8-2012

Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice

Heather C. Hill  
*Harvard Graduate School of Education*

Kristin Umland  
*University of New Mexico*

Erica Litke  
*Harvard Graduate School of Education*

Laura R. Kapitula  
*Grand Valley State University, kapitull@gvsu.edu*

Follow this and additional works at: [https://scholarworks.gvsu.edu/sta_articles](https://scholarworks.gvsu.edu/sta_articles)

**Recommended Citation**

Hill, Heather C.; Umland, Kristin; Litke, Erica; and Kapitula, Laura R., "Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice" (2012). *Peer Reviewed Articles*. 2.  
[https://scholarworks.gvsu.edu/sta_articles/2](https://scholarworks.gvsu.edu/sta_articles/2)

*This Article is brought to you for free and open access by the Statistics Department at ScholarWorks@GVSU. It has been accepted for inclusion in Peer Reviewed Articles by an authorized administrator of ScholarWorks@GVSU. For more information, please contact scholarworks@gvsu.edu.*
Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice

HEATHER C. HILL
Harvard Graduate School of Education

KRISTIN UMLAND
University of New Mexico

ERICA LITKE
Harvard Graduate School of Education

LAURA R. KAPITULA
Grand Valley State University

Multiple-choice assessments are frequently used for gauging teacher quality. However, research seldom examines whether results from such assessments generalize to practice. To illuminate this issue, we compare teacher performance on a mathematics assessment, during mathematics instruction, and by student performance on a state assessment. Poor performance on the written assessment predicted poor classroom performance; likewise, strong performance on the written assessment predicted strong classroom performance. However, classroom performance of teachers in the middle of the written assessment distribution varied considerably. These results also held true using student outcomes as a criterion, leading to implications for both research and policy.

Measuring the quality of teachers and teaching is important to a broad array of educational stakeholders. The initial teacher certification process in most states involves an exam meant to eliminate candidates unfit for teaching. Once certified, principals and other school personnel conduct observations of practice in order to make tenure and retention decisions. Salary and pay are based on a different set of measures—years of experience in the classroom, degrees attained, and sometimes teacher “value-added” scores produced from student performance on state assessments. Promotions to subject-matter specialist,
coaching positions, or professional development responsibilities may be based
on some combination of personality, motivation, classroom performance, ac-
dademic degrees, and external credentials such as National Board Certification.

However, many view these current efforts to measure teachers and teaching
as imperfect, inadequately capturing true variation in teaching quality. Despite
being commonly used as a determinant of salary increases, teacher experience
and degrees have been found to hold only modest explanatory power in models
predicting student outcomes (Goldhaber and Brewer 1997; Hanushek 1996;
Kane, Rockoff, and Staiger 2008; Monk 1994; Wayne and Youngs 2003). Ad
hoc observations performed by principals are little better; one widely cited
report notes that in several large districts, only 1% of teachers received “un-
satisfactory” ratings (Weisberg et al. 2009). Value-added scores are seen by
many to suffer from a lack of reliability and validity (Amrein-Beardsley 2008;
exams are generally seen as only weak predictors of student outcomes (Han-
ushek 1996) and may negatively affect teacher labor markets (Angrist and
Guryon 2008; Goldhaber 2007). Importantly, few studies have compared re-
sults from certification assessments to classroom practice; instead, most cer-
tification assessments are validated through “job analysis,” which justifies the
inclusion of topics and tasks presented on the assessment but fails to examine
whether performance on these assessments predict future practice.

Thus, evidence about existing indicators of teacher quality suggests that
there is a clear need for better measures of this construct as well as more
information about the validity of those measures. To this end, this report
explores the utility of a relatively new mathematics multiple-choice assessment
by determining the amount of congruence between teacher scores on the
assessment, performance in several domains of mathematical instructional

HEATHER C. HILL is a professor at the Harvard Graduate School of
Education; e-mail: heather_hill@harvard.edu. Her research interests in-
clude teacher quality, mathematics instruction, and teacher mathematical
knowledge. KRISTIN UMLAND is an associate professor, Department of
Mathematics and Statistics, University of New Mexico; e-mail: umland@
math.unm.edu. Her research interests include teacher mathematical knowl-
edge, teacher professional development, and value-added modeling. ERICA
LITKE is a graduate student at the Harvard Graduate School of Education;
email: egl278@mail.harvard.edu. Her research interests include secondary
mathematics instruction, issues of equity in mathematics, and the classroom
implications of education policy. LAURA KAPITULA is an assistant professor,
Department of Statistics, Grand Valley State University; e-mail: lkapitula@g
mail.com. Her research interests include quantitative evaluation, value-added
modeling, and influence diagnostics.

490 American Journal of Education
practice, and student mathematics outcomes. Because this assessment bears a resemblance to certification tests used in the United States, it sheds light on the use of such tests to screen candidates for suitability for teaching. This report also investigates whether the multiple-choice assessment might be useful in helping to identify expert teachers for leadership roles, for example, coaches, mentor teachers, or subject matter specialists.

Our findings suggest that this multiple-choice assessment, known as the Mathematical Knowledge for Teaching (MKT) instrument, can be used as a rough first cut to categorize teachers for further inquiry. Specifically, this instrument does well in predicting teacher performance, especially around proposed cut-points; we elaborate on this theme in a pair of case studies. However, the MKT does not discriminate well for those with test scores in the middle of the distribution. Implications for measuring teacher and teaching quality are discussed.

Framing the Problem

Over the past century, a wide variety of measurement methods have been deployed for measuring teacher and teaching quality. To simplify our review of this terrain, we focus on two areas: methods used to certify candidates for practice and methods used to reward expert teachers with promotion into content-specific leadership positions.

Certification Decisions

All states require the majority of prospective teachers to pass an exam prior to entry to teaching. Forty-six states use some version of the Praxis Series assessments, either to screen candidates prior to enrollment in an education program, as a component of the certification process, or both. While the Praxis is developed and managed by the Educational Testing Service (ETS), individual states select which assessment(s) to use and, notably, set cut-scores that determine whether candidates will pass or fail. States that do not use the Praxis often employ a customized assessment produced by Pearson or the newer Pearson product titled the National Evaluation Series (NES). Publicly available mathematics items from such assessments suggest that they capture basic competencies as well as disciplinary (i.e., mathematical) sensibilities. To assess the latter, for instance, one NES practice item asks about the relationship between natural numbers, whole numbers, and integers, and another asks respondents to identify the linear function from a set of four equations (NES 2010 [Pearson Education, Inc. 2012]).
Teacher Quality and Quality Teaching

In general, there is little information regarding the relationship of initial teacher licensure exam scores to teachers’ eventual instructional quality. There is no record that any test in the Praxis Series has been validated against measures of teacher practice. More research has been done comparing licensure scores to student outcomes, but here the results are mixed. Some studies (Clotfelter et al. 2004; Ferguson 1991) demonstrate a modest relationship between licensure scores and student outcomes, but others do not (Buddin and Zamarro 2009). Older assessments, such as the National Teacher Exam (NTE), have received only marginally more attention. In a careful review, Wilson and Young (2005) note that of six studies examining relationships between teacher NTE scores and student and principal ratings, student outcomes, and other correlates, most found no relationship. They conclude: “We are inclined to believe that although the NTE may be correlated with success in university coursework or teacher candidates’ performance in those courses, it does not predict teacher effectiveness” (605).

In recent years, some scholars have been critical of the use of assessments for initial certification. Angrist and Guryon (2004) and Goldhaber (2007) both note that some individuals with poor test scores are nevertheless successful in producing student outcomes while others with high scores are not. As well, the procedure used to determine cut-scores has been subject to criticism. Cut-scores are typically established by a panel of experts, which uses information about expected item performance, expected examinee performance, or both in order to determine a passing score (Hambleton and Pitoniak 2006; Zieky and Perie 2006). Notably, most methodologies to establish cut-scores do not seek to ascertain the relationship between cut-scores and actual evidence of on-the-job performance. We also note that across states there is wide variance in cut-scores for most Praxis tests and, within states, variance over time as policy makers react to changing teacher labor markets.

Selecting Expert Teachers

Until recently, the selection of expert teachers was not a policy issue; teachers worked under a standard salary schedule with few opportunities for advancement without leaving the classroom. However, teacher development policies have recently focused on deploying expert teachers as coaches, mentors, specialists, and professional developers (e.g., Mangin and Stoelinga 2010; York-Barr and Duke 2004). Little information exists about how decisions are made to promote teachers into these positions; while many advocates suggest that teacher leaders should be carefully selected, the guidelines produced by these advocates list qualities that are subjective or for which there are few valid measures (such as content and pedagogical expertise and interpersonal skills).
Little (1990) suggested that the selection process for mentor teachers is quite varied and might include peer or supervisor recommendations, interviews, observations, simulations, or a portfolio process, and Steiner and Kowal (2007) noted that criteria for selection of instructional coaches are not consistent across states or even within districts.

Clearly, selection into leadership roles should be predicated on multiple criteria: content expertise, knowledge of how to communicate that content expertise to others, interpersonal skills, pedagogical skill, effort, and commitment. However, previous research has shown that in the area of content expertise, at least, selection criteria may not be fully executed. For instance, elementary school teachers who are in the mathematics leadership pipeline—individuals who provide mathematics professional development, mathematics-focused peer mentoring, or who serve on a mathematics committee—differ little from those who are not in terms of their mathematical knowledge for teaching (Hill 2010). We assume that most districts would prefer that mathematically knowledgeable teachers fill such positions; we explore the feasibility of a screening assessment that would provide information on this criterion below.

Summary

In summary, our review suggests that there is little evidence of disciplined inquiry into the validity of teacher scores on teacher certification assessments and that decisions to promote teachers into leadership roles appear to proceed on a mostly ad hoc basis. The investigations that exist—older research on the NTE and current research on the Praxis—suggest that the predictive power of multiple choice assessments is modest at best. We argue that, as a field, we need to understand whether multiple-choice assessments can substantially predict or correlate with performance in mathematics teaching. Such information is necessary in the case of certification, and it could be useful in the case of promotion into leadership roles. We argue, as well, that the field needs templates for investigating the validity of scores from written assessments; few investigations of this type have occurred to date.

To answer these challenges, we examine a newer multiple-choice assessment of teacher quality in mathematics—one that taps job-embedded mathematical knowledge and skills—in order to understand the extent of convergence between teacher performance on this assessment, teacher performance in the classroom, and student outcomes. Thus, this study is unusual in two ways: the test itself is more aligned with the skills and knowledge deployed by teachers in classrooms than typical certification assessments, and, unlike all teacher
certification assessments available today, we attempt to discern the validity of teacher scores on this assessment vis-à-vis other key criteria. Specifically, we ask the following:

(1) Can a multiple-choice mathematics assessment predict:
   (a) teacher performance in key domains of classroom practice?
   (b) student outcomes?

(2) What would be the effect of setting cut scores on the written assessment to make decisions regarding fitness for practice and potential for instructional leadership?
   (a) Can performance on the written assessment identify low-skill teachers? Can it accurately identify those who should not be admitted to practice or retained on the job? Might it wrongfully exclude competent teachers from practice?
   (b) Can performance on the written assessment identify high-skill teachers? Can it accurately identify those who should be recognized for outstanding performance in the practice domain? Might it wrongfully exclude any exceptional teachers from recognition as outstanding?

We investigate these questions below.

Method

Our general approach to answering these questions was to collect and compare three sources of information on teachers of mathematics: scores on our mathematics assessment; scores on an observational instrument that quantifies the mathematical quality of instruction; and, for a sample of middle school teachers, value-added scores produced from a state assessment of students’ mathematical knowledge. We asked whether scores on the multiple-choice assessment would provide information with which to make accurate decisions for initial certification/adequacy of practice as well as promotion into mathematics-specific leadership positions. We inquired about the viability of cutpoints on the MKT assessment, and we also developed two case studies of “marginal” teachers, both of whom had extreme written scores yet average instructional scores. These marginal cases allowed us to assess the extent of harm that would be done should scores on the mathematics assessment misidentify teachers for termination or promotion.
The sample for investigating the relationship between teacher knowledge and practice consists of 10 elementary teachers in two districts and 24 middle school teachers in a third district in the United States. Two districts enrolled students of lower socioeconomic status (SES) and one enrolled moderately affluent students. For the elementary study, teachers were recruited based on willingness to participate in the project and attendance at a professional development program. For the middle school study, we sampled schools that consistently produced low, medium, and high gains in student achievement and then successfully recruited 24 of the mathematics teachers in those schools (for more details, see Hill et al. [2011]). Thus, this is a convenience rather than a representative sample, although at both levels teachers represented the variation in the US teaching population well (Hill et al. 2008, 2011).

This is admittedly a small sample of in-service teachers, and thus this study is limited in several regards. Correlations in small samples are imprecise, and we cannot definitively estimate the magnitude of the relationships we seek to understand. Nevertheless, the small sample size imparts several advantages. First, we wanted multiple sources of in-depth data on each teacher in the sample in order to accurately characterize teachers’ knowledge and practice; inaccurate characterizations (e.g., low reliability) would leave us open to the possibility that any lack of relationship was due to measurement error. Yet accuracy is expensive; the high reliabilities described below were arrived at through extensive data collection—up to nine lessons per teacher, a lengthy survey, and roughly 3 hours of interviews. While more cursory data collection would have allowed a larger sample size, we argue that a carefully executed study is of equal value and provides information that other studies cannot.

Data were collected in the years 2003–4 (elementary) and 2008 (middle school) by administering surveys, videotaping classroom practice, conducting interviews, and gathering curriculum materials. Substantially similar video collection protocols were used in both the elementary and middle school samples, including the collection of videotaped lessons in groups of two or three over a period of several months to ensure diversity in mathematical content. For the elementary sample, nine videotaped lessons were collected per teacher in three waves; for the middle school sample, six lessons were collected per teacher in two waves. For more details on the data collection, please see Hill et al. (2008, 2011).
The written assessment used in this study was the Mathematical Knowledge for Teaching (MKT) instrument created by the Learning Mathematics for Teaching project at the University of Michigan (Hill et al. 2005). The project’s literature and publications describe the specific forms of the instrument used here as encompassing two major domains of mathematical knowledge for teaching: common content knowledge and specialized knowledge for teaching. The former is thought to be mathematical knowledge used across professions, while the latter is mathematical knowledge specific to the work of teaching (Ball et al. 2008). The MKT has been shown to positively predict student outcomes (Hill et al. 2005; Rockoff et al. 2008). A previous study examined the relationship between the MKT and specific elements of the mathematical quality of instruction (e.g., teacher errors, richness of the mathematics) in the sample of 10 elementary teachers also examined here and found moderate to strong \((r = .30–.80)\) correlations. This article expands upon that analysis by inspecting this relationship in middle school, working with an additional 25 cases, comparing MKT scores to teacher value-added scores, and taking a different perspective. Whereas in the previous work on the MKT and mathematics teaching the authors examined the extent of the relationship in order to generalize to theory about teaching and teacher preparation, here we test the limits of the relationship for making job-related decisions.

We used this assessment, rather than a commercially available certification exam, for several reasons. First, only partial and retired certification exam forms are officially available from companies like ETS and Pearson, and the scoring procedures used by these companies, in particular conversions from raw score to scale scores, are not publicly available. Second, the MKT and certification assessments are quite similar in format. An examination of released items from the Praxis and Pearson revealed that both assessments rely heavily or exclusively on multiple-choice items, all involve mainly content and problem contexts relevant to teaching, all are scaled using Item Response Theory (IRT), and reliabilities of the certification assessments and MKT forms used here are similar, at about .95.

However, our inspection of released items from commercial exams suggests that they only inconsistently capture knowledge that is “specialized” to teaching mathematics (see also Hill et al. 2007). We expected, therefore, that the MKT would have greater predictive power for classroom instruction and student outcomes given the alignment between the assessment and practice. This means that the analysis presented here may overstate the relationships as compared to the relationships one would find using a typical certification exam instead of the MKT. Given our interest in whether multiple-choice exams can predict qualities of classroom practice, this is not necessarily fatal...
to our inquiry; however, results cannot be generalized to all current certification exams.

In both the elementary and middle school samples, teachers completed an MKT survey appropriate to their grade span (K–5, 6–8). The middle school MKT was measured by 159 items on number, operations, and proportional reasoning. We obtained an estimate of teachers’ written MKT scores by entering their survey responses into an IRT model and scoring this data alongside data obtained as part of a nationally representative sample of middle school teachers (see Hill 2007). Two-parameter IRT models were used to score teacher responses, with a resulting reliability of .97. For this study, teachers’ MKT scores are expressed as percentile ranks as compared with the larger nationally representative sample.

The elementary MKT survey contained 77 problems that were relatively evenly divided across number and operations, geometry, and algebra. These 10 teachers’ pencil-and-paper responses were scored with a much larger data set (n = 636) that included teachers who had signed up for professional development but had not attended, teachers who attended math professional development, and others. Percentile rankings for these 10 teachers are drawn from this larger data set. This larger sample was nonrandom and not representative of the general population of elementary teachers; however, it does place the current study’s elementary teachers in a wider context. We also know from comparing data from the nonrandomly sampled group to a subsequent random sample of teachers that the nonrandomly sampled group outscored the random sample by roughly 3.6% over a set of identical items. Thus the differences between the representative sample and the one used here for norming were, on average, not large. The IRT reliability of the elementary MKT measure is .94.

Procedure for Measuring the Quality of Instruction

To examine classroom practice, we used an observational instrument designed as a companion to the MKT assessment, titled the Mathematical Quality of Instruction (MQI; Learning Mathematics for Teaching Project 2011). The MQI instrument provides separate scores for several aspects of mathematics instruction: the richness of the mathematics present in the class (e.g., explanations, representations, explicitness about language and reasoning), the mathematical errors made by the teacher, the teacher’s ability to provide productive responses to students’ mathematical ideas, and student participation in meaning-making and reasoning.

Scoring occurred by parsing videotapes into roughly 5-minute segments according to breaks in instruction, then having raters assign each segment a
value for each item. Two raters viewed and scored each lesson and then reconciled their scores to agreement. The 19 raters for this effort were randomly assigned to one another and to lessons, and they viewed multiple lessons from the same teacher only by chance. Interrater agreement of more than 80% was obtained before scoring proceeded, and a generalizability study showed that teacher-level reliability for dimension-level items ranged from between .63 and .91 (see table 1). These dimension-level items used with the elementary and middle school samples were virtually identical.

In this article, we also focus on two more general ratings. One is the overall assessment of each lesson's MQI. This three-point Likert-style rating was applied by observers to the whole lesson, with “low” corresponding to lessons with significant teacher mathematical errors, “medium” corresponding to lessons with fewer such errors yet mostly routine instruction, and “high” reserved for lessons with no errors as well as significant mathematical richness through explanation, representations, and strong interactions with students. The generalizability study showed that the teacher-level reliability of the MQI rating is .90 (six lessons) or .93 (nine lessons).

The other measure is raters’ lesson-based evaluation of teachers’ mathematical knowledge for teaching (lesson-based MKT), also scored on a three-point Likert scale and using the same procedure outlined above. This estimate differed from overall MQI in that observers could consider evidence that the teacher was mathematically knowledgeable beyond what was observed in the majority of the lesson. An example would be a lesson that was mostly focused on procedural practice but during which the teacher briefly demonstrated remarkable command of both mathematical explanations and student thinking about mathematics. The lesson would be assigned a high lesson-based MKT score and an average MQI score. This rating was only available for the 25 teachers in the middle school sample. Although this measure was not included in the generalizability study, interrater agreement of 80% was reached before coding proceeded.

We note that both our written measure of teachers’ MKT and our observational...
measure of their teaching practice were highly reliable. Standard errors for teacher MKT scores, for instance, were typically under 0.20 standard deviations. This level of accuracy is critically important; if an unreliable measure fails to predict teaching performance, the failure could be attributed simply to measurement error. We return to this point in the conclusion.

**Measuring Teacher Value-Added Scores**

As described in Hill et al. (2011), we constructed value-added scores for all mathematics teachers in the middle school district with complete data for nine or more students ($n = 222$) and then extracted the studied teachers for whom we had student outcomes data ($n = 24$) from the resulting data set. This value-added analysis relied on 2 years of state test scores (2006–7 and 2007–8) for students. The state test, the Standardized Mathematics Assessment (SMA) was a mix of constructed response and multiple-choice items. An inspection of released items shows they assess a mix of procedural and conceptual knowledge; the latter help align the student assessment with the MKT assessment described earlier. Form and interrater reliabilities range between .85–.90 (Pearson Psychometric Research Services 2008), and vertical equating between adjacent-year tests implied student gain scores can be calculated from adjacent-year test data. Students were excluded from these models if they spent less than a full year at the same school, had evidence of switching mathematics teachers midyear, or had missing values for the previous years’ standardized math, reading, or science scores. Because there is significant disagreement in the literature regarding the most appropriate specification of value-added models (see McCaffrey et al. 2003), eight different models were initially fit to the data. All models explored were covariate adjustment models; data were not available to construct longitudinal models. Scores generated from models using teacher fixed effects or empirical Bayes’s estimates of teacher effects were correlated strongly ($r = .98–.99$), and thus we discuss only teacher empirical Bayes’s estimates here. Because models with student background variables and school fixed effects resulted in different rankings, we present one from each class.

Model 1: simple model.—Adjusting for prior student scores, teacher random effect,

$$y_{ijkl} = \beta_0 + \beta x_{ijkl} + \tau_j + \xi_{ijkl} + \varepsilon_{ijkl},$$

where $y_{ijkl}$ is the SMA scale score for the $i$th student in the $j$th classroom of the $k$th teacher at the $l$th school in 2007–8 and $x_{ijkl}$ is a vector consisting of the $ijkl$’s student’s previous year’s SMA scale score (SMA06), SMA06 squared, SMA06 cubed, grade indicator variables, previous year’s standardized reading scale score (SRA06), previous year’s standardized science scale score (SSA06),
and SMA06, SRA06, and SSA06 by grade indicators interaction terms; \( \beta \) is a parameter vector. The \( \tau_{ij} \sim N(0, \sigma_{\tau}^2) \) are teacher random effects and \( \xi_{ijkl} \) are classroom- and student-level error terms assumed to be independent, where

\[
e_{ijkl} \sim N(0, \sigma_{e}^2),
\]

\[
\xi_{ijkl} \sim N(0, \sigma_{\xi}^2).
\]

**Model 2: school fixed effects model.**—Adjusting for prior student scores, teacher random effect (as with the simple model) plus a school fixed effect,

\[
y_{ijkl} = \beta_0 + \beta x_{ijkl} + \tau_{ij} + \phi c_i + \xi_{ijkl} + e_{ijkl},
\]

where \( c_i \) is a vector of indicator variables for each school and the \( \phi \) are the parameters indicating the school fixed effects.

**Model 3: student background adjusted model.**—Adjusting for prior student scores and a teacher random effect (as with the simple model) plus student background variables,

\[
y_{ijkl} = \beta_0 + \beta x_{ijkl} + \tau_{ij} + \lambda s_{ijkl} + \xi_{ijkl} + e_{ijkl},
\]

where \( s_{ijkl} \) is a vector of student covariates including indicator variables for accelerated/enriched, algebra, free or reduced priced lunch (FRL), English language learner (ELL), special education (SPED), Spanish test language (STL), indicator variables for each ethnicity, and a SPED by SMA06 interaction; \( \lambda \) is a parameter vector. We related value-added scores to the MQI and lesson-based MKT measures described above using a Spearman rank-order correlation.

**Setting Cut-Scores**

To simulate the use of a written assessment to screen teachers for specific purposes, we selected cut-points on the MKT. As described above, this process is typically done by state education officials, who make judgments based on how specific categories of teacher candidates (i.e., a “marginal candidate”) would perform on specific items. Because this project had available actual measures that categorized teachers as having high- or low-quality instruction, we asked if the distributions of teacher and teaching measures suggested specific cut-points. A review of lesson-based MKT and MQI scores for these data suggested that approximately nine of the 35 teachers (26%) had scores below 1.5, meaning that half or more of their lessons were judged to have significant teacher errors or unclear presentations of the mathematics or to lack a mathematical point. We also observed that roughly seven of the 35 teachers (20%) were rated as very strong, with overall lesson scores of 2.5 or above. This
suggested that roughly 25% of the teachers should have been excluded from teaching, while another 20% would make good candidates for leadership positions based on the mathematical quality of their teaching. We also plotted MKT and MQI against one another to determine whether there were any natural cut-points in the data, that is, whether teacher performance improved dramatically (or at least became more variable) at a certain percentile from the bottom. This analysis suggested roughly the same cut-points, and thus we set the lower-bound cut-score at the 25th percentile and the upper-bound cut-score at the 75th percentile to facilitate comparison.

**Development of Case Studies**

Once we had conducted the quantitative analysis, we chose two cases for further investigation. Case study selection was premised on the belief that by investigating the most extreme teachers in detail—that is, a teacher identified as having low MKT scores but with the most acceptable MQI scores and a teacher with high MKT scores but the lowest MQI scores—we could critically examine the assumption that the use of MKT scores would lead to proper classification of teachers. To do so, we qualitatively analyzed each case study teacher’s instruction, developing a sense for both strengths and weaknesses and an overall judgment about the accuracy of the MKT assessment in placing that teacher. We also used interview transcripts, including those from two 60-minute interviews that asked teachers to describe their practice and their views of mathematics, mathematics teaching, and the students they currently teach. These interviews also included a cognitive tracing interview similar to those used in Hill et al. (2008) to illuminate teachers’ thinking about the MKT items contained on the survey.

**Results**

We begin by discussing the overall predictive value of the MKT to practice and student outcomes, then move to a discussion of using hypothetical cut-points on the MKT to identify teachers for specific purposes. To explore these cut-points further, we construct two case studies of extreme teachers—teachers whose MKT scores placed them in one category but whose observational scores suggested they might belong in another.
Table 2

Correlations between Written Mathematical Knowledge for Teaching (MKT) Instrument Scores and Observational Scores for Elementary and Middle School Teachers

<table>
<thead>
<tr>
<th>Written MKT</th>
<th>.65**</th>
<th>.58**</th>
<th>.28</th>
<th>-.50**</th>
<th>.53**</th>
<th>.20</th>
<th>-.22</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson MKT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesson MQI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—Pearson correlations are reported. N = 34 teachers, except for Lesson MQI and Student Participation, which were only scored for the middle school sample.

* p < .05 (two-tailed test).

** p < .01 (two-tailed test).

Table 3

Correlations between Measures of Teacher Knowledge and Practice and Value-Added Scores

<table>
<thead>
<tr>
<th>Written MKT</th>
<th>.58**</th>
<th>.72**</th>
<th>.41*</th>
<th>.51*</th>
<th>.25</th>
<th>.27</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lesson MQI</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lesson MKT</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note.—MKT = Mathematical Knowledge for Teaching. MQI = Mathematical Quality of Instruction. Pearson correlations are reported. N = 24.

* p < .05 (two-tailed test).

** p < .01 (two-tailed test).

Overall Measure Performance

The correlation between the MKT written assessment and the assessment of practice domains is shown in Table 2. The written MKT was strongly related to our lesson-based guess of MKT, at 0.65 (n = 35, p < .01) and moderately related to lesson MQI, at 0.58 (n = 24, p < .01). This suggests that, as the instrument developers had designed, the lesson-based guess at MKT enabled raters to ignore lesson-specific “error” and focus on the anticipated knowledge.
level of teachers, which we suspect better generalized to the teachers’ other lessons. We also found that the written MKT was a moderate predictor of middle school teachers’ value-added scores, at 0.25–0.51, depending upon the model (table 3).

In terms of specific dimensions from our MQI instrument, we found that the written MKT scores were only related significantly to two dimensions: teacher mathematical errors \((n = 35, r = -0.50, p < 0.01)\) and productive teacher responses to students’ mathematical ideas \((n = 35, r = 0.53, p < 0.05)\). The negative coefficient on errors demonstrates that higher-scoring MKT teachers have on average fewer errors in the classroom than lower-scoring teachers. Three others dimensions, student participation in meaning-making and reasoning, classroom work is not connected to mathematics, and richness of the mathematics, were signed in the anticipated direction but were not significant.

**Identifying Low-Skill Teachers**

Next we turn to the question of whether the written MKT assessment would correctly identify low-skill teachers. Recall that this question has two components: Would it wrongfully exclude competent teachers from practice? Would it accurately identify all those who should not be admitted to practice?

Figures 1–3 suggest that, as related to most of our observational measures, including overall MQI, lesson-based MKT, and responding to students and teacher errors, the use of the written MKT assessment to make hiring and retention decisions would not wrongfully exclude teachers from the profession. Teachers below the 25th percentile in MKT, in other words, were rated low on the overall observational metrics, were prone to making mathematical errors, and were also not likely to use student ideas in instruction either well or often. One exception was Bianca, who scored in the 1st percentile on MKT but who had an error score of less than .10. A review of her lessons, however, suggests that her instruction proved problematic in ways likely reflected in her low overall scores (1.80 for both lesson-based MKT and MQI) but not on the errors dimension. Another exception is Wilhelmina, a teacher with a 7th percentile MKT score and an average MQI score of 2.0. We investigate her instruction in more depth below.

However, figures 1–3 suggest that using the MKT as a screening tool would not exclude all potentially problematic teachers from practice. Specifically, a significant portion of teachers in the 25th to 75th MKT percentiles—roughly six to eight teachers in each panel—scored in the same MQI/lesson-MKT range as teachers who would be excluded from teaching. Some teachers had significantly troublesome observational scores. Four elementary teachers in
Fig. 1.—MKT percentile versus MQI for 25 middle school teachers.

Fig. 2.—MKT percentile versus overall lesson score for 10 elementary and 25 middle school teachers.
FIG. 3.—MKT percentile versus four dimensions of instruction for 10 elementary and 25 middle school teachers.
Teacher Quality and Quality Teaching

this range, for instance, had error scores in the same problematic range as our low-MKT teachers (fig. 2). Five middle school teachers in this range had overall MQI scores of less than 1.5 (fig. 1), suggesting that at least half their lessons contained significant mathematical problems or were off-track.

Figure 3 also suggests that for some of the MQI dimensions, notably richness of the mathematics and student participation in meaning-making and reasoning, lower-MKT teachers were not significantly different from higher-MKT teachers. This suggests that lower-MKT teachers emphasize meaning and pose challenging tasks and questions for students at roughly the same rate as high-MKT teachers. This finding is not dissimilar from those of others who have investigated the relationship of MKT to teacher practice (e.g., Stein and Kaufman 2010) and may result from several factors. Some evidence from our work with the broader data set suggests that curriculum materials (and lesson-focused professional development) can to some degree scaffold low-MKT teachers successfully into these practices (see Charalambous and Hill, forthcoming). We also conjecture that engagement in these activities is a matter of dispositions toward teaching as well as knowledge and skill—that is, low-MKT teachers may engage in these practices because of their views on mathematics and student learning. We also note, however, that the upper-right panel of figure 3 shows that low-MKT teachers were engaging in these activities with significant amounts of mathematical error and imprecision. As one of the case studies below illustrates, student participation and richness in a low-MKT classroom may occur but with significant problems in the presentation of the mathematics.

Finally, the plot of MKT and student value-added scores (fig. 4) tells much the same story. Low-MKT teachers tended to have student value-added scores below the district mean, with only one or two exceptions, depending upon the model. A handful of teachers in the 25th to 75th MKT percentile had similarly low scores. While using the district mean is perhaps a lenient criterion, and our data clearly skew toward high value-added teachers, these are optimistic findings. Given the difficulties in making causal inferences based on value-added scores (Raudenbush 2004), the possibility of test preparation behaviors artificially inflating scores (Koretz 2008), and the amount of noise and potential bias in these scores (McCaffrey et al. 2003), finding that exclusion decisions based on MKT and value-added scores would only conflict for one teacher in the first two models is encouraging.

These findings show that in this sample low-MKT teachers had generally problematic practice, from the point of view of being prone to teacher mathematical errors, having little evidence of responding productively to students’ mathematical ideas, and generally producing low student value-added scores. A qualitative analysis of these low-MKT teachers found several lesson segments in which teachers solved problems or defined terms incorrectly and many

506 American Journal of Education
FIG. 4.—MKT percentile versus VAM scores from three different models for 24 middle school teachers.
more in which teachers provided inaccurate or incomplete mathematical explanations, used imprecise mathematical language, rejected correct student solution methods, and provided confusing or otherwise inadequate presentations of content. In many cases, raters’ notes indicated that they felt the teacher did not have a strong conceptual knowledge of the lesson content. While the MKT assessment would not exclude all such teachers with problematic practice, this evidence suggests that it would not unfairly exclude teachers. Below, we use a case study of one low-MKT teacher who appeared at the margin of the observational measures (Wilhelmina) to examine this proposition in more detail.

**Identifying Exemplary Teachers**

Our next question asks whether we can use the MKT assessment to identify teachers with strong mathematics instruction. This question again has two components: whether teachers with upper-quartile MKT scores have above-average instruction and whether MKT scores identify all teachers with similarly strong instruction. Figures 1–3 show that upper-quartile MKT teachers do tend to have strong instruction in the sense that there are few errors, strong use of student ideas, and strong overall ratings of MQI and lesson-based MKT. Qualitative analyses suggest that these teachers tended to focus instruction on mathematical meaning, identified and emphasized key mathematical points and connections for their students, used mathematical language accurately, and pressed their students for the skillful use of mathematical language. The same pattern occurs for student participation in the development of the mathematics, where upper-quartile MKT teachers were rated as having consistently high student participation in novel tasks, questioning, and explanations. Once again, the scatterplot of richness versus MKT shows more dispersion in the sense that high-MKT teachers exhibited differing levels of mathematical explanation and practices in their teaching.

Figures 1–3 also demonstrate that not all teachers with exemplary observed practice scored well on the MKT assessment. Notably, two teachers who were assigned a high lesson MKT and one who was assigned a high lesson MQI (2.5) scored toward the middle of the distribution on the written assessment. An inspection of value-added scores also suggests that upper-quartile MKT teachers scored above average in student value added, with no exceptions, and all above-average MKT teachers also scored strongly on the student outcome measure, regardless of model and with only a few exceptions.

These findings show that, in this sample, a strong MKT score ensured low-error instruction, a strong ranking for responding to students, and generally high-quality instruction. Student outcomes were also very strong. Whether all
teachers in this range could serve equally well as coaches or instructional leaders is an issue we take up below in the examination of a borderline teacher.

Figures 1–3, taken together, suggest that the lower-bound cut-score cannot be moved to improve classification accuracy. While additional teachers might be identified as lacking by moving the cut-score for one measure, others may be wrongly excluded by any movement, and in some cases, the data themselves suggest that 25% is a nearly optimal threshold for this sample (e.g., see errors). Moving the higher-bound MKT cut-score to the right would correctly exclude more teachers with more average instructional scores from leadership positions, suggesting that if policy makers wished to be conservative, a higher cut-score would be appropriate.

**Borderline Cases: Does Qualitative Analysis Contravene These Findings?**

We selected one teacher from each of the lowest and highest quadrants who would be likely to challenge our conclusions regarding potential policy uses of this multiple-choice assessment. We profile the teaching of Wilhelmina, a sixth-grade teacher who had roughly average lesson-based MKT and MQI scores, at 1.83 and 2.00, but was at the 7th percentile in MKT. At issue with Wilhelmina was whether using the MKT assessment with a 25th percentile cut-score might inappropriately screen her out of the teaching profession. We also profile Marie, also a sixth-grade teacher, who also had roughly average lesson-based MKT and MQI scores but was at the 89th percentile in MKT. At issue with Marie is whether her strong MKT score would falsely indicate that she was an appropriate candidate for a mathematics leadership position at her school or in her district.

**Wilhelmina.**—In reviewing Wilhelmina’s lessons, we saw a teacher with a strong command of her classroom and positive instructional “moves” but who delivered a superficial treatment of mathematics. Her MQI score reflected a high level of student participation in the mathematics, yet her imprecision with language, incomplete explanations of mathematical ideas, and overreliance on procedural approaches suggested that Wilhelmina’s MKT score and teaching quality were not completely divergent. We judged Wilhelmina to be a borderline case, and we discuss the implications of this designation below.

Wilhelmina’s lessons featured significant student engagement. For example, in a lesson on multiplying fractions, she elicited students’ prior experiences with the topic, then encouraged students to develop their own ideas about how the multiplication of fractions works rather than simply presenting the algorithm. In another lesson, she pushed students to provide explanations for the models of fraction multiplication they had generated. In an introductory lesson on two-dimensional geometry, Wilhelmina had students examine several
polygons and nonpolygons and pressed them to use these figures to articulate the key characteristics of a polygon.

Yet a careful review of Wilhelmina’s lessons also revealed repeated instances of imprecise use of language that indicated potential conceptual misunderstandings. An example comes from a lesson on multiplication of fractions. Her textbook defined multiplication of fractions in the following way: “½ of ¼ . . . mean[s] the operation of multiplication, or ½ × ¼.” During the lesson, she repeatedly stated that “of means multiply,” yet used this mantra in both appropriate and inappropriate ways. For example, she repeated multiple times that the wording “¾ pounds of fudge” signified multiplication when in fact it signified an amount of fudge. In another problem, she presented a situation in which someone has ¾ of a sandwich, cuts it into three pieces and eats two out of three pieces. Here, the problem situation would be represented as “⅔ of ¾,” indicating that it can be correctly represented by ⅔ × ¾. However, Wilhelmina suggested that “three-fourths of a sandwich” indicates that this problem requires multiplication and then wrote “¾ of” on the board; she then emphasized the word when she said “2 out of 3,” again implying it indicates multiplication, and she wrote “¾ of ⅔.” While ⅔ × ¾ and ¾ × ⅔ are computationally equivalent, only the first correctly models the situation presented in the story problem. In her teaching, Wilhelmina inappropriately used key words to justify her choice of a mathematical operation rather than relying on the structure of the context.

Another common feature of Wilhelmina’s instruction was a reliance on procedural algorithms without much attention to explanations and meaning. For example, later in the lesson on multiplication of fractions, Wilhelmina reviewed a problem from the textbook: “Linda is making bows to put on a wreath. Each bow uses 2⅓ yards of ribbon. A spool of ribbon contains 15 yards of ribbon. Choose the number of whole bows she can make from one spool. (F) 6 (G) 7 (H) 12 (J) 35.” Although this problem was part of a unit on multiplication, it can be represented by division as 15 ÷ 2⅓. Wilhelmina began discussing the problem by telling students that she would accept an answer of 35, which she found by multiplying 2⅓ and 15, because she herself had initially chosen that answer. She then said, “I think they wanted us to add. But if we do it, if we say 2⅓ + 2⅓, what do I get? Anybody remember, common denominators? I already have common denominators. So what’s 2 plus 2?”

Wilhelmina went on to instruct the students to repeatedly add 2⅓ until there was not enough ribbon remaining for more bows. The bulk of this discussion revolved around the computations. While Wilhelmina’s approach was computationally sound, she missed significant opportunities to develop key mathematical ideas related to the lesson. First, even after she noticed that her original answer was incorrect, she assured students that she would accept...
a method and answer that was incorrect simply because she had made that same mistake. More importantly, however, she never explained why the problem can be solved with addition or how this approach relates to multiplication or division. We argue that this sends a message to students that thinking about the structure of the problem and choosing an appropriate method was not important. This focus was representative of a large portion of her teaching; while she frequently asked students to explain how they solved problems, she almost always elicited or provided a list of procedural steps without an explanation for why a particular operation was appropriate.

Wilhelmina’s written MKT assessment and cognitive tracing interview supported this observation in that they revealed she was most comfortable with taking a procedural approach to problems. For example, when choosing between a correct nonstandard approach to a problem and one describing the mechanical steps, she almost always preferred the steps, stating that it was the one that she felt most comfortable with.

In summary, Wilhelmina is a borderline case. While we do not judge her instruction to be as problematic as that of other teachers in our sample, in the sense that her students were not learning incorrect mathematical procedures or enduring incoherent lessons, she had significant gaps in her understanding of middle school mathematics. She made frequent small mistakes and was often imprecise in her mathematical language and explanations. Whether a state would wish to employ her depends upon its view of the teacher corps and the purpose of the certification exam. Our qualitative analysis suggests she worked hard at and cared about teaching, had very good classroom management skills, demonstrated procedural fluency in the mathematics she teaches, and had some elements of good practice in place; in other words, she had potential. However, she did not yet meet a minimum standard for accurate delivery of content. If the purpose of the exam is to exclude only the most problematic of teachers, perhaps only the ones whose instruction suggests such weak teaching knowledge and skill that no amount of professional development would aid, then her MKT score would incorrectly exclude her from teaching. If the purpose of the exam is to exclude teachers who are not currently ready for a moderate to high level of practice, however, the assessment would not incorrectly exclude Wilhelmina.

Marie.—Examining Marie’s case allowed us to ask a different question: if this exam were used to select teachers into leadership, coaching, and mathematics specialist positions, would it be incorrect in this particular case? Marie scored very strongly on the written MKT assessment (89th percentile), but she was rated by observers as having only average MQI.

In reviewing Marie’s lessons, we saw a teacher with strong mathematical content knowledge, precise use of mathematical language, and the ability to communicate that knowledge clearly to her students. While she consistently
encouraged students to explain their thinking, her teaching featured a strong emphasis on procedural mastery, reliance on shortcuts and mnemonics, and an emphasis on getting quickly to the right answer. Thus, we concluded that while Marie’s MKT was reflective of her strong content knowledge, her lessons did not show the amount of richness or student participation in mathematical reasoning that districts might expect from a mentor teacher, coach, or other content leader.

Throughout her lessons, Marie demonstrated a strong knowledge of the mathematics she was teaching in the sense that she rarely made errors and appeared to have a solid command of the mathematics she taught. Marie also demonstrated a strong use of mathematical language in her classroom. For example, in a review of homework covering two-dimensional geometry, she insisted on precise geometric definitions and pushed students to refine their answers using the appropriate mathematical terms. Much of one of the lessons was focused on the terminology for geometric transformations, and she had students create a “foldable” to use for the definitions, breaking down each part of each definition to make important distinctions, for example, between change in orientation and change in location.

Marie frequently requested students to explain their work—asking in particular how they approached and solved problems. However, she emphasized explanations as they pertained to specific tasks and seldom generalized those explanations to key mathematical ideas of the content under consideration. This may be related to another aspect of Marie’s instruction—an emphasis on mnemonics to remember mathematical procedures rather than a habit of unpacking and explaining the reason why those procedures work. For example, in a lesson in which students were reviewing dividing fractions, she reminded students, “What do we remember? Copy, switch, flip.” She later added, “I call it invert and multiply.” Rather than developing students’ conceptual understanding of potentially difficult concepts, Marie’s instruction privileged “tricks,” which were often idiosyncratic, to help students overcome difficulties.

In Marie’s classes, we observed students acknowledging and practicing the given rules, but we did not observe students requesting explanations for the mathematics behind the procedures. Students in the observed lessons seldom explained why their procedures worked or engaged in comparisons of multiple solution methods. When students made an error in a procedure, Marie often addressed the individual student one-on-one rather than using the error as a teaching tool for the class. She also tended to address student errors by having the students identify where in the procedure they went wrong, but she made no attempt to unearth conceptual misunderstanding. During her interview, Marie called herself “a devotee of Saxon math,” and our observations of her teaching style suggested she would be better matched to that curriculum than the one that she was using.
In summary, Marie had a high level of content knowledge, as evidenced by her MKT score, but she did not deploy it to provide her students with explanations and meaning for the procedures and problems they studied. In some schools and districts, Marie’s view of teaching mathematics would be aligned with the dominant paradigm, and she would serve as an excellent mentor for new teachers. In other schools and districts, Marie’s instruction would be at odds with prevailing philosophies regarding mathematics instruction. The validity of the decision to nominate her nomination to a leadership role, then, would depend at least in part upon the match between her ideas about student learning and those of the school or district in which she taught.

Summary of Case Studies

In looking at Wilhelmina and Marie’s lessons, it became apparent that an MKT score was in some cases not sufficient to predict the quality of a teacher’s instruction. In both cases, the teachers’ procedurally-oriented approach to teaching affected the quality of their instruction. In Wilhelmina’s case, a reliance on procedures might have actually served to improve her observational scores; without delving deeply into concepts, she avoided some of the problems that arise when teachers with marginal content knowledge are challenged to teach conceptual material to students. This is not an unusual pattern; other work (e.g., Hill et al. 2008) has described the case of a teacher with a low MKT but acceptable procedural instruction. By contrast, Marie’s focus on procedures in the absence of developing mathematical explanations and meaning capped her observational scores. In tandem, these cases suggest that, in addition to teachers’ mathematical knowledge, measures of their orientation toward the teaching of mathematics would be an appropriate factor to consider when selecting teachers for coaching or mentorship roles.

Discussion and Conclusion

This study has clear limitations. We chose one specific written mathematical assessment for teachers out of the many we could have selected to investigate. This decision has two tangible results. First, we did not investigate the actual relationship of a real certification assessment to practice, which limits the generalizability of our findings. Moreover, the use of the MKT assessment may overstate the relationship of test scores to teacher practice because the MKT assessment was built directly from a grounded study of practice (Ball and Hill 2008) rather than from the lists of competencies more frequently used by large testing firms. That the MKT was possibly closer to actual practice
and our predictive power was still imprecise in several areas suggests to us that it is not likely that commercially available assessments would do better in terms of identifying teacher candidates who are unprepared to enter teaching.

Another limitation of this study is its sample and sample size. We could not recruit teachers who did not pass state certification, and our in-depth examination of teachers’ practice necessitated a small sample as the demands of collecting, coding, and analyzing videotaped data are heavy. Nevertheless, this report is a first attempt to examine how teacher scores on a test that resembles those used in teacher certification generalizes to practice. The fact that this work has not been done to date with modern multiple-choice teacher certification exams is a significant problem for the profession of teaching and for those who seek to understand and shape teacher labor markets. If certification exams are to act as professional gatekeepers (Abbott 1988), they must accurately and effectively prevent ill-prepared candidates from practicing. Because certification exams affect labor markets, knowing what percentage of candidates to exclude (and therefore replace) is critical.

One criterion for evaluating the success of our hypothetical exam was based on the assumption that states would prefer a teacher exam that selects out only candidates whose teaching does not meet a minimum criterion for competence while at the same time identifying all teachers within that pool. The MKT met the first of these conditions but not the second. Given the number of mathematical errors they made, excluding Wilhelmina and other low-MKT teachers from teaching would be justifiable for many based on disciplinary grounds. Furthermore, low-MKT teachers evidenced weak and often significantly problematic practice; their student outcomes fell below the district average, perhaps as a result. Just as importantly, using the MKT assessment as a filter for entering the teaching profession would not exclude a strongly promising teacher from practice.

However, the MKT assessment failed to identify all teachers with such significantly weak practice. Several of the elementary teachers in particular had moderate MKT (25th to 75th percentile) but extremely weak lessons on average; several middle school teachers evidenced similar patterns. We doubt that this misplacement was due to measurement error itself as our measure reliabilities were high and their corresponding standard errors were low. Instead, we view this as a problem of generalization of the MKT assessment to practice in the sense that the assessment does not filter out all aspects of problematic practice. This would suggest that the assessment should be revised to capture the types of difficulties we observed in the moderate MKT-teachers classrooms. In the shorter term, the assessment could be used as is if prospective teachers with scores lower than the 25th percentile were to be excluded from teaching and
if teachers in the next two quartiles were closely monitored at entry to practice, particularly if their students’ outcomes fell below the district average.

At the other end of the spectrum, the MKT assessment performed relatively well, identifying most but not all teachers with strong observational scores. If the MKT was used as a certification exam, policy makers might use its scores to identify beginning teachers who require less mathematics-specific supervision (although still, surely, supervision in other areas). If used as a screen for leadership positions, our results suggest that the instrument might be used effectively in conjunction with other measures, such as attitudes about mathematics instruction. In this scenario, a written assessment would be administered to candidates for such positions as part of the preliminary screening process, similar to the way in which specialist exams are used in medicine or accounting. Although professional advancement decisions are lower-stakes, having relatively inexpensive and efficient means of identifying or qualifying such teachers could help ensure the integrity of induction and coaching efforts.

This all said, these results do suggest that multiple-choice exams can be predictive of classroom practice and student performance. Future work should address whether this relationship is as strong when we compare preservice teachers’ scores on an MKT-like certification exam with their later practice and student outcomes. If there remains a significant relationship, this would have several implications for the measurement of teacher and teaching quality. We discussed two in the introduction to this article: developing certification assessments that better demarcate well-prepared and poorly prepared individuals and ensuring that the promotion of individuals into leadership roles is done with the assurance of at least adequate subject matter knowledge. In our view, this is an improvement over the currently operating system. Additional improvements may be realized by combining an MKT-like assessment with additional sources of information, including observations of teaching itself.

Our results are also suggestive with regard to current passing rates on most state certification exams. Whereas many reports suggest that few teachers fail the state mathematics exams, our observational analysis suggested that as many as 25% of the sample should not have gained entry to teaching. We base this assertion on teacher performance on the items that capture errors, imprecision in language, and lack of clarity; in our view, this is the most clear-cut and defensible criterion in setting minimal standards for teaching quality. Few find it acceptable when teachers make many uncorrected errors in front of students, and yet several teachers in our sample did so on a regular basis.

Our results also suggest that although teaching proficiency is multidimensional, that multidimensionality is unlikely to be captured by a single multiple-choice certification assessment. The clearest example is Marie’s case, where a predisposition toward seeing mathematics as a set of rules and procedures resulted in error-free but procedural lessons. While this would be acceptable
in some schools or districts, it would not be in others; however, this is a dimension not detected by her MKT score. We also saw that teachers may be strong on some dimensions but weak on others; for instance, Wilhelmina did involve students in mathematical reasoning and meaning-making, but her procedures ultimately led to superficial understandings of ideas and procedures. Given that the MKT did not reflect the strengths and weaknesses of her practice, it is unlikely that conventional certification assessments would either.

The imprecision in the predictive value of the MKT results for the middle of the spectrum, along with the above-noted multidimensionality and worrisome treatment of mathematical content by some teachers, suggest that the current licensure arrangements should be augmented by intensive and rigorous pretenure observations. We argue that these observations could examine elements of teaching not easily gauged by multiple-choice assessments and could also be used as an assurance of high-quality mathematical content, if not fully developed pedagogical skills. Methods for collecting remote observations on a wide scale have been pioneered by private firms such as Teachscape and thereNow; for other instruments, such as Framework for Teaching (Danielson Group 2011), training for live observers is available.

These results also bear suggestions for the measurement of teacher competency at different stages of the career ladder. If these results were to hold up in large samples of first-year teachers, it would suggest that worries about certification assessments may be overstated. Goldhaber (2007), for instance, may have found substantial numbers of failed-certification/high value-added teachers simply because the value-added scores are so imprecise. Studies that also include measures of actual teaching practice in addition to examining the relationship between certification exams and value added measures are needed to sort this out.

Clearly, studies of this sort that investigate in-use teacher examinations in various states are needed. Evidence from this project suggests that such studies are not impossible to mount, and many of our study limitations suggest ways in which future studies might be improved. Reductions in the cost of videotaping teaching and progress in improving the efficiency of coding schemes should further make such studies possible. If teacher certification exams are going to survive critiques and continue as tools of public policy, they must be subject to this form of scrutiny.

Note

This research was funded by NSF grants REC-0207649, EHR-0233456, and EHR-0335411. The authors would like to thank the teachers who participated in this study and the district mathematics coordinator who enabled the study.
References


Teacher Quality and Quality Teaching


