-based approach to teaching, or other unobserved factors, teachers in District C were observed presenting students with high-level tasks much less frequently than teachers in the other districts.

This analysis suggests that a teacher’s selection of rigorous tasks for classroom instruction may have a large, positive impact on the kind of student learning measured by these NCLB-era tests. Results from cross-tabulation indicate that teachers utilizing high-level tasks in a given school year also tended have high-level implementation scores and facilitate more elaborate classroom discussions that school year. For these reasons, the selection of rigorous tasks might appropriately be called a high leverage practice, which Ball and colleagues describe as a teaching practice “in which the proficient enactment by a teacher is likely to lead to comparatively large advances in student learning” (Ball, Sleep, Boerst, & Bass, 2009, p. 460). In work similar to the present study, Stein and Lane (1996) also conclude that selection of high-level mathematical tasks is associated with greater increases in student learning. However, these researchers also assert that the benefits to student learning are somewhat eroded when these high-level tasks are implemented at a lower level of cognitive demand. While Stein and Lane’s study employs cross-sectional data in which schools are presented as the unit of analysis, our analyses likely estimate more precisely the learning benefits of high-level tasks and the implementation and discussion which follow, given that our data allow for longitudinal analysis at the teacher level. While our findings do not suggest that student learning is reduced in classrooms where high level tasks are implemented at a low level, it is also important to note that
the assessments of student learning used in these analyses from the QUASAR study differ substantially from ours. Stein and Lane employed a researcher-designed assessment consisting of open-ended tasks intended to assess mathematical problem solving, reasoning, and communication at the middle-school level (Lane, et al., 1996). By contrasts, our measures of student learning are derived from NCLB era state tests, which have been characterized as overwhelmingly testing basic skills and neglecting to measure higher order thinking or in depth knowledge (Darling-Hammond & Adamson, 2010; Yuan & Le, 2012) and therefore likely have different psychometric properties than the tests used by QUASAR researchers. As such, we interpret our findings as reinforcing Stein and Lane’s conclusions about the importance of task rigor, more generally.

Additionally, the results from this study reinforce other work suggesting that many large-scale standardized tests used for accountability purposes are relatively insensitive to the quality of instruction which occurs in the classroom (D’Agostino, et , 2007; Popham, 2007; Polikoff, 2010, 2014; Ruiz-Primo, et al., 2002), particularly for instructional practices intended to cultivate complex student thinking (Grossman et al, 2014) or when compared with observational measures (Guarino & Stacy, 2012; Walkington & Marder, 2014). In recent years, NLCB’s mandates for every-child, every-year testing have represented a large and reoccurring financial obligation for states, leading them to rely heavily on standardized multiple-choice tests which foreground students’ ability to recall facts and perform basic skills (Darling-Hammond & Adamson, 2010). While the widespread adoption of the Common Core State Standards holds promise for increasing students’ opportunities to learn, states taking the additional step of participating in the CCSS testing consortia also spread the costs of assessment development and implementation while improving assessment quality (Chingos, 2012). However, if financial or political
considerations induce states to defect from the assessment consortia and instead choose to implement legacy tests or newer tests with a similar focus on lower level knowledge and skills, the resulting test-based accountability system might not reward the kinds of instructional practices connected to higher order thinking and skills. The tests of the NCLB era, though less expensive and more easily scored, are likely to be perceived as increasingly antiquated and inadequate given our expectations and aspirations for our education systems. This becomes even more apparent as researchers and reform advocates propose that schools of the near future cultivate and assess not only higher order cognitive skills such as critical thinking, problem solving, and communication, but also “21st century skills” in the interpersonal and intrapersonal domains, including teamwork, leadership, creativity, intrinsic motivation, and grit (e.g. National Research Council, 2012; Soland, Hamilton, & Stecher, 2013; William and Flora Hewlett Foundation, 2012);

Ultimately, the results from this study have interesting and important implications for education policy and practice. The selection of high-level task in mathematics is associated with an outsized impact on student test scores on these NCLB-era tests. Furthermore, complex and sophisticated whole-class discussions are much more likely to occur when the lesson centers on a high-rigor task. These findings highlight the importance of mathematics teachers’ ability to access, recognize, and integrate high-rigor tasks into their lessons. The context of this study suggests that district policies mediate teachers’ access to high-level tasks, such that district leadership and structures should support teachers’ access to and use of high-rigor tasks in mathematics. In the end, more work is needed in both policy and research to ensure that curriculum, instruction, and assessment are aligned in ways that best promote the ambitious
goals and vision we have articulated for our nation’s students and our education system as a whole.
References

[Citations for Author 1, 2, & 3 removed to maintain integrity of the review process]


http://dx.doi.org/10.2307/749781


http://dx.doi.org/10.1162/edfp_a_00104


http://dx.doi.org/10.1080/10627190709336945

WILL THIS BE ON THE TEST

University, Stanford Center for Opportunity Policy in Education. Retrieved from the Stanford Center for Opportunity Policy in Education website:


http://dx.doi.org/10.3102/01623737020002095


http://blogs.edweek.org/edweek/curriculum/2014/06/the_portion_of_students_not.html

http://dx.doi.org/10.1257/aer.100.2.250


http://dx.doi.org/10.3102/0013189x14544542


http://www.greatlakescenter.org/docs/Think_Twice/TT_Guarino_MET.pdf


http://dx.doi.org/10.3102/0013189x025004012


http://dx.doi.org/10.2307/749690


http://dx.doi.org/10.1086/499693


http://dx.doi.org/10.1111/j.1745-3992.2003.tb00124.x


http://dx.doi.org/10.1086/596998


http://dx.doi.org/10.1037/e644712011-001


Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects?. 


http://dx.doi.org/10.3102/01623737026003237


http://dx.doi.org/10.1787/9789264096660-en


http://dx.doi.org/10.3102/0002831210362589


http://dx.doi.org/10.1111/j.1745-3992.2010.00189.x


http://dx.doi.org/10.3102/0162373714531851


http://dx.doi.org/10.1177/0895904896010001004

http://dx.doi.org/10.3102/01623737019002185

http://dx.doi.org/10.1080/10986060802229675

http://dx.doi.org/10.3102/0028312033002455

http://dx.doi.org/10.1080/1380361960020103


Yuan, K., & Le, V. (2012). *Estimating the percentage of students who were tested on cognitively demanding items through the state achievement tests.* Santa Monica, CA: RAND Corporation. Retrieved from the RAND website:

http://www.rand.org/content/dam/rand/pubs/working_papers/2012/RAND_WR967.pdf
Footnotes

1. Specifically, the five instruments chosen for classroom observation were the Framework for Teaching (FFT), the Classroom Assessment Scoring System (CLASS), the Protocol for Language Arts Teaching Observations (PLATO), the Mathematical Quality of Instruction (MQI) instrument, and the UTeach Teacher Observation Protocol (UTOP) (see Kane & Staiger, 2012).

2. Student-level controls include indicator variables for race/ethnicity, gender, free- or reduced price lunch status, grade, English language learner status, special education status, and attendance of math tutoring- or support classes scheduled during the school day. Also included are two prior years of end of year assessment scores in mathematics, including quadratic and cubic terms for each of these scores.

3. Class-level controls include class size, class average prior achievement in years Y-1 and Y-2, the standard deviation of the prior achievement in for Y-1 and Y-2 across the class. Class-level demographics controlled for include percent of students in class English language learners, special education, free-and reduced price lunch, African-American, White, or “Other” ethnicity. Each of these linear terms is also accompanied by its quadratic term in the estimation equation.

4. In addition to utilizing these three more global rubrics, we also scored classroom observations on five additional rubrics which categorized participation levels and made more nuanced distinctions regarding the specific discourse patterns used by students and teachers during whole-class discussions (i.e. the Accountable Talk rubrics) (Boston, 2012). However, we decided to exclude these from analysis here given that, both theoretically and in our data, scores on these rubrics describing the participation and sophistication of particular teacher- and student discourse patterns correlated strongly with scores registered on the discussion rubric.
5. Kappa scores are measures of reliability based on percent agreement but adjusted for the probability of chance agreement given the actual distribution of the data (J. Cohen, 1960).

6. The effects on student learning of interactions between scores on these three rubrics were tested, but none of the results yielded differences of any practical significance.

7. Teacher value-added is scaled in terms of standard deviations of the distribution of student scores, and Hattie (2008) describes a conversion rate of change in achievement tests in standard deviation and instructional days added to or taken away from the traditional 180 day school year, assuming a linear relationship between achievement and days attended. Specifically, Hattie (2008) equates 0.25 standard deviations of growth along a normalized distribution of test scores as equivalent to one year (approximately 180 days) of instruction. This conversion rate is also used in Kane and Staiger, 2012.
Table 1. Percent agreement (kappa statistic in parentheses) for the three academic rigor rubrics of the Instructional Quality Assessment (IQA), by study year. Overall reliability measures compared with those from classroom observation instruments as reported data from the Measures of Effective Teaching (MET) Project (source Park, Chen, & Holtzman, 2014). Classroom observations instruments from the MET project include the Classroom Assessment Scoring System (CLASS), Framework for Teaching (FfT), Mathematical Quality of Instruction (MQI), and the Protocol for Language Arts Teaching Instruction (PLATO).

<table>
<thead>
<tr>
<th></th>
<th>Y1</th>
<th>Y2</th>
<th>Y3</th>
<th>Y4</th>
<th>Avg. Y1-4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Potential</td>
<td>59.4</td>
<td>56.9</td>
<td>75.0</td>
<td>59.1</td>
<td>62.6</td>
</tr>
<tr>
<td></td>
<td>(0.37)</td>
<td>(0.29)</td>
<td>(0.63)</td>
<td>(0.36)</td>
<td>(0.41)</td>
</tr>
<tr>
<td>Task Implementation</td>
<td>78.1</td>
<td>78.5</td>
<td>89.3</td>
<td>63.6</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td>(0.51)</td>
<td>(0.37)</td>
<td>(0.75)</td>
<td>(0.29)</td>
<td>(0.48)</td>
</tr>
<tr>
<td>Discussion</td>
<td>78.1</td>
<td>69.2</td>
<td>67.9</td>
<td>70.5</td>
<td>71.4</td>
</tr>
<tr>
<td></td>
<td>(0.71)</td>
<td>(0.58)</td>
<td>(0.55)</td>
<td>(0.59)</td>
<td>(0.61)</td>
</tr>
<tr>
<td>Yearly Average</td>
<td>71.9</td>
<td>68.2</td>
<td>77.4</td>
<td>64.4</td>
<td>70.5</td>
</tr>
<tr>
<td></td>
<td>(0.53)</td>
<td>(0.41)</td>
<td>(0.64)</td>
<td>(0.41)</td>
<td>(0.50)</td>
</tr>
<tr>
<td>MET CLASS</td>
<td>34</td>
<td>57</td>
<td>76</td>
<td>59</td>
<td>70.5</td>
</tr>
<tr>
<td></td>
<td>(0.21)</td>
<td>(0.24)</td>
<td>(0.51)</td>
<td>(0.45)</td>
<td>(0.50)</td>
</tr>
<tr>
<td>MET FFT MQI PLATO</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Descriptive statistics of dependent and independent variables employed in this study (standard deviations in parenthesis).

<table>
<thead>
<tr>
<th></th>
<th>Dist A</th>
<th>Dist B</th>
<th>Dist C</th>
<th>Dist D</th>
<th>Complete Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-School TVA</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.11)</td>
<td>(0.11)</td>
<td>(0.17)</td>
<td>(0.08)</td>
<td>(0.12)</td>
</tr>
<tr>
<td>Within-District TVA</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.02</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.09)</td>
<td>(0.11)</td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
<tr>
<td>IQA Composite (normalized)</td>
<td>0.37</td>
<td>0.03</td>
<td>-0.41</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td></td>
<td>(0.86)</td>
<td>(0.68)</td>
<td>(0.58)</td>
<td>(0.83)</td>
<td>(0.78)</td>
</tr>
<tr>
<td>IQA Subscale: Task Potential</td>
<td>3.18</td>
<td>3.12</td>
<td>2.63</td>
<td>3.09</td>
<td>3.00</td>
</tr>
<tr>
<td></td>
<td>(0.60)</td>
<td>(0.72)</td>
<td>(0.64)</td>
<td>(0.76)</td>
<td>(0.72)</td>
</tr>
<tr>
<td>IQA Subscale: Task Implementation</td>
<td>2.74</td>
<td>2.37</td>
<td>2.18</td>
<td>2.42</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>(0.70)</td>
<td>(0.52)</td>
<td>(0.41)</td>
<td>(0.64)</td>
<td>(0.60)</td>
</tr>
<tr>
<td>IQA Subscale: Discussion</td>
<td>2.06</td>
<td>1.87</td>
<td>1.46</td>
<td>1.92</td>
<td>1.82</td>
</tr>
<tr>
<td></td>
<td>(1.12)</td>
<td>(0.87)</td>
<td>(0.93)</td>
<td>(0.94)</td>
<td>(0.98)</td>
</tr>
<tr>
<td>High Task (Task = 3 or 4)</td>
<td>0.89</td>
<td>0.81</td>
<td>0.43</td>
<td>0.76</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>(0.31)</td>
<td>(0.39)</td>
<td>(0.50)</td>
<td>(0.43)</td>
<td>(0.43)</td>
</tr>
<tr>
<td>Observations</td>
<td>94</td>
<td>134</td>
<td>114</td>
<td>119</td>
<td>461</td>
</tr>
</tbody>
</table>
### Table 3. Distribution of IQA subscale scores by level: Counts and percentages

<table>
<thead>
<tr>
<th></th>
<th>Level 0</th>
<th>Level 1</th>
<th>Level 2</th>
<th>Level 3</th>
<th>Level 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Counts</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Potential</td>
<td>0</td>
<td>1</td>
<td>115</td>
<td>226</td>
<td>119</td>
</tr>
<tr>
<td>Task Implementation</td>
<td>0</td>
<td>5</td>
<td>282</td>
<td>152</td>
<td>22</td>
</tr>
<tr>
<td>Discussion</td>
<td>51</td>
<td>98</td>
<td>211</td>
<td>83</td>
<td>18</td>
</tr>
<tr>
<td><strong>Percentages</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task Potential</td>
<td>0.0%</td>
<td>0.2%</td>
<td>25.2%</td>
<td>49.0%</td>
<td>25.8%</td>
</tr>
<tr>
<td>Task Implementation</td>
<td>0.0%</td>
<td>1.1%</td>
<td>61.2%</td>
<td>33.0%</td>
<td>4.8%</td>
</tr>
<tr>
<td>Discussion</td>
<td>11.1%</td>
<td>21.3%</td>
<td>45.8%</td>
<td>18.0%</td>
<td>3.9%</td>
</tr>
</tbody>
</table>
Table 4. Cross tabulation of task level and implementation level, discussion score, with row percentages reported in parentheses under frequencies.

<table>
<thead>
<tr>
<th>Implementation Score</th>
<th>Low (i.e. ≤2)</th>
<th>High (i.e. &gt;2)</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Low (i.e. ≤2)</td>
<td>109</td>
<td>7</td>
<td>66.2***</td>
</tr>
<tr>
<td></td>
<td>(94.0)</td>
<td>(6.0)</td>
<td></td>
</tr>
<tr>
<td>Task High (i.e. &gt;2)</td>
<td>178</td>
<td>167</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(51.6)</td>
<td>(48.4)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Discussion Score</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>χ²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task Low (i.e. ≤2)</td>
<td>22</td>
<td>35</td>
<td>51</td>
<td>7</td>
<td>1</td>
<td>136.0***</td>
</tr>
<tr>
<td></td>
<td>(19.0)</td>
<td>(30.2)</td>
<td>(44.0)</td>
<td>(6.0)</td>
<td>(0.9)</td>
<td></td>
</tr>
<tr>
<td>Task High (i.e. &gt;2)</td>
<td>29</td>
<td>63</td>
<td>160</td>
<td>76</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(8.4)</td>
<td>(18.3)</td>
<td>(46.4)</td>
<td>(22.0)</td>
<td>(4.9)</td>
<td></td>
</tr>
</tbody>
</table>

Note. ***= p ≤ 0.001.
Figure 1 Distribution of observations of IQA typology, by district over time.
Table 5: Hierarchical Linear Model with school- and teacher level random effects and district level fixed effects. (Note: the base group for regression columns 3 and 4 is task, implementation, and discussion scores of $(\leq 2, \leq 2, 0)$

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Within-School</td>
<td>TVA</td>
<td>TVA</td>
<td>TVA</td>
<td>TVA</td>
<td>TVA</td>
<td>TVA</td>
</tr>
<tr>
<td>TVA</td>
<td>0.010*</td>
<td>0.013*</td>
<td>0.031**</td>
<td>0.043***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td>(0.006)</td>
<td>(0.007)</td>
<td>(0.010)</td>
<td>(0.012)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Task=3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation</td>
<td>-0.001</td>
<td>-0.003</td>
<td>(0.015)</td>
<td>(0.012)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 3 or 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>-0.006</td>
<td>-0.013</td>
<td>(0.015)</td>
<td>(0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>-0.001</td>
<td>-0.005</td>
<td>(0.013)</td>
<td>(0.016)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td>0.002</td>
<td>0.004</td>
<td>(0.017)</td>
<td>(0.020)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>=3 or 4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Task</td>
<td>0.030</td>
<td>0.041***</td>
<td>(0.010)</td>
<td>(0.011)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>District FE</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.001</td>
<td>-0.002</td>
<td>-0.023</td>
<td>-0.028</td>
<td>-0.026***</td>
<td>-0.035</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.019)</td>
<td>(0.016)</td>
<td>(0.025)</td>
<td>(0.012)</td>
<td>(0.021)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>N</th>
<th>461</th>
<th>461</th>
<th>461</th>
<th>461</th>
<th>461</th>
</tr>
</thead>
<tbody>
<tr>
<td>$AIC$</td>
<td>-925.076</td>
<td>-728.213</td>
<td>-923.363</td>
<td>-729.448</td>
<td>-932.564</td>
<td>-737.523</td>
</tr>
<tr>
<td>$BIC$</td>
<td>-892.009</td>
<td>-695.146</td>
<td>-869.629</td>
<td>-675.713</td>
<td>-899.497</td>
<td>-704.456</td>
</tr>
</tbody>
</table>

Standard errors adjusted for clustering at the teacher level in parentheses

* $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
Table 6: Results from teacher fixed effects regression. (Note: the base group for regression columns 3 and 4 is task, implementation, and discussion scores of \( \leq 2, \leq 2, 0 \))

<table>
<thead>
<tr>
<th></th>
<th>(1)</th>
<th>(2)</th>
<th>(3)</th>
<th>(4)</th>
<th>(5)</th>
<th>(6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Within-School TVA</td>
<td>Within-District TVA</td>
<td>Within-School TVA</td>
<td>Within-District TVA</td>
<td>Within-School TVA</td>
<td>Within-District TVA</td>
</tr>
<tr>
<td>IQA</td>
<td>0.007</td>
<td>0.008</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Composite</td>
<td>(0.006)</td>
<td>(0.008)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task=3</td>
<td></td>
<td></td>
<td>0.031**</td>
<td>0.043**</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.013)</td>
<td>(0.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task=4</td>
<td></td>
<td></td>
<td>0.026^</td>
<td>0.034^</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>(0.014)</td>
<td>(0.018)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Implementation</td>
<td></td>
<td></td>
<td>-0.004</td>
<td>-0.009</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 3 or 4</td>
<td></td>
<td></td>
<td>(0.012)</td>
<td>(0.015)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
<td></td>
<td>-0.000</td>
<td>-0.014</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 1</td>
<td></td>
<td></td>
<td>(0.019)</td>
<td>(0.023)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
<td></td>
<td>0.008</td>
<td>0.003</td>
<td></td>
<td></td>
</tr>
<tr>
<td>= 2</td>
<td></td>
<td></td>
<td>(0.018)</td>
<td>(0.023)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discussion</td>
<td></td>
<td></td>
<td>0.003</td>
<td>-0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>=3 or 4</td>
<td></td>
<td></td>
<td>(0.022)</td>
<td>(0.028)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High Task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.030**</td>
<td>0.040**</td>
</tr>
<tr>
<td>(=3 or 4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>(0.010)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Intercept</td>
<td>-0.003***</td>
<td>-0.005***</td>
<td>-0.028^</td>
<td>-0.029</td>
<td>-0.026**</td>
<td>-0.034***</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.015)</td>
<td>(0.019)</td>
<td>(0.008)</td>
<td>(0.009)</td>
</tr>
<tr>
<td>N</td>
<td>461</td>
<td>461</td>
<td>461</td>
<td>461</td>
<td>461</td>
<td>461</td>
</tr>
<tr>
<td>( R^2 )-overall</td>
<td>0.014</td>
<td>0.018</td>
<td>0.024</td>
<td>0.018</td>
<td>0.029</td>
<td>0.028</td>
</tr>
</tbody>
</table>

Standard errors in parentheses
^+ p < 0.10, * p < 0.05, ** p < 0.01, *** p < 0.001