



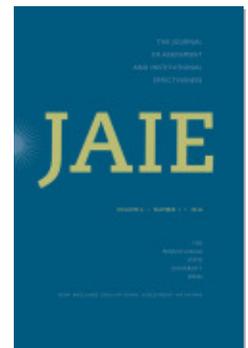
PROJECT MUSE®

Assessing Teacher Education through Mathematics Pupil
Performance: A Case Study about Implementation in Response
to External Pressure

Mary Yakimowski, Mary Truxal

Journal of Assessment and Institutional Effectiveness, Volume 6, Number
1, 2016, pp. 28-49 (Article)

Published by Penn State University Press



➔ For additional information about this article

<https://muse.jhu.edu/article/623782>



Assessing Teacher Education through Mathematics Pupil Performance: A Case Study about Implementation in Response to External Pressure

MARY YAKIMOWSKI AND MARY TRUXAL

ABSTRACT

Policymakers and accreditation agencies are now seeking to gauge the effectiveness of teacher preparation programs by following teacher candidates into their professional practice and, further, by linking to their pupils' academic performance. However, the task of gathering and analyzing such data is complex, especially within states that have not received federal funding to link the pupil test database to individual teachers by higher education institute where they attended. In this case study, researchers examine mathematics pupil performance in grades 3–8, as measured by the state-mandated assessment, and make connections to a specific university teacher education program. The results of this longitudinal study of pupil performance are shared in order to evaluate the specific teacher preparation program and provide a model for those who investigate the impact of teacher preparation programs. Additionally, obstacles faced and challenges of such a quantitative study for a higher education institution are shared. Keywords: Accreditation, institutional effectiveness, teacher education, mathematics, teacher performance

Introduction

The purpose of this study is to examine whether teacher education experiences in higher education have an impact on pupil performance in mathematics.¹

In an earlier work, Yakimowski and colleagues (2010) noted the importance of such an effort: “The quality and education of teachers is extremely complex and ever changing in this 21st century global world. Now, it is imperative that we look at the ‘product,’ that is, the PreK-12 pupil performance that is the result of teacher preparation at higher education institutions.”

Policymakers and the accreditation agency, the Council for the Accreditation of Educator Preparation (CAEP), are moving in the direction of seeking to assess the performance of teacher preparation candidates after they graduate and begin professional practice. Federal policymakers are moving in a similar direction as evidenced by recent changes in the Title II reporting requirements. In teacher preparation programs and other professional fields (e.g., nursing, business, athletic training, physical training), accreditation agencies do require reporting of candidate standardized test scores, employment records, and survey results from employers and/or alumni themselves. However, the field of education is taking this effort farther by addressing how the alumni contribute to their impact in the field. In the field of education, this impact often is defined as K-12 pupil learning. To gauge the amount of impact, data must be collected and analyzed demonstrating that higher education’s teacher preparation programs produce candidates that impact K-12 pupil performance (as often measured by standardized test scores).

There are challenges involved in measuring the impact of a teacher preparation program on the K-12 pupils that its graduates teach. For example, obtaining the necessary data can be complex for higher education institutions in states that did not “win” the federal funding needed to link the student testing databases to individual teachers, and then to the higher education institutions those teachers attended. Another challenge is the advent of new testing requirements associated with the Common Core State Standards (Council of Chief State School Officers and National Governors Association 2010). In this study, we did manage to gather the necessary data without federal assistance. We did this by seeking data from five local education agencies that generally hired many of our alumni. All agreed to assist us in investigating pupils’ achievement patterns in classes taught by graduates of the teacher preparation program. Also, because there were not yet sufficient years of data from assessments related to Common Core, we backtracked to the previous testing results to get multiple years of data.

In this case study, we will determine whether variations in mathematics pupil performance may be significantly related to formal teacher preparation,

with a particular focus on the teacher education program of one university of interest (UI), the University of Connecticut, as a point of reference. At the University of Connecticut, the Neag School of Education incorporates two components within the teacher preparation program: the Integrated Bachelor's/Master's (IB/M) program and the Teacher Certification Program for College Graduates (TCPCG) program. Based on principles established by the Holmes Partnership and work of the National Network of Educational Renewal, the IB/M teacher education program was established to prepare pre-service teachers to meet the needs of all students in all types of learning environments. Building on the existing strengths of the IB/M program, TCPCG was developed for college graduates who wish to gain teacher certification. Both components have evidence that they share a strong commitment to high standards and extensive clinical experience, as well as a concern for the development of reflective and analytic practitioners, for urban and multicultural issues in education, and for teacher leadership. The IB/M components educate students in the junior, senior, and fifth (master's) year, while TCPCG educates those students already with a bachelor's degree during the master's year. Within a public institution, this nationally accredited program graduates approximately 180 individuals annually, with about 86% of the alumni remaining in Connecticut and teaching with their initial certification.

Specifically, our objectives for this research study are to:

Measure the impact of teacher education experiences on pupil performance in mathematics.

Interpret the findings and provide recommendations for developing a model to evaluate teacher preparation programs in higher education institutions.

Then we will discuss the logistics of constructing such a study and the complexities that should be considered in addition to evaluating the effectiveness of the program in conjunction with other factors such as graduation and certification rates, alumni and employer survey data, and employment information.

Review of the Relevant Literature

Evidence exists to support the argument that teacher quality is a crucial element in the success of students in school (Darling-Hammond 2006; Hill, Rowan, and Ball 2005; Sanders 1998). High-quality teacher education

programs take on even more importance (Bransford, Darling-Hammond, and LePage 2005; Darling-Hammond 2006) as teaching becomes an increasingly complex endeavor. Hammerness et al. (2005) described key elements of teacher preparation programs and shared a conceptual model for a system of change from preservice to novice to master teacher. One defining characteristic in this conceptual model is the assessment of PreK-12 pupil performance. Accordingly, it follows that pupil performance can be an indicator of the quality of teacher preparation programs. However, critics note the lack of empirical evidence connecting teacher education programs with pupil outcomes (Crowe 2010; Grossman 2008). To address this concern, it is necessary to develop methods that can be used across institutions to track programs over time, “while respecting the complexity of linking initial preparation to eventual outcomes such as student achievement” (Grossman and McDonald 2008, 199).

Although there is little research investigating pupil outcomes in relation to teachers from specific preparation programs, there are methods for tracking pupil outcomes. For example, with the No Child Left Behind Act (NCLB) acting as a catalyst, there has been a marked interest in examining growth achievement models over the last several years. States such as Tennessee (e.g., Barone 2009) and Colorado (Colorado Department of Education 2016), along with organizations such as the National Gifted Association (Olszewski-Kubilius and Clarenbach 2012), have sought to move in this direction, though controversy exists regarding how these models should be applied in practice. Yakimowski and colleagues noted such shortcomings more than five years ago: “While this work is part of a broader movement on the part of university-based teacher education in the United States, most of the work on how teachers affect the academic outcomes of their pupils has been conducted by economists using various value-added models” (2010, 1). Such value-added models have been difficult to understand at best, requiring one simply to “trust” that they work. Providing a more transparent model could help to take the mystery out of making connections between pupil and teacher performance. However, although the effect of teachers on pupil achievement has received considerable attention in recent years (Lockwood et al. 2007; Wallace 2009), findings are limited.

To develop an enhanced and more transparent model for connecting pupil performance with teacher preparation, it makes sense to consider performance in a key curricular area that is assessed in a standardized manner—mathematics is such an area. Connecting mathematics pupil

performance with teacher preparation is important for at least two reasons. First, mathematics is a critical content area of focus. A message that has remained consistent across this decade from a wide variety of government, professional, and academic sources is that mathematics plays a significant role in preparing students with the problem-solving and analytical skills necessary for real life, higher education, and career pursuits (Council of Chief State School Officers and National Governors Association Center for Best Practices 2010; National Council of Teachers of Mathematics 2000, 2003, 2006; National Research Council 2001). Second, teachers' pedagogical content knowledge has been found to be critically important for supporting students' mathematical learning and performance (Ball 2003; Ball, Lubienski, and Mewborn 2001; Fennema and Franke 1992; Hill, Rowan, and Ball 2005; Shulman 1987). Although research has suggested practices for enhancing teachers' mastery of mathematical content and pedagogy (Wallace 2009), there is no consensus on how best to prepare teachers to achieve these ends and, in turn, facilitate student learning (Bransford, Darling-Hammond, and LePage 2005; Kirtman 2008). Research that links mathematics pupil performance and teacher preparation programs may begin to answer questions about how best to prepare teachers.

This challenge comes at a time when teacher preparation programs are undergoing greater scrutiny. Some teacher education researchers suggest that traditional teacher preparation programs are vulnerable to criticism due to lack of empirical evidence demonstrating that "how teachers are prepared does make a difference and that it makes a difference to the outcomes that the public cares most about—student learning" (Grossman 2008, 12). We contend that the quality of either university-based or alternative teacher preparation programs does matter and that it can impact pupil achievement.

In order to gauge the effectiveness of teacher preparation programs relative to pupil learning, and mathematical achievement in particular, it is essential to follow teachers from specific programs into professional practice and investigate their pupils' mathematical performance. The one remaining accreditation agency for all teacher preparation programs, CAEP, similarly is moving in the direction of having institutions report on the impact of their alumni on K-12 pupil performance (Council for Accreditation of Educator Preparation 2013). However, much of the available research is limited in scope to analysis of the preparation process rather than mathematics pupil achievement as operationalized by longitudinal performance on NCLB assessment outcomes (No Child Left Behind Act of 2001, U.S.C §

6301 [2002]). Additionally, only a small number of states receive Race to the Top federal funds to merge the state-level pupil databases with the teacher databases. Some, including even these “fortunate” states, do not further allow for the connectivity of the teacher database with the records of higher education institutions.

In terms of mathematics performance in particular, the mathematics education community has recommended reform-oriented, standards-based practices for educating elementary and middle-school teachers of this discipline (e.g., Council of Chief State School Officers and National Governors Association Center for Best Practices 2010; National Council of Teachers of Mathematics 2000, 2006; National Research Council 2001), but few studies have tied teacher education practices to pupil achievement. This shortcoming may stem from the complexity of gaining access to the necessary K-12 pupil achievement data. As evident from these existing gaps in the literature, there is still much to learn from research about the impact of various teacher preparation practices on pupil mathematics attainment. Thus, to provide links between teacher preparation and pupil performance, this study follows educators from one specific teacher preparation program into their professional years and compares their pupils’ mathematics achievement outcomes with the pupil performance of teachers (within the same school districts) who did not participate in that preparation program.

Along with overall achievement, this model examines mathematical strands and objectives, further elucidating connections between pupil performance and teacher education programs. This research has the potential to provide a model for gathering empirical evidence of effectiveness of teacher preparation programs with respect to student performance. Furthermore, Connecticut did not have access to the Race to the Top federal funds, thus demonstrating that it is possible to obtain necessary data without federal assistance.

Methodology

This research uses a quantitative mode of inquiry. The major research question follows: Is there a significant difference between teachers who graduated from one teacher preparation program and graduates from other institutions in terms of performance on the statewide mathematics assessment (i.e., the state test) which is used, in part, to determine adequate yearly progress as stipulated by federal legislation? More specifically,

are there significant differences and/or relationships between teachers who graduated from this teacher preparation program and other teachers in terms of (a) overall pupil mathematics performance, (b) five domain scores, (c) strand scores within each domain, (d) vertical scale scores, and (e) proficiency levels on the mathematics portion of the state test?

Sample and Data Sources

While we will refer through this monograph to the university of interest (or UI), the data were obtained on teachers who graduated from the University of Connecticut. We used data from both the IB/M and TCPCG components of their teacher education preparation program.

Only five states receive Race to the Top funding, which has provided resources to link the pupil test data with other data such as teachers and higher education files. Although Connecticut did not receive this funding, we were still able to obtain necessary data. We did this by approaching five local education agencies (i.e., school districts), and all agreed to assist us in investigating pupils' achievement patterns in classes taught by graduates of our teacher preparation program. We obtained a total of 49,402 data records from these five districts. It is important to note further that these five districts were not selected through random sampling, but data were collected from an intentional, convenience sample of districts that employed a high number of alumni from the UI, as determined by a previous study by Yakimowski and colleagues (2010). The third- through eighth-grade pupil data from the 2007–2008 and 2008–2009 school years were then validated and matched. Of further note is that testing these hypotheses required clean and comparable data across years. Connecticut had been administering its statewide testing program since the early 1980s. This testing program went through four significant revisions during that time. The comparability across each significant revision was not possible. Additionally, because of changes to federal legislation, this state replaced its testing program and joined one of two state consortia involved with developing, piloting, and adopting for use a new assessment (i.e., Smarter Balance). At this time, longitudinal data were still not available on this new assessment.

Instrumentation

Individual pupil results from one of the state's assessment programs were obtained for the same two academic years. Specifically, at the time of the research, public school pupils in grades 3–8 were required to

participate in a statewide test in Connecticut. The test assessed essential reading, writing, mathematics, and science skills, as identified in the state's curriculum framework, focusing on content that pupils at each grade level can reasonably be expected to master. It was intended to fulfill the following purposes:

- Set high expectations and standards for pupil achievement;
 - Test a comprehensive range of academic skills;
 - Disseminate useful test achievement information about pupils, schools, and districts;
 - Identify pupils in need of intervention;
 - Assess equitable educational opportunities; and
 - Continuously monitor pupil progress in grades 3–8 over time.
- (Joldersma 2007)

The curriculum frameworks help to identify the knowledge, skills, and understanding needed for pupils, as well as to provide guidance in designing curriculums for schools and districts (Connecticut State Department of Education 2008a).

Procedures

We collected the individual pupil information on the state test including the overall raw scores for the total mathematics portion of the test and results from each of the five domains. These domain scores from the mathematics portion were for numerical and proportional reasoning (NP), geometry and measurement (GM), working with data: probability and statistics (DPS), algebraic reasoning: patterns and functions (AR), and integrated understanding (IU). We collected all information on the 25 strands that were spread across the domains (i.e., ma1 through ma25).

Numerical and Proportional (NP)

1. Place Value
2. Pictorial Representations of Numbers
3. Equivalent Fractions, Decimals and Percents
4. Order, Magnitude and Rounding of Numbers
5. Models for Operations
6. Basic Facts
7. Computation with Whole Numbers and Decimals

8. Computation with Fractions and Integers
9. Solve Word Problems
10. Numerical Estimation Strategies
11. Estimating Solutions to Problems
12. Ratios and Proportions
13. Computation with Percents

Geometry and Measurement (GM)

14. Time
15. Approximating Measures
16. Customary and Metric Measures
17. Geometric Shapes and Properties
18. Spatial Relationships

Working with Data: Probability and Statistics (DPS)

19. Tables, Graphs and Charts
20. Statistics and Data Analysis
21. Probability
22. Classification and Logical Reasoning

Algebraic Reasoning: Patterns and Functions (AR)

23. Patterns
24. Algebraic Concepts

Integrated Understanding (IU) (May include content from one or more of the four domains)

25. Mathematical Applications

We then gathered proficiency-level (Below Basic, Basic, Proficient, Goal, Advanced) information and, using a conversion table provided by the Connecticut State Department of Education, to convert raw test scores to “vertically scaled scores.” This vertical scale information “allows for valid interpretations of growth across time using tests different in content, length, and item difficulty” (Connecticut State Department of Education

2008b, 3). Originally, the test did not afford the ability to evaluate growth as students moved up the grades because their performance was based on specific skills tied to the respective grade level. Starting with the fourth generation of this test, however, state department officials completed a linking process, called vertical scaling, to measure growth across grades. The range of vertical scores is from 100 to 800, and each of the vertical scale's score points represents the same theoretical position for each grade. Vertically scaled scores allowed us to track not just individual year performance, but also growth of the pupils across years.

We gathered the teacher grouping variable (UI vs. non-UI) and made comparisons by matching within and then comparing across all the districts. Our connection of pupil performance to higher education institution was noteworthy for a number of reasons. Primarily, this present research marked the first effort of its kind undertaken in Connecticut, as teacher and student files are not linked, and the state is not a recipient of Race to the Top federal funding. A second major difference is that the district and university staff worked together to get accurate comparative data. This was due to the relationship-building activities focused on strengthening trust between each district and the university over several years.

Parameters and limitations proposed by Yakimowski et al. (2010) were kept in mind when developing the research design used in this study. For example, cautions about the limitations and advantages of placing all of the individuals not attending the UI together in the "non-UI" group needing "matched" data, and those who left the state to teach are explored further in this manuscript.

Analyses

We calculated descriptive measures (measures of central tendency and distribution) to examine the overall mathematics achievement, domains, and strand scores.² We also ran a correlation analysis using strand scores within each domain on mathematics tests for each group of teachers. The averages of each domain score and their corresponding strand scores were compared to see whether pupils with teachers who graduated from the UI teacher preparation program performed differently than those whose teachers did not attend the UI.

We then conducted a proportional analysis using proficiency-level scores from student performance on the mathematics tests for each group of teachers and compared the results from the two groups. A proportional analysis

TABLE I. Analyses used to examine the research questions

| Research Question | Scores | Type | Other |
|---|-------------------|---|--------------------------------|
| What is the overall mathematics achievement for pupils educated by UI alumni? Is it any different from pupils of alumni from other institutions? | Overall raw score | Descriptive, <i>t</i> -test | 1 year |
| What are the five mathematics domains scores for pupils educated by UI alumni? Is performance in the domains for these pupils any different from pupils of alumni from other institutions? | 5 domain scores | Descriptive, <i>t</i> -test; Correlation analysis | 1 year |
| What are the mathematics strand scores of pupils of UI alumni? Is this performance any different from pupils of alumni from other institutions? | 25 strands | Descriptive, <i>t</i> -test Correlation analysis | 1 year |
| What is the pattern in pupil proficiency status for those educated by UI alumni? Is the performance any different from pupils of alumni from other institutions? For example, are UI alumni pupils at a higher proportion in the goal and advanced categories and the Non-UI in the Below Basic and Basic categories? | Proficiency level | Proportion analysis | 1 year |
| What is the overall pupil mathematics achievement for those educated by UI alumni after controlling for initial difference on earlier achievement? Is it any different from pupils of alumni from other institutions? | Vertical scale | ANCOVA | Covariate: last year's results |

was used to analyze this ordinal variable—mathematics performance level—in order to explore the degree of differences that may exist. For example, we could determine the percentage of students in the Below Basic level for each grouping.

Finally, an analysis of covariance (ANCOVA) was completed to examine differences in student performance on the test with the teacher variable as the two-level independent variable (UI vs. non-UI), using the interval measure of the 2008–2009 mathematics vertical scale (MAVS) with the matched 2007–2008 MAVS serving as a covariate. These analyses were conducted in order to statistically control for initial differences (i.e., 2007–2008) prior to the current school year of instruction with a teacher by adjusting for 2008–2009 mathematics achievement. The research questions and a summary of the corresponding analyses are presented in table 2.

Results

The overall mathematics performance, domain scores, and strand information indicate that the UI alumni pupil performance is higher than among pupils of alumni from other institutions. For example, the overall score for UI was 106.0 (SD = 22.8) compared to Non-UI of 95.3 (SD = 26.8).

In examining each strand under the respective domains, we found that the mean scores on five domains and their corresponding strands in 2008–2009 are generally higher for pupils who had UI teachers than for those students who had non-UI teachers with NP having the largest differences. For example, Domain 1: Numerical and Proportional was 53.3 for UI compared to 46.4 for Non-UI. While the range indicates similar performance, examining one standard deviation from the mean tells us where most of the group performed.

In addition to examining the descriptive information by domain and by strand, we ran a correlational analysis to look at the relationship between pairs of strands. For example, do pupils who perform well in algebraic reasoning also perform well in geometry and measurement? Did the performance in one strand relate to performance in another strand under the same domain? As an example, we found a higher relationship for algebraic reasoning (i.e., AR) and other domains for pupils with UI teachers than among those with non-UI instructors, which suggests that pupils with teachers from the UI would be more likely to relate the knowledge and skills in algebraic reasoning with those in other domains. However, correlations between integrated understanding (i.e., IU) and other domains for pupils

TABLE 2. Overview of the findings tied to each research question

| Research Question | Scores | Results |
|--|-------------------|---|
| What is the overall pupil mathematics achievement educated by UI alumni? Is it any different from alumni from other institutions? | Overall raw score | UI score (106.1) > Non-UI score (95.3); statistically significant |
| What are the five mathematics domains scores educated by UI alumni? Is the pupil performance any different from alumni from other institutions? | 5 domain scores | UI Domain scores > Non-UI Domain scores; statistically significant; e.g., for Domain1, UI score (53.3) > Non-UI score (46.4); statistically significant |
| What are the 25 mathematics strand scores from UI alumni? Is the pupil performance any different from alumni from other institutions? | 25 strands | UI Strand scores > Non-UI Strand scores; statistically significant; e.g., for strand 9, UI score (5.6) > Non-UI score (4.6); statistically significant |
| What is the pattern in pupil proficiency status by those educated by UI alumni? Is the pupil performance any different from alumni from other institutions? For example, are UI at a higher proportion in the goal and advanced categories and the Non-UI in the Below Basic and Basic categories? | Proficiency level | 76% of UI and 60% of Non-UI in the Goal and Advanced categories; 9% of UI and 20% of Non-UI in the Below Basic and Basic categories |
| What is the overall pupil mathematics achievement educated by UI alumni after controlling for initial difference on earlier achievement? Is it any different from alumni from other institutions? | Vertical scale | UI score (536.7) > Non-UI score (525.6); statistically significant after controlling for initial difference on earlier achievement |

with UI teachers were lower than among those who were non-UI, which suggests that pupils with UI teachers were less likely to relate the knowledge and skills in integrated understanding with those in other domains.

In examining proficiency levels, the proportional analysis results show that only about 9% of pupils with UI teachers scored at the Basic and Below Basic levels, compared to about 20% of pupils with teachers who were

not UI graduates in 2008–2009. Also, about 76% of pupils with the UI teachers scored at or beyond the Goal level in 2008–2009, compared to about 60% of pupils with teachers who were not UI graduates (see table 3).

Tied to the vertical scale, where growth across years can be determined, we can see that the UI teachers had higher performance than non-UI teachers on the mathematics performance (i.e., MAVS) for both 2007–2008 ($n = 816$, $m = 534.2$; $n = 9072$, $m = 513.0$) and 2008–2009 (564.2 vs. 541.3); furthermore, the differences on the 2008–2009 MAVS between the two groups of teachers are statistically significant after controlling for initial differences on the MAVS in 2007–2008. After controlling for initial differences, the adjusted means on these 200–800 vertical scales were significantly different (549.1 vs. 542.6), indicating that the UI had higher levels of mathematics achievement than the non-UI programs. Significant differences in vertical-scale scores between pupils of UI and non-UI teachers demonstrate greater academic growth, not just higher performance for one year.

Conclusions, Lessons, and Discussion

We sought to measure the impact of teacher education experiences among our university’s graduates on pupil performance in mathematics, and then interpret the findings and provide recommendations for developing a model to evaluate teacher preparation programs in higher education institutions. Our monograph further attempted to discuss the most pertinent logistics of constructing and executing such a study and the complexities that should be considered in evaluating the effectiveness of the program.

TABLE 3. Proportional analysis showing performance level on the 2008–2009 mathematics assessment by group and overall total

| Level | n | |
|-------------|---------------|-------------|
| | Non-UI School | UI |
| Below Basic | 1022 (9.1%) | 47 (4.2%) |
| Basic | 1204 (10.7%) | 49 (4.4%) |
| Proficient | 2295 (20.4%) | 172 (15.4%) |
| Goal | 3659 (32.6%) | 404 (36.2%) |
| Advanced | 3057 (27.2%) | 445 (39.8%) |
| Total | 11237 (100%) | 1117 (100%) |

TABLE 4. ANCOVA Results: 2008–2009 mathematics pupil performance on the adjusted vertical scale scores based on initial differences on the 2007–2008 scores

| | <i>F</i> | <i>p</i> |
|--------------------------|----------|----------|
| Between-Subjects Effects | | |
| Intercept | 3914.3 | 0.001*** |
| MAVS2007–2008 | 32283.6 | 0.001*** |
| Teacher Group | 36.6 | 0.001*** |

*** Statistically significant at .001 level

In order to address external requests (e.g., policymakers and accreditation needs), this study did investigate the performance of pupils whose teachers graduated from a specific teacher preparation program—thereby providing a means of further evaluating the effectiveness of the program in conjunction with other factors such as graduation and certification rates, alumni and employer survey data, and employment information. Although the mathematics education community has recommended reform-oriented, standards-based practices for educating elementary and middle school mathematics teachers (Council of Chief State School Officers and National Governors Association Center for Best Practices 2010; National Council of Teachers of Mathematics 2000; National Research Council 2001), few studies have tied teacher education practices to pupil performance, particularly in mathematics. This research demonstrates that quantitative analysis can be used, albeit cautiously, to examine pupil performance for teachers associated with a specific teacher preparation program.

We demonstrate that it is possible to measure the relationship between graduates of a teacher preparation program and mathematics performance of their K-12 pupils. In this case study, teachers who graduated from a preparation program at the UI were shown to have significantly higher pupil performance than other teachers in the same five school districts. The examination of these positive results begs us to answer another question: Why were they better? Perhaps the candidates in this institution were better. Perhaps the curriculum in the UI was better. Perhaps it was because of the long-standing relationship between this higher education institution and these school districts. Perhaps it was a combination of all. Certainly, this question would be of value to study further. In addition, curricular changes could be piloted—for example, focusing on strands such

as integrated understanding that, for UI, were demonstrated to be relatively lower than other domains and strands within the mathematics assessment.

Thus, these findings suggest that the teacher education practices and policies of this program may be worthy of further examination using qualitative inquiry.

There were many lessons learned by undergoing this study. For example,

Pupil databases across and within districts change every few years, and they do not necessarily readily communicate with one another.

Access to pupil performance data linked to subject-matter teachers (e.g., mathematics, reading/language arts, and science) is at best problematic at the district level in the elementary grades.

Teacher education programs must consider working together and with districts to find a consistent means to measure impact on pupil performance for content areas where there is no statewide achievement measure such as social studies, foreign language, music, and art.

Pupil mobility is inconsistent across districts, thus confounding the complexity of the data collection and the analyses.

When conducting a longitudinal cohort study, analyses may be biased for teachers. For example, teachers working in urban districts might start with pupils' scores being low and then have the potential for significant growth. In a community with greater financial resources, teachers could be educating pupils who are starting with scores close to the ceiling of the test, with much less room for growth.

Teacher mobility is of note as instructors can change grade level and/or schools within and across districts from year to year. This mobility results in missing data in a longitudinal cohort study.

There is a need for a stable vertically scaled statewide assessment system to facilitate analyses across grades to demonstrate pupil academic growth.

There is a need to differentiate among categories of special education and English-learner classifications and incorporate assessments for grouping purposes.

Small sample sizes are unavoidable for higher education institutions graduating many teachers who are employed in many districts rather than a handful of districts. Also, an institution needs to factor the number of graduates each year into its research design.

It is important to ensure complete confidentiality to all participating school districts, teachers, and pupils. The "trust" does not happen overnight; it takes many years to develop.

While we faced many challenges and learned much that was unanticipated, this study also provided a vehicle through which to investigate the impact on pupil performance generated by teacher preparation programs other than our own—thereby adding a basis for potentially informing statewide mathematics teacher education policy and practice.

Though we did accomplish means of further assessing the effectiveness of the program of the UI and generated new “learnings” (i.e., interesting knowledge and insight that we gained through the various stages of conducting the study), we also confronted a plethora of challenges involved in measuring the impact of a teacher preparation program’s graduates on K-12 pupils. Aside from financial and labor resources to conduct such a study, we are obliged to acknowledge these complexities.

First, it took an immense amount of effort to obtain the information from the districts. Second, we acknowledge that some demographic information, such as the institution’s entrance criteria, should be obtained to potentially use as covariates. Third, additional or alternative analytical techniques should be considered—for example, including districts, schools, and then teacher factors in a hierarchical design. There are also complexities involved in conducting a longitudinal study. These include pupil mobility (as pupils transfer from one school to another, both within and across districts, the result is missing data within a longitudinal cohort study), teacher mobility (as teachers change grade level and/or schools within and across districts, the result is also missing study data), and the need for a stable vertically scaled statewide assessment system to facilitate analyses across grades and demonstrate pupil academic growth. Also, the need for qualitative information to supplement the quantitative analyses is paramount.

We caution against a higher education institution, policymakers, and/or accreditation agencies relying on evaluation of pupil performance as their *only* tool for assessing the impact of teacher preparation programs. We propose that a higher education institution implementing an evaluation of its teacher education program should use multiple sources of data (i.e., triangulate data) that periodically include pupil performance. Higher education institutions generally examine the number of teaching certificates awarded to their alumni. They might look as well at their alumni’s employment records, professional awards/recognitions and involvement (at the local, state, regional, and/or national level), promotions, and retention. It also might be wise to routinely conduct an alumni survey and/or employer survey.

The more traditional view of assessing impact omits the major stakeholder group—the pupils. Our study showed how to access and analyze

some objective data—in this case, statewide assessment—to examine impact on pupil learning and achievement. Perhaps other measures—such as amending the alumni survey to provide evidence for pupil learning—must also be examined.

Pupil achievement studies linked to teacher preparation programs should be *cautiously* conducted to provide feedback for programs, school administrators, and researchers. For the school administrators, this type of research can provide evidence of the effectiveness of the higher education institution from which the applicant graduated. For the teacher preparation program, it can provide some perspective on what in the mathematics field is taught well and what needs to be reconsidered; for researchers, the features of a teacher preparation program that successfully captures “bang for the buck” can be investigated. Overall, higher education programs dedicated to training teachers do need to continue to strive to be evidence-based. By so doing, all stakeholders benefit as we strive for continued improvement of the educational opportunities of our nation’s pupils.

As noted by Education Secretary Arne Duncan at an annual meeting of the American Association of Colleges of Teacher Education in February of 2010 in Atlanta, “To put it in the simplest terms, we believe teacher-preparation programs should be focused on results.” Aligned with Secretary Duncan’s charge, we do contend that we must continue to strive to build an evidence-based teacher preparation model that is directly linked to pupil academic performance; however, we also recognize that such a model will add only one piece, albeit an important one, to the puzzle of authentic program evaluation for teacher education.

Bridging across three departments in the Neag School of Education and three additional schools/colleges at the University of Connecticut, this study has prompted our faculty within our educator preparation program to undertake further discourse on how we can judiciously examine our alumni’s impact on P-12 schools. This is indeed timely for our colleagues in other institutions as both national and state accreditation organizations are similarly looking to move in that direction. For example, the fourth standard of CAEP requires each teacher education program to demonstrate the impact of P-12 student learning and development, classroom instruction, and schools (2013). State policymakers are moving in this direction, too. For example, the Connecticut State Department of Education is drafting a proposal to deliver to the board of education next fall that requires teacher education programs show P-12 pupil impact.

When enrolled in teacher preparation programs at higher education institutions, students are assessed on their mastery of the knowledge, skills, and dispositions necessary for effective teaching. There is a push to go further to measure the impact of those who went through a teacher preparation program on the job with P-12 student learning and development. By having tried to demonstrate impact on pupil performance, this study now provides this university with a solid understanding of the obstacles, limitations, and advantages to moving in this direction. Thus, we are sharing this case study with other higher education institutions and policymakers because what seems so simple—that is, measuring the impact of teacher preparation on pupil achievement—is quite complex, time consuming, and only provides a partial answer. Due diligence to this zeitgeist is paramount.

MARY E. YAKIMOWSKI's scholarly work has focused on assessment, program evaluation, urban education and research in the schools. She now serves as assistant dean for the College of Education at Sacred Heart University in Fairfield, Connecticut. She is past vice president of AERA, past president of DRE, CTN, NATD/NAAD, and a winner of several national best papers awards.

MARY TRUXAW is an associate professor of mathematics education at the Neag School of Education at the University of Connecticut. She works with preservice and in-service teachers and conducts research at the intersection of mathematics education, teacher education, and language.

Notes

1. Throughout this article, the term “pupils” will refer to K-12 students, while “candidates” will refer to those who partake in a preparation program, such as the teacher education pre-service preparation program offered by a higher education institution.
2. Unlike the vertical scales, the domain and strand scores do not have a statistical mechanism to measure growth over time.

References

- Ball, D. L., S. T. Lubienski, and D. S. Mewborn. 2001. “Research on Teaching Mathematics: The Unsolved Problem of Teachers’ Mathematical Knowledge.” In *Handbook of Research on Teaching*, 4th ed., ed.

- V. Richardson, 433–56. Washington, DC: American Educational Research Association.
- Barone, C. 2009. “Are We There Yet? What Policymakers Can Learn from Tennessee’s Growth Model?” Education Sector Technical Reports, March. Retrieved from http://txcc.sedl.org/orc/resources/Are_We_There_Yet.pdf.
- Bransford, J., L. Darling-Hammond, and P. LePage. 2005. “Introduction.” In *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do*, ed. L. Darling Hammond, J. Bransford, P. LePage, K. Hammerness, and H. Duffy, 1–39. San Francisco, CA: Jossey-Bass.
- Colorado Department of Education. 2016. “Schoolview.” Retrieved from <http://www.cde.state.co.us/schoolview/coloradogrowthmodel>.
- Connecticut State Department of Education. 2008a. *Connecticut Mastery Test: Fourth Generation Mathematics Handbook*. Hartford: Connecticut State Department of Education.
- . 2008b. *Connecticut Mastery Test: Vertical Scales*. Hartford: Connecticut State Department of Education.
- Council for Accreditation of Educator Preparation. 2013. “CAEP Accreditation Standards and Evidence: Aspirations for Educator Preparation.” Downloaded from <http://edsources.org/wp-content/uploads/commrpt.pdf>.
- Council of Chief State School Officers and National Governors Association Center for Best Practices. 2010. *Common Core State Standards for Mathematics*. Retrieved from http://www.corestandards.org/assets/CCSSI_Math%20Standards.pdf.
- Crowe, E. 2010. “Why Is Looking at Teacher Performance on Students So Important? The Status of Research and Policy.” Roundtable presentation at the meeting of the American Educational Research Association, Denver, CO, May.
- Darling-Hammond, L. 2006. “Constructing 21st-Century Teacher Education.” *Journal of Teacher Education* 57:300–314.
- Fennema, E., and M. L. Franke. 1992. “Teachers’ Knowledge and Its Impact.” In *Handbook of Research on Mathematics Teaching and Learning*, ed. D. A. Grouws, 147–64. New York: Macmillan.
- Grossman, P. 2008. “Responding to Our Critics: From Crisis to Opportunity in Research on Teacher Education.” *Journal of Teacher Education* 59 (1): 10–13.

- Grossman, P., and M. McDonald. 2008. "Back to the Future: Directions for Research in Teaching and Teacher Education." *American Educational Research Journal* 45 (1): 184–205.
- Hammerness, K., L. Darling-Hammond, P. Grossman, F. Rust, and L. Shulman. 2005. "The Design of Teacher Education Programs." In *Preparing Teachers for a Changing World: What Teachers Should Learn and Be Able to Do*, ed. L. Darling Hammond, J. Bransford, P. LePage, K. Hammerness, and H. Duffy, 390–441. San Francisco, CA: Jossey-Bass.
- Hill, H. C., B. Rowan, and D. L. Ball. 2005. "Effects of Teachers' Mathematical Knowledge for Teaching on Student Achievement." *American Educational Research Journal* 42 (2): 371–406.
- Joldersma, K. B. 2007. *The Connecticut Mastery Test: Fourth Generation Technical Report*. Hartford, CT: Connecticut State Department of Education.
- Kirtman, L. 2008. "Pre-service Teachers and Mathematics: The Impact of Service-Learning on Teacher Preparation." *School Science and Mathematics* 108 (3): 94–102.
- Lockwood, J. R., D. F. McCaffrey, L. S. Hamilton, B. Stecher, V. Le, and J. F. Martinez. 2007. "The Sensitivity of Value-Added Teacher Effect Estimates to Different Mathematics Achievement Measures." *Journal of Educational Measurement* 44 (1): 47–67.
- National Council of Teachers of Mathematics. 2000. *Principles and Standards for School Mathematics*. Reston, VA: National Council of Teachers of Mathematics.
- . 2003. "NCATE/NCTMA Program Standards: Programs for Initial Preparation of Mathematics Teachers." Retrieved from http://www.nctm.org/uploadedFiles/Standards_and_Positions/CAEP_Standards/NCTMSECONStandards.pdf.
- . 2006. *Curriculum Focal Points for Prekindergarten Through Grade 8 Mathematics: A Quest for Coherence*. Reston, VA: National Council of Teachers of Mathematics.
- National Research Council. 2001. *Adding It Up: Helping Children Learn Mathematics*. Mathematics Learning Study Committee. Washington, DC: National Academy Press.
- Olszewski-Kubilius, P., and J. Clarenbach. 2012. "Unlocking Emergent Talent: Supporting High Achievement of Low-Income, High-Ability Students." Washington, DC: National Association for

Gifted Children. Retrieved from http://www.jkcf.org/assets/1/7/Unlocking_Emergent_Talent.pdf.

Sanders, W. L. 1998. "Value-added Assessment." *School Administrator* 11 (55): 24–27.

Shulman, L. S. 1987. "Knowledge and Teaching: Foundations of the New Reform." *Harvard Educational Review* 57 (1): 1–22.

Wallace, M. R. 2009. "Making Sense of the Links: Professional Development, Teacher Practices, and Student Achievement." *Teachers College Record* 111 (2): 573–96.

Yakimowski, M. E., S. Brown, M. Kehrhahn, R. Sen, W. Xia, and M. Eastwood. 2010. "An Examination of Grades 3–8 Reading Achievement Using a Longitudinal Design in Educational Expansions Research." Paper presentation at the meeting of the American Educational Research Association, Denver, CO.