# Working with What They Have: Professional Development as a Reform Strategy in Rural Schools

Nathan Barrett
*Tulane University*

Joshua Cowen
*Michigan State University*

Eugenia Toma
Suzanne Troske
*University of Kentucky*

*In-service teacher professional development has been used to improve teacher effectiveness. In Kentucky, the National Science Foundation funded a large professional development program called the Appalachian Math and Science Partnership (AMSP) to provide content-based professional development to teachers in rural schools. We show that students assigned to AMSP teachers at a baseline year realized significant math gains not only in that year of assignment but in the following year as well. No gains are evident two and three years after being assigned to AMSP teachers. We frame both the program and its results in the context of teaching careers in rural schools, arguing that limited access to outside labor markets implies that successful professional development may be a key component of improving education in rural locales.*

In the current educational environment, scholars and policymakers alike are focused on improving student outcomes. Their efforts have been particularly directed toward sources of inequality, typically defined on the basis of student racial/ethnic identity and geographic locale. A variety of reforms aimed at increasing the number of viable school choices for underprivileged students, holding schools accountable for results, or unifying academic standards across states and regions have formed the basis for policy change in recent years. In addition, and in recognition that teacher quality varies markedly across contexts (e.g.,

Aaronson, Barrow, & Sander, 2007; Rockoff, 2004; Rivkin, Hanushek, & Kain, 2005), many of the latest reforms are directed specifically at the teaching profession.

Nearly all the literature on teacher quality and the achievement gap has focused on the differences between suburban and urban (and particularly inner-city) schools. It is well-documented that urban schools with primarily minority students, students of lower socioeconomic status, and students with low academic performance are generally served by less effective teachers (Boyd, Lankford, Loeb, & Wyckoff, 2005; Chester & Beaudin, 1996; Goldhaber & Hansen, 2009; Hanushek, Kain, & Rivkin, 2004; Lankford, Loeb, & Wyckoff 2002; Loeb, Darling-Hammond, & Luczak, 2005). Rural schools remain under-examined relative to their suburban and urban counterparts across a variety of reform dimensions (Arnold, Newman, Gaddy, & Dean, 2005; Ballou & Podgursky, 1995; Ingersoll & Rossi, 1995; Miller 2012; Sherwood, 2000). The absence of emphasis on rural locales in the educational policy literature generally, and in the teacher quality literature specifically, is especially glaring given the possibility that improvements to the teaching workforce are among the more direct ways in which policymakers may plausibly influence student achievement in these areas.

In rural locales, reforms based on school choice and accountability may either be infeasible, of limited long-term impact, or at least in need of application to a particular rural context (Cowen, Butler, Fowles, Streams, & Toma, 2012; Miller, 2012). For example, one tenet of the federal No Child Left Behind (NCLB) reform is that schools will respond to accountability pressures to improve outcomes. Although the literature varies somewhat, studies of NCLB and other similar performance-based regimes in typically large, urban settings have generally shown improvement in student test scores (Carnoy & Loeb, 2002; Dee & Jacob, 2011; Hanushek & Raymond, 2005; Jacob, 2005; Rockoff & Turner, 2010). In such locales, however, one of the primary sanctions for sustained low performance is the threat of school reorganization; closure; or, especially, competition from charter schools and other alternatives. These options are often not available in rural districts, particularly those in which only one school at each level serves the community's schoolchildren.

Moreover, even more broad-based teacher quality reform strategies may be difficult to implement in rural contexts. The most recent of these reforms emphasize the use of teacher evaluation based on student achievement. In more than 20 states, teacher employment is at least partly contingent on evidence of student learning (Winters & Cowen, 2013a), but implicit in any effort to improve the teacher workforce by dismissing ineffective teachers is the idea that more effective teachers are available to take their place (Rothstein, 2012; Winters & Cowen, 2013b). Research on teacher staffing in urban areas has emphasized the difficulty in recruiting and retaining high quality teachers for at-risk children (Boyd et al., 2005; Chester & Beaudin, 1996; Goldhaber & Hansen, 2009; Hanushek et al., 2004; Lankford et al., 2002; Loeb et al., 2005), and there is reason to believe that rural locales face similar, yet unique, challenges in this regard. Policymakers have devised programs with the express intent of making teaching a more attractive profession in rural communities, thus improving the workforce by bringing new employees from outside these areas (Streams, Butler, Cowen, Fowles, & Toma, 2011). In Kentucky, for example, the Kentucky Education Reform Act of 1990 (KERA) removed the locally based finance structures that had characterized the state, thereby equalizing expenditures between school districts statewide, including expenditures on teacher salaries. Although the immediate impact of this systemic reform was indeed parity in financial outlays, the reform did not result in changes to patterns of teacher entry or exit between locales (Cowen et al., 2012), and the expenditures themselves began to diverge again after only a few years (Streams et al., 2011). This evidence underscores the possibility that the fundamental problem facing teacher quality reforms in rural locales is simply that teaching is largely a "home-grown" workforce. Studies in a variety of contexts have confirmed teacher placement patterns within limited geographic distance from home high schools or colleges—a phenomenon that is especially apparent in teaching, and acute in rural schools (Boyd et al., 2005; Fowles, Butler, Cowen, Streams, & Toma, 2014; Miller 2012; Reininger 2012).

All these research findings imply that any improvement to teacher quality in rural locales must include—and perhaps be centered around—developing the skills of teachers who are already committed to their classrooms and schools. In this article, we consider a particular professional development effort that partnered teachers in Appalachian school districts with university faculty members to create an intense, content-based training experience in mathematics and science subjects. Building off earlier work, our research question in this article tests the hypothesis that students whose teachers initially received intensive, content-based math and science professional development benefit academically in future years as well. Specifically, we ask this research question: *Do students whose teachers received math/science content-based professional development in a given year have higher mathematics scores one, two, or three years following exposure than students whose teachers never had such training?*

Such longer-term improvement would represent a potential framework for systematic improvement of student outcomes in a particular content area and from a rural region in which students have been historically disadvantaged. Our evidence indicates that gains in math achievement were indeed realized for at least one additional year following teacher training, although we do not see evidence that such gains lingered for two and three years following the training. After providing background on professional development in rural areas generally and in the program under study in particular, we discuss our research design and results. We conclude by framing our findings as evidence that well-targeted interventions linking teacher knowledge to student results are promising avenues for educational policy in rural locales.

## Background: Professional Development in Appalachian Kentucky

Teacher quality is particularly problematic in Appalachian regions of Kentucky where some of the state's lowest student achievement exists. Studies of Kentucky school staffing have found that most teachers who teach in Appalachia receive their baccalaureate degree from an Appalachian institution and are more likely to obtain first employment in that region than are those who attended a college or university outside the region, controlling for other factors (Fowles et al., 2014). Other research has also found that teachers in general do not move between Appalachian and non-Appalachian locales, especially in the most rural areas (Cowen et al., 2012). The few Appalachian

teachers who do take a first teaching position outside of Appalachia tend to have higher credentials. In other words, the teachers who take positions in Appalachian schools not only have the weakest credentials upon entry, but are the same teachers who spend their entire careers in Appalachian school districts. One area of ongoing work is the potential for professional development programs to improve teacher knowledge and, ultimately, effectiveness.

Most states require teachers to participate in some form of continuing education known as in-service training or professional development. The underlying or implicit belief is that participation improves the quality of the teacher and will lead to improved student achievement. A large literature has considered professional development impacts in a number of frameworks and methodologies. Some studies test specific elements of professional development such as content vs. pedagogical training for teacher improvement (Desimone, 2009; Loucks-Horsley, Stiles, Mundry, Love, & Hewson, 2009; Wayne, Yoon, Zhu, Cronen, & Garet, 2008). Others focus on teacher perceptions of whether professional development made them more effective (Garet, Porter, Desimone, Birman, & Yoon, 2001). The latter research finds that professional development increased teachers' self-reported knowledge and skills. Ball and Cohen (1999); Hill and Ball (2004, 2009); and Hill, Rowan, and Ball (2005) go beyond the self-reported perceptions of teachers to develop measures of teacher knowledge enhanced by professional development programs. In these studies, teachers of mathematics show measured gains in mathematics knowledge from professional development that was focused on one subject or was a summer-length institute. Desimone, Porter, Garet, Yoon, & Birman (2002) and Desimone (2009) link the type of professional development to observed changes in the practice of teachers. They find that especially focused instructional practice resulted in more use of the learned practices in the classroom. Similarly, Penuel, Fishman, Yamaguchi, and Gallagher (2007) argue that interventions that stress curriculum implementation may be among the most successful. Penuel, Gallagher, and Moorthy (2011) have also emphasized providing models of teaching for participants in effective professional development programs. Foster, Toma, and Troske (2013) consider professional development in a school building and whether it improves student achievement. They find positive results for math achievement in middle school math, and that for these middle schools the professional development was cost-effective for the results achieved. Grigg, Kelly, Gamoran, and Borman (2013) find evidence that inquiry-based science practices increased for teachers who participated in a district-wide development program in Los Angeles, but that these changes were limited to select areas of instruction. This research also provides evidence that professional development may actually have negative impacts on student achievement; novice teachers

appeared to become more effective, while veteran teachers realized a large negative impact (Borman, Gamoran, & Bowdon, 2008). Thus, despite the considerable attention to professional development among both policymakers and practitioners, Wayne et al. (2008) argue that we know little about whether professional development generally delivers positive effects on student achievement. As a result, they call for more methodological diversity in trying to evaluate the effectiveness of professional development programs, and, in particular, they call for experimental and quasi-experimental study designs.

As part of the vision of NCLB, and in response to concerns over our declining national competitiveness especially in science, technology, engineering, and math (STEM) subjects, the National Science Foundation (NSF) launched an initiative in 2002 to improve the quality of teaching in the STEM areas. NSF has provided support for professional development in the Appalachian regions of Kentucky along with over 30 other states. Since 2002, NSF has allocated over $800 million to this initiative. The program is focused on the creation of math and science partnerships (MSP) between institutions of higher education and K-12 schools to increase the quality of teachers, but it has as its ultimate goal the improvement in student outcomes in STEM subject areas. A requirement of each partnership is STEM faculty (most commonly math and science and some engineering and computer science faculty) involvement from participating institutions of higher education. While there are over 40 targeted and comprehensive funded partnerships around the country, one of the largest of the initial programs was located in the central Appalachian states of Kentucky, Tennessee, Virginia, and West Virginia and is known as the Appalachian Math and Science Partnership (AMSP). This program forms the basis of our research here.

The AMSP received an initial five-year grant of $22.5 million from NSF and began implementation in the 2002-2003 school year. The program was phased in with the peak level of teacher participation in 2005-2006, a slight decline in 2006-2007, and then a gradual phase-out. AMSP began as a partnership among 38 central and eastern Kentucky school districts, nine Tennessee school districts, and five western Virginia school districts, the Kentucky Science and Technology Corporation, and 10 higher education institutions located in these three states, although our analysis focuses only on the eastern Kentucky districts because of data availability. Schools within the districts voluntarily participated in the program as did the teachers at those schools. Consequently, not all teachers or schools in a selected district participated.

In the AMSP, the higher education faculty designed and delivered training programs for K-12 teachers of math and science. The programs covered, for example, content training in algebra, geometry, physics, and biology. The training programs were offered in a variety of settings. In

some cases, K-12 teachers traveled to the institution of higher education for training, but in most cases, the higher education faculty traveled to an Appalachian site accessible to K-12 teachers across multiple schools and districts. The training varied in terms of hours per session and the number of sessions offered for a particular course type (i.e., biology or algebra). Some of the training occurred during the regular semester, and other training took place in longer periods over the K-12 summer break. In a small number of cases, the sessions focused on content pedagogy. Our analysis does not distinguish between programs, and we only assess the effect on student math scores.

The central Appalachian region is especially interesting because of its poor, rural population and longstanding achievement gap between the more isolated rural schools and those in urban and less isolated areas of these states. The AMSP program was funded and developed on the implicit assumption that this achievement gap exists, in part, because teachers in central Appalachia are less prepared to teach math and science than are teachers in other areas. Given that teachers in the K-12 schools of this region are not adequately achieving good student outcomes and these teachers are already in the school systems, concentrating on in-service training offers an alternative to improving teacher quality strictly through recruitment and pre-service training.

### Data Collection Process

The data for our earlier work and for this article were collected from four sources: the local school districts, the Kentucky Department of Education (KDE), the Kentucky Education Professional Standards Board (EPSB), and the AMSP administrators. Kentucky, at the state level, did not collect information that allowed particular students to be matched to specific teachers over the time period of this professional development program and the time period covered in our analysis. To successfully confirm which students were in a particular teacher's classroom, we obtained the cooperation of local school districts in the Appalachian portion of the state that were the target of the AMSP professional development activities.

We invited all Appalachian districts to provide classroom roster data regardless of whether or not the school superintendents had officially agreed for their schools to participate in the partnership program. The roster data for each school and each school year listed the course, the teacher for the course, and all students who were enrolled in that course. Ten school districts in eastern Kentucky provided useable data for this project, although not all districts provided data for the same years. As a result, we have a mixed panel of data at the district level.

The study then required matching data from several state-level administrative databases to the class roster data.

The matching process was complicated by the fact that the roster data were gathered for district-only use and did not include common student and teacher identification numbers which could be easily matched to state agency data, and none were developed with the idea of being used for evaluation purposes. Much of the matching required use of names and birth dates or other person-specific characteristics to make the match. Recently, Kentucky has assigned unique student identifiers that will allow the type of student-teacher matching described in this article.

KDE provided individual student demographic characteristics and test score data for this analysis. We were provided with test score data prior, during, and after the student's having a teacher who received the professional development. From these data we also collected student characteristics such as gender, race, and whether the student received free and reduced-price lunch. EPSB provided teacher-level data from the 2000-2001 until the 2007-2008 school year. We explicitly use teacher experience and highest degree achieved. All other characteristics of the teachers, such as Praxis scores, gender, and race, are time invariant and were captured in the teacher fixed effects, as well as in the matching model described below. We also had school-level characteristics that were time varying such as school-level average test scores, total enrollment per pupil, spending, and student-teacher ratios.

The last piece of data for our evaluation came from the AMSP staff. The AMSP program provided us with data on which teachers took the professional development, in what type of activity they participated, and in which years and for how many hours they were involved. The teachers in our study mostly came from small, rural districts, typically with a single high school and middle school and multiple elementary schools. The schools that were sampled were in small county-based school districts and from districts located in cities within counties, usually the county seat. Nine of the 10 districts in this sample were county-based. Of these 10 districts, six districts formally participated in AMSP activities while four districts did not officially participate. These districts are similar to one another in income, education, and population density, although they are more isolated culturally and geographically than rural districts outside of Appalachia. Regardless of official district-level participation, teachers in all districts were permitted to enroll in the training activities, and in the non-participating districts, teachers crossed district lines for the training activities even though their superintendent had not formally joined AMSP. AMSP fulfilled the Kentucky teacher requirement to participate in a minimum of four days of professional development annually. Participation rates tended to be higher in those districts whose superintendents agreed to participate. Ten percent of the teachers in our analysis who took the professional development were from non-AMSP districts. All protocols

for data collection were approved in advance by the Institutional Review Board at the University of Kentucky.

## Sample Description

For this study, we use student-level testing data from the 2000-2001 to 2010-2011 school years. In the school years 2000-2001 to 2004-2005 Kentucky tested students in different grade levels with different subject tests. The state also used two different standardized tests. The Kentucky-designed test was the Commonwealth Accountability Testing System (CATS), and the state also used the Comprehensive Test of Basic Skills (CTBS/5), a nationally norm-referenced test. Prior to 2005-2006, the state tested math in grades 3, 5, 6, 8, 9, and 11; reading in grades 3, 4, 6, 7, 9, and 10; and science in grades 4, 7, and 11. A revised iteration of testing was introduced in the 2006-2007 school year when Kentucky started testing all students in grades 3 to 8 in math and reading, math and science in grade 11, and reading in grade 10. The school year 2005-2006 was a transition year between the old and new testing systems and resulted in less reliable test score data. The focus of this study is on student achievement in math. This testing schema limits the number of years we can observe student math achievement scores. It also required us to use reading and science scores as lagged independent variables along with nonconsecutive year math lags, for which we adjust in the models below. Ultimately, our elementary sample includes student math outcomes in grade 5 for 2002-2003 through 2004-2005 and grades 4 and 5 for 2006-2007 through 2010-2011. Our middle school sample includes student math outcomes in grades 6 and 8 for 2002-2003 through 2004-2005 and grades 6, 7, and 8 for 2006-2007 through 2010-2011.

As in many states, the scaling of the tests also changed over the examined period. The change in the state's test in 2007 was such that the scale of scores in years prior could not be reliably reconciled with those of 2007 and onward. In addition, each grade-level test involved scores with different scales. A 500 on a fifth-grade math test was not designed to be equivalent to a 500 on an eighth-grade math test. Therefore, grade levels must be examined separately for evaluation purposes. Given the changing scale of the test score data over the time period we observe and the multiple exams (state and national), we convert raw test scores to Z-scores based on state averages for each grade. Z-scores are a nonlinear transformation and are frequently used for standardizing student test score data across multiple exams and scaling challenges.

Table 1 displays descriptive statistics for the analysis sample below.[1] These statistics are reported at the student,

[1]These statistics are reported for the sample used to estimate Equation 3. The data are drawn from a larger set of student and teacher characteristics used to estimate Equations 1 and 2 per Barrett, Butler, & Toma (2012).

school, and teacher levels. Several variables are worth discussion. The percentage of students in the sample who were assigned to a teacher who participated in AMSP professional development is higher than may be expected. However, it is important to note that the methodology used to control for selection bias into the program was designed at the teacher level, where presumably each teacher can then have a different number of students to which they are assigned. The number of students in the sample who are eligible for free and reduced-price lunch status is high, but it is lower than the overall percentage in the schools they attend. Also, approximately 97% of the sample is White. These figures are not unexpected given the geographical area where the program takes place. The math index is calculated based on a formula developed by the KDE. It identifies eight performance levels and then calculates the percentage of students within each level multiplied by that level's weight. Those eight values are then summed to provide an index on a scale that ranges from 0-140. The teachers included in the sample have a slightly lower average level of experience and fewer master's degrees than the population of teachers in the schools included in the analysis. Finally, we include the average spending per student at the school level.

## Construction of Teacher Sample and Selection Issues Related to Initial AMSP Assignment

As described above, given the opportunity, teachers were allowed to choose whether they participated in AMSP. We assume that this choice is not independent of either prior effectiveness or, by extension, earlier measures of student outcomes associated with each teacher. We also assume that teachers anticipate some perceived benefit with respect to their own effectiveness with future students. Such selection issues would cause even our estimates of exposure to an AMSP teacher in year $t$ to be biased.

Barrett, Butler, and Toma (2012) describe these issues in great detail in their study of the first-year AMSP impact. Their solution, which we employ exactly here, is to create an analysis sample of teachers matched based on their propensity to receive AMSP treatment. As they note, propensity scores will generate unbiased treatment impact estimates if the observable variables used to generate the matches are also those related to the outcomes of interest. However, as they describe, and we have noted here, eligible teachers are likely to have elected AMSP based at least in part on one largely unobserved variable directly related to student outcomes—namely, prior underlying teacher effectiveness.

The solution proposed by Barrett, Butler, and Toma (2012) employs a two-step process to generate a pool of control teachers whose students are compared to those students of an AMSP treatment teacher at time period, $t$. This pool is generated by measuring individual teacher effects on

Table 1

*Descriptive Statistics for the Analysis Sample*

| Characteristics | Mean | Standard Deviation |
|---|---|---|
| **Students** | | |
| Percent of Students with AMSP Teacher | 51.87 | 49.21 |
| Percent Free or Reduced Lunch | 60.85 | 48.81 |
| Percent Female Students | 49.46 | 50.00 |
| Percent Asian Students | 0.23 | 4.81 |
| Percent Black Students | 2.13 | 14.43 |
| Percent Hispanic Students | 0.46 | 6.76 |
| Percent Native American Students | 0.03 | 1.62 |
| Average Class Size | 24.47 | 6.69 |
| **Selected Schools** | | |
| Math Index Score | 73.56 | 13.81 |
| Percent Free or Reduced Lunch | 68.06 | 26.03 |
| Student-Teacher Ratio | 15.40 | 1.93 |
| Number of Students | 557.06 | 151.63 |
| Percent of Teachers with Master's Degree | 75.45 | 11.48 |
| Average Years of Experience of all Teachers | 12.02 | 1.80 |
| Spending Per Student | $5,452.88 | $853.18 |
| **Selected Teachers** | | |
| Highest Degree Earned[a] | 3.39 | 1.05 |
| Years of Experience | 11.33 | 8.10 |
| Experience Squared | 194.03 | 224.11 |
| Percent in Elementary School | 20.35 | 40.26 |

*Note.* Number of observations: 18,944 students. [a] Highest degree is coded as follows: 1=Bachelors, 2=5th year, 3=Planned 6th year, 4=Master's, 5=Rank 1.

student test scores, *Z*, *k* years prior to the teacher's entering the AMSP treatment. We estimate:

Equation 1

$$(Z_{it-k} - Z_{i,t-(k+1)}) = \beta_0 + \beta_2 D_{i,t-k} + \beta_3 TC_{ijt-k} + \beta_4 SC_{it-k} + \theta_j + c_j + u_{it-k}$$

where *D* is a vector of student characteristics, *TC* is a vector of time-varying teacher characteristics, and *SC* is a vector of school characteristics. Equation 1 includes a teacher fixed effect, $\theta_j$ with $c_j$ representing the teacher-clustered error term. The estimates of this fixed effect, $\widehat{\theta}_j$ across all available years are recovered as estimates of each teacher's

underlying effectiveness, or "value-added" *up until the last possible pre-AMSP year*, denoted here as *t-1*.[2] Equation 1

[2]Similar value-added models are used in the literature and, increasingly, in practice to gauge teacher outcomes. Scholars continue to debate the use of value-added, especially with respect to issues of bias and efficiency in the proper specification of the models involved. In particular are objections raised by Rothstein (2010) related to non-random teacher-student sorting. Several studies (e.g. Guarino, Reckase, & Wooldridge, 2011; Kane & Staiger, 2008; Koedel & Betts, 2011) have explored these issues, and while the debate continues, the use of value-added in both scholarship and policy has grown.

represents our specification of perhaps the most generally used model on teacher effectiveness in the literature thus far (Guarino, Reckase, & Wooldridge 2011; Koedel & Betts, 2011).

Next, we follow Barrett, Butler, and Toma (2012) by including $\widehat{\theta}_j$ in our estimation of the propensity of an individual and eligible teacher to elect AMSP:

Equation 2
$$PD_{jt} = \alpha_1 TC_{jt} + \alpha_2 SC_{it} + \alpha_3 \widehat{\theta}_{j,t-1} + \varepsilon_t$$

where $PD$ is a binary indicator of participation in AMSP of the entire sample of teachers, the vectors $TC$ and $SC$ are the observable teacher and school characteristics, and $\widehat{\theta}$ is prior teacher effectiveness. The vector $TC$ includes dummy variables for teacher race and gender, as well as for initial Praxis scores for each teacher; it also includes the time-varying experience and education levels for each teacher.

Equation 2 is estimated as a probit, and we use the subsequent $PD$ predictions to generate propensity scores that represent the estimated probability of a given teacher's AMSP participation. We provide results of estimating

Equation 2 in Table 2, and see Barrett, Butler, and Toma (2012) for additional computational issues. We note here that several school factors were statistically significantly related to participation: Teachers from smaller schools were more likely to participate, as were those with more free/reduced lunch students, lower per-student expenditures, and lower school-level math index scores. Since the AMSP was explicitly targeted to lower performing and particularly at-risk schools in the Appalachian area, these are exactly the relationships we would expect a priori. In addition—and of direct relevance for our interpretation of one pattern below—even after controlling for these school-level characteristics and controlling for teachers' observable credentials, our prior teacher effectiveness estimate significantly and negatively predicted participation. Teachers who were less effective before the AMSP program were more likely to accept the program offer. This is, as with the school characteristics, evidence that the program's targeting worked as the developers intended.

After estimating teachers' probability of participating via Equation 2, non-AMSP teachers were then matched to AMSP teachers based on their estimated propensity score

Table 2

*Results from Propensity Score Prediction of Teacher AMSP Participation (Probit)*

| Independent Variables | Coefficient | | Standard Error |
|---|---|---|---|
| **Teacher Characteristics** | | | |
| Estimated Previous Effectiveness | -0.363 | *** | 0.078 |
| Max Degree Held | 0.028 | | 0.029 |
| Experience | 0.011 | | 0.014 |
| Experience Squared | -0.001 | | 0.0004 |
| | | | |
| **School Characteristics** | | | |
| Average Experience | 0.008 | | 0.016 |
| Enrollment (000's) | -0.145 | *** | 0.020 |
| Percent Master's | -0.001 | | 0.002 |
| Percent of FRP Students | 0.437 | *** | 0.131 |
| Expenditure per Student (000's) | -0.145 | *** | 0.025 |
| Student-Teacher Ratio | 0.139 | *** | 0.016 |
| Student-Computer Ratio | -0.007 | | 0.023 |
| Math Index | -0.014 | *** | 0.002 |
| N | 5830 | | |
| Log Likelihood | (-814.25) | *** | |

*Note*. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. See also Barrett, Butler and Toma (2012) for these estimates and additional computational details.

and we generate our sample of treated and control teachers at time period, $t$. Matching was done by employing nearest neighbor matching without replacement and stratifying on year so that AMSP teachers were matched to only non-AMSP teachers that could have participated in the same year. The students enrolled in the classrooms of these treated and control teachers at $t$ represent our students of interest, whose outcomes we measure in the years following $t$. Thus our final analytic sample contains two distinct groups of students: those who had AMSP teachers at $t$ and the matched control group of students whose teachers did not participate.[3]

## Modeling Sustained AMSP Impacts

The Barrett, Butler, and Toma (2012) evaluation of the AMSP program found generally positive impacts[4] on student outcomes in the academic year in which teachers participated. In this article, we extend those results and consider whether the program's effects lingered beyond the treatment year. As described above, AMSP was primarily a content-based development program. Teachers learned STEM content directly from experts in their subject area. The program's objective was to improve student outcomes by improving the content knowledge of the teachers who instructed them. One measure of that objective is whether students of AMSP teachers learn more in the subject area than students with non-participating teachers, and a second measure is how long that advantage persists over time.

## Formulation of Research Question and Primary Model of Interest

Given student achievement as the primary outcome of interest, there are two specific ways to formulate the question of sustained impacts. The first is the AMSP program's longer term impact on teachers themselves. In this version of the question we would ask: Do the test scores of students assigned to teachers previously trained in AMSP improve in subsequent years? A second formulation would be: Do the outcomes of a student who had an AMSP teacher at one time continue to improve after he or she has left the classroom of the AMSP teacher?

The first formulation is directed toward the question of sustained effectiveness for the teacher who was trained. The second formulation concerns the durability of the initial

treatment impact on students. Both are relevant formulations for education policy. If the primary goal of any professional development program is, as the name implies, to extend a teacher's capabilities, it would seem that a minimum criterion for judging the AMSP would be evidence that teacher improvement can translate to new students in the yearly cycle of student-teacher classroom assignments. Even if a teacher's own effectiveness waned, however, it is still possible that whatever the benefit of the initial AMSP program in terms of student outcomes, the students whose learning was enhanced in the year their teacher was trained may continue to make progress later on.

From an empirical standpoint, both formulations present difficult estimation problems. The difficulties in addressing the first formulation, unfortunately, appear to be intractable. As Barrett, Butler, and Toma (2012) argued in detail, and as we note above, we assume that teachers' decisions to participate in AMSP were at least correlated (if not outright confounded) with their own effectiveness as an instructor. Such effectiveness we assume to have a systematic but unobserved component even after adjustment for teacher observables like experience and certification. Barrett, Butler, and Toma (2012) employed a novel estimation strategy in addressing this initial form of teacher selection, and the matching process described above is an extension of that approach.

However, the earlier Barrett, Butler, and Toma (2012) approach deals only with teachers in the initial AMSP participation year. Here lies a particular problem for estimating the sustained impact of the program on an AMSP teacher's future students if teacher effectiveness—measured in terms of student test scores—is related to new students' assignment to teachers. In particular, if students are assigned to teachers in part because the teacher was exposed to AMSP in the past, we cannot estimate the AMSP impact on new students without some way (i.e., an instrumental variable) to break that relationship in expectation.

The second formulation—considering whether individual students' test scores improve in the years after they had an AMSP teacher—presents similar but surmountable difficulties under the assumption that the Barrett, Butler, and Toma (2012) strategy minimizes bias for estimates of the AMSP impact in the year in which teachers participated. We make this formulation here. First, we denote $t$ as the year a student's teacher participated in AMSP or was a member of the control group. The first test scores on which we might observe an AMSP impact are in the same year, when student $i$ is assigned to the AMSP teacher or a control teacher. Thus, $t+1$, $t+2$, and $t+3$ denote outcomes 1, 2, and 3 years after student $i$ was assigned to the AMSP teacher, respectively. Formally, we wish to estimate the model:

---

[3]Please see Appendix A for the pre- and post-match balance statistics of the AMSP participating teachers and their comparison group.

[4]We use the word "impact" throughout the article for ease of exposition. We note, however, that identification of causal impacts/ effects generally requires experimental designs, while our design here is quasi-experimental.

Equation 3

$$Z_{i,t+1,t+2,t+3} = \beta_0 + \delta AMSP_{ijt} + \beta_1 Z_{i,t-1} + \beta_2 D_{it} + \beta_3 TC_{ijt} + \beta_4 SC_{ist} + \theta_s + c_s + u_{it}$$

where student $i$'s outcomes $Z$ in year $t+1$, $t+2$, $t+3$ after assignment to a teacher $j$ who received AMSP training or who were in the control group in year $t$ are also a function of prior achievement in the last pre-treatment year ($t$-$1$), student demographics $D$, and time-varying teacher and school characteristics $TC$ and $SC$, respectively.[5] The model also includes fixed effects  for each school. Finally, $c$ represents a school-clustered error term and $u$ the remaining error component that is idiosyncratic to student $i$. The lagged score term $Z_{t-1}$ is actually a vector of not only the prior outcome subject (mathematics) but also science and reading to serve as additional controls. To enter the sample, students must have had a score from at least one exam at $t$-$1$. For those for whom we do not have the lagged dependent variable, we use the lagged science or reading as proxies, along with dummy variables to indicate them as additional controls. Further below, we provide specifications that exclude these students altogether, with little difference in the observable results.

In Equation 3, the parameter of interest is $\delta$, the difference in outcomes between students who had an AMSP teacher and those whose teachers were not in the program at the same past point $t$. The AMSP designation remains with each student for all times after $t$ regardless of future student, teacher, or school changes that occur. Thus, the subscripts in Equation 3 for the $D$, $TC$, and $SC$ vectors remain fixed at $t$. Also fixed, albeit at the last pre-treatment year $t$-$1$, is a baseline outcome $Z$, so that the model in Equation 3 represents growth from the last pre-AMSP year to some post-AMSP period $t+1$, $t+2$, or $t+3$.

This approach, which is analogous to the "intent-to-treat" (ITT) parameter in randomized control trials of policy interventions, necessarily restricts our question to the future impact of initial assignment to an AMSP teacher in the past. Students who are assigned to the AMSP teacher in years beyond $t$, whose assignment to teachers after the teachers are trained could be a function of their AMSP experience. As long as the initial assignment strategy discussed above minimizes selection bias in the year the student was actually exposed to an AMSP teacher, we do not worry about bias in our estimates *of that initial exposure impact* on subsequent outcomes. Since this study is not experimental, however, we cannot fully rule out such bias, and thus note that a cautious consideration of $\delta$ is as an observed rather than a causal difference in outcomes.

This specification also leads to interpretation problems

_____

[5]Teacher characteristics include teacher experience and education level; teacher race, gender and initial Praxis scores are explicitly included in the propensity score match described above.

to the extent that any systematic differences between students who did and did not have an AMSP teacher at $t$ are subsumed in the original AMSP designation in Equation 3. What Equation 3 requires us to do, in other words, is attribute all outcome differences to be part of the treatment regardless of whether it was structurally a component of the AMSP development program. In the most extreme case, students could be differentially assigned based on earlier AMSP status to teachers of differing effectiveness later on. More generally, we might worry that, for example, principals were more likely to assign students who had an AMSP teacher in year $t$ to a better teacher in the following year. If so, our estimate of $\delta$ in Equation 3 will overstate the true long-term benefit of AMSP as a professional development design even if it provides an unbiased estimate of differential assignment at $t$. On the other hand, if students were assigned to less effective teachers based on having been taught by an AMSP teacher in year $t$, then our estimate of  will understate the true impact of the professional development. We tested for potential systematic assignment by estimating the relationship between a teacher's prior value added and prior student exposure to an AMSP teacher while controlling for student and school characteristics. Results of this estimation suggest that principals are not systematically assigning students to teachers based on the student's prior exposure to AMSP and the teacher's prior effectiveness.

**Results**

**Main Results**

We estimate Equation 3 across a series of cross-sectional data composed of the analytic sample as it proceeds in the years following $t$. For example, to consider the AMSP impact in the first follow-up year after teachers were trained, we estimate Equation 3 on students with AMSP and non-AMSP teachers, replacing the outcome in the treatment year with the next year's outcome at $t+1$. We estimate similar models for $t+2$ and $t+3$, the latter of which represents the last year in which we can plausibly track students before the sample disintegrates as the youngest students move into high school, where state testing was less frequent.

Table 3 depicts our primary results from estimates of Equation 3. The table moves consecutively from left to right reporting outcomes from year $t$, which are substantively similar to those in Barrett, Butler, and Toma (2012), and on to $t+1$, $t+2$, and $t+3$. Since the outcome is standardized by grade and year against the state averages, the coefficients are directly interpretable in standard deviations. The key finding in Table 3 is that not only did teachers' AMSP participation have a positive impact on students in the year

Table 3

*Main Results of AMSP Impacts on Students over Time*

| VARIABLES | BASE YEAR (t) | t+1 year | t+2 year | t+3 year |
|---|---|---|---|---|
| **AMSP Teacher at** *t* | 0.109*** | 0.088* | -0.070 | 0.042 |
| | (0.0388) | (0.0443) | (0.0712) | (0.0436) |
| **Math t-1** | 0.434*** | 0.466*** | 0.447*** | 0.392*** |
| | (0.0211) | (0.0187) | (0.0268) | (0.0216) |
| **Reading t-1** | 0.193*** | 0.168*** | 0.145*** | 0.174*** |
| | (0.0110) | (0.0098) | (0.0179) | (0.0118) |
| **Science t-1** | 0.175*** | 0.167*** | 0.179*** | 0.231*** |
| | (0.0103) | (0.0119) | (0.0172) | (0.0187) |
| **Free/Reduce Lunch** | -0.124*** | -0.155*** | -0.235*** | -0.179*** |
| | (0.0191) | (0.0225) | (0.0248) | (0.0356) |
| **Female** | -0.015 | 0.035** | 0.046** | 0.034 |
| | (0.0213) | (0.0160) | (0.0199) | (0.0218) |
| **Asian** | 0.096 | 0.356*** | 0.733*** | 0.469** |
| | (0.2187) | (0.1040) | (0.2043) | (0.1912) |
| **Black** | -0.132*** | -0.042 | -0.119* | -0.293*** |
| | (0.0389) | (0.0528) | (0.0662) | (0.0879) |
| **Hispanic** | -0.034 | -0.084 | -0.049 | -0.087 |
| | (0.0875) | (0.1434) | (0.1900) | (0.1924) |
| **Native** | -0.284* | 0.212 | -0.008 | 0.136 |
| | (0.1607) | (0.5209) | (0.0751) | (0.2827) |
| **Other** | -0.043 | -0.086 | 0.176 | 0.123 |
| | (0.0768) | (0.1033) | (0.1640) | (0.1491) |
| **School Controls** | ✔ | ✔ | ✔ | ✔ |
| **Teacher Controls** | ✔ | ✔ | ✔ | ✔ |
| | | | | |
| **Constant** | -1.422* | -0.378 | -0.275 | -0.145 |
| | (0.7438) | (0.6902) | (1.3023) | (1.0345) |
| **Observations** | 18,944 | 13,805 | 7,961 | 7,949 |
| **R-squared** | 0.368 | 0.418 | 0.357 | 0.392 |
| **Number of schools** | 78 | 78 | 71 | 67 |

*Note.* Robust standard errors in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$. Models include controls for school characteristics (size, mean teacher experience, proportion of teachers with master's degrees, proportion of free/reduced lunch students, students/teacher ratios, and average math index scores) as well as teacher degree and experience. Also included are indicators for missing student prior scores by subject as described in text.

of participation, but the effect on students appears to have lingered into the following year as well. Not only does it linger, but the impact remains remarkably strong relative to participation year—nearly 0.09 standard deviations compared to 0.11 standard deviations in the participation year. Due to rounding the estimate of the *t+1*, impact is only significant at p<0.10, but it is very close to the common p<0.05 threshold.

Per our discussion of our analytical framework above, the coefficient on follow-up years should be interpreted with caution. These estimates address the somewhat narrow question, "What is the effect of AMSP participation on test scores of students in the years following the AMSP program?" Since this formulation requires the AMSP designation at *t* to follow students into *t+1* and beyond, it implies that any factors differently affecting students whose teachers had AMSP at *t* from students whose teachers did not have AMSP are subsumed in the AMSP designation. The more time passes, the less this designation may meaningfully differentiate between students with these varied teacher experiences unless the AMSP impact at *t* and *t+1* is actually powerful contribution to student learning.

## Additional Results and Robustness Checks

**Sample stability over the analysis period.** One potentially problematic issue comes with the change in the sample over time. Although we tracked students after *t*, not all students had valid test scores in later years, primarily because they reached terminal or non-tested grades as described in the data section above. Our inclusion of a grade-level dummy (elementary referenced to middle school as of time *t*) should account for any systematic differences that could drive *t+1* and later years forward. If, for example, the program is simply less effective for elementary students, we might worry that null results in *t+2* and *t+3* are driven by the fact that it is precisely these students who still have test scores in later years for us to study. To consider this issue we simply estimate the models on the sub-set of students in schools for whom we should expect scores in later years. These results are included as the first column labeled "reduced sample" under each treatment header in Table 4. These results are very consistent with our estimates from the full sample.

**Different metrics for lagged outcomes.** Given our reliance on the pre-treatment vector to anchor each post-treatment year, another problem arises for those students who, as described above, tested in different subjects at *t-1*. As noted, we have test scores for at least one subject at *t-1*, mathematics, reading, or science, for all students who have mathematics test scores at *t* or beyond. To avoid excluding from the analysis students without mathematics scores in *t-1*, our main results in Table 3 include indicators for these students as part of the student-level variables in Equation 3. That specification in the main results should control for different lagged achievement levels for these students. To confirm that the results in Table 3 are not driven in some way by the inclusion of these students even after including indicator variables in the main results, we simply estimate Equation 3 on the sub-sample of students for whom we have no missing pre-AMSP scores in mathematics. The results are remarkably consistent with the main results, as indicated by the columns (labeled "all lags") for each year in Table 4. Thus the lingering treatment effect in *t+1* appears not to be driven by sample instability in later years, or by our differential inclusion of prior scores.

**Repeated exposure to an AMSP teacher.** The issue of repeated exposure to AMSP teachers over time raises the most direct challenge to our ability to estimate the lingering impact of receiving an AMSP teacher at *t*. Although more than 80% of our students only received an AMSP teacher once (at *t*), a considerable number of students (18%) actually had at least one teacher after *t* who trained in AMSP. If AMSP impacts are positive in *t*, we would expect a priori that a second year with a similarly trained teacher would generate positive impact estimates at that time. If a student's probability of assignment to a new AMSP teacher at, say,

*t+1* is partly a function of assignment to an AMSP teacher at *t*, then our results in *t+1* are driven by this repeated exposure and not by linger impacts of the original teacher's participation at *t*.

To check for this we estimated Equation 3 with additional controls that account for students who received multiple treatments. If our positive estimate of exposure to an AMSP teacher at *t* were generating positive results in *t+1*, *t+2*, or *t+3* because some students essentially got a second chance to learn from an AMSP-trained teacher, we would expect the significant coefficient on our estimate of the impact at *t* to be diminished or eliminated altogether by including these controls. As explained above, we are not confident in our ability to model any post-*t* selection mechanisms that would sort teachers and students together in a way related to earlier AMSP exposure, so we cannot interpret those controls as treatment effects in their own right. Their inclusion should, however, give us an indication of how greatly our estimates of the sustained effect of initial exposure to an AMSP teacher at *t* are being driven by gains made by the students who actually had sustained exposure to an AMSP teacher.

Table 5 provides the results of estimating Equation 3 with additional controls for whether the student received an AMSP teacher after *t*. Table 5 also includes this specification for the two additional robustness checks described in the preceding sub-sections. All results are comparable to those in Tables 3 and 4. These similarities provide some assurance that our primary results—in particular, the lingering AMSP impact in *t+1*—are not driven by the sub-sample of students who ultimately received more than one AMSP teacher.

## Discussion and Conclusion

In the remote, rural environment that we examine here, schools rely almost exclusively on teachers who were trained nearby (Fowles et al., 2014). Once hired, teachers either remain in their original school or exit the workforce entirely (Cowen et al., 2012). In Appalachian Kentucky, there are few additional opportunities to draw outside professionals into the classroom, and dismissing existing teachers based on low performance may substantially reduce the pool of available teachers to staff these schools. Research on teachers in rural locales remains underdeveloped in academic literature, and studies of other areas of educational reform—school choice, accountability, and academic standards, to name the most prominent—are similarly geared toward urban and suburban systems.

There is a long history of an educational achievement gap in the central Appalachian region. This gap is especially apparent in the areas of science and mathematics. Few students in the central Appalachian region score at the proficient level or above in mathematics and/or science as defined by the assessment standards developed in

Table 4

*AMSP Impacts for Alternative Samples*

| VARIABLES | BASE YEAR (t) | | t+1 year | | t+2 year | | t+3 year | |
|---|---|---|---|---|---|---|---|---|
| | reduced sample | All lags | reduced sample | All lags | reduced sample | All lags | reduced sample | All lags |
| AMSP Teacher at *t* | 0.116*** | 0.092*** | 0.086* | 0.098** | -0.069 | 0.011 | n.a. | 0.024 |
| | (0.0394) | (0.0338) | (0.0457) | (0.0467) | (0.0709) | (0.0841) | | (0.0398) |
| Math t-1 | 0.431*** | 0.428*** | 0.467*** | 0.467*** | 0.441*** | 0.463*** | | 0.391*** |
| | (0.0218) | (0.0216) | (0.0194) | (0.0190) | (0.0269) | (0.0278) | | (0.0224) |
| Reading t-1 | 0.192*** | 0.189*** | 0.169*** | 0.158*** | 0.149*** | 0.123*** | | 0.163*** |
| | (0.0113) | (0.0101) | (0.0100) | (0.0099) | (0.0183) | (0.0155) | | (0.0111) |
| Science t-1 | 0.177*** | 0.187*** | 0.166*** | 0.186*** | 0.177*** | 0.200*** | | 0.246*** |
| | (0.0105) | (0.0101) | (0.0118) | (0.0121) | (0.0173) | (0.0197) | | (0.0170) |
| Free/Reduce Lunch | -0.124*** | -0.108*** | -0.153*** | -0.111*** | -0.244*** | -0.169*** | | -0.136*** |
| | (0.0197) | (0.0140) | (0.0225) | (0.0193) | (0.0251) | (0.0199) | | (0.0350) |
| Female | -0.011 | -0.025 | 0.039** | 0.031** | 0.050** | 0.017 | | -0.012 |
| | (0.0223) | (0.0154) | (0.0164) | (0.0144) | (0.0202) | (0.0166) | | (0.0233) |
| Asian | 0.098 | 0.026 | 0.423*** | 0.226* | 0.734*** | 0.733*** | | 0.442** |
| | (0.2183) | (0.2390) | (0.0874) | (0.1187) | (0.2049) | (0.2549) | | (0.2172) |
| Black | -0.120*** | -0.065 | -0.039 | -0.066 | -0.113* | -0.107 | | -0.185* |
| | (0.0402) | (0.0433) | (0.0551) | (0.0557) | (0.0674) | (0.0732) | | (0.1074) |
| Hispanic | -0.065 | 0.102 | -0.101 | -0.008 | -0.089 | 0.191 | | -0.112 |
| | (0.0941) | (0.0711) | (0.1462) | (0.1355) | (0.1923) | (0.1567) | | (0.2420) |
| Native | -0.322 | -0.138 | 0.200 | 0.995*** | -0.034 | 0.048 | | 0.246 |
| | (0.1959) | (0.2105) | (0.5202) | (0.0687) | (0.0753) | (0.0996) | | (0.3361) |
| Other | -0.033 | 0.019 | -0.094 | -0.010 | 0.175 | 0.202 | | 0.342* |
| | (0.0776) | (0.0820) | (0.1079) | (0.0943) | (0.1637) | (0.2264) | | (0.1907) |
| School Controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Teacher Controls | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Constant | -1.526* | -0.875 | -0.310 | 0.034 | -0.200 | -1.439 | | 0.251 |
| | (0.7698) | (0.6446) | (0.7530) | (0.6999) | (1.3138) | (1.7156) | | (1.2284) |
| Observations | 17,873 | 14,686 | 12,920 | 11,351 | 7,622 | 5,735 | | 5,855 |
| R-squared | 0.365 | 0.495 | 0.419 | 0.503 | 0.354 | 0.476 | | 0.479 |
| Number of schools | 67 | 69 | 67 | 68 | 67 | 62 | | 56 |

*Note.* Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Models include controls for school characteristics (size, mean teacher experience, proportion of teachers with master's degrees, proportion of free/reduced lunch students, students/teacher ratios, and average math index scores) as well as teacher degree and experience. Also included are indicators for missing student prior scores by subject as described in text.

Table 5

*AMSP Impacts with Alternative Samples and Controlling for Multiple AMSP Teachers*

| VARIABLES | t+1 year | | | t+2 year | | | t+3 year | | |
|---|---|---|---|---|---|---|---|---|---|
| | full w/ add'l part | reduced sample w/ add'l part | All lags w/ add'l part | full w/ add'l part | reduced sample w/ add'l part | All lags w/ add'l part | full w/ add'l part | reduced sample w/add'l part | All lags w/add'l part |
| AMSP Teacher at *t* | 0.079* | 0.076 | 0.100** | -0.071 | -0.077 | 0.002 | 0.049 | n.a. | 0.017 |
| | (0.0442) | (0.0460) | (0.0479) | (0.0725) | (0.0723) | (0.0893) | (0.0452) | | (0.0409) |
| t1_treatment | 0.010 | 0.014 | -0.014 | 0.080 | 0.149** | 0.065 | 0.001 | | 0.015 |
| | (0.0423) | (0.0464) | (0.0284) | (0.0749) | (0.0637) | (0.0797) | (0.0510) | | (0.0641) |
| t2_treatment | | | | 0.103 | 0.096 | 0.073 | 0.074* | | 0.022 |
| | | | | (0.2035) | (0.2111) | (0.3375) | (0.0435) | | (0.0533) |
| t3_treatment | | | | | | | 0.031 | | 0.054 |
| | | | | | | | (0.0452) | | (0.0378) |
| Math t-1 | 0.457*** | 0.456*** | 0.466*** | 0.438*** | 0.431*** | 0.463*** | 0.391*** | | 0.391*** |
| | (0.0199) | (0.0206) | (0.0190) | (0.0263) | (0.0262) | (0.0278) | (0.0203) | | (0.0224) |
| Reading t-1 | 0.160*** | 0.161*** | 0.158*** | 0.145*** | 0.149*** | 0.123*** | 0.172*** | | 0.162*** |
| | (0.0103) | (0.0104) | (0.0098) | (0.0171) | (0.0174) | (0.0156) | (0.0114) | | (0.0113) |
| Science t-1 | 0.178*** | 0.181*** | 0.186*** | 0.181*** | 0.180*** | 0.201*** | 0.230*** | | 0.246*** |
| | (0.0123) | (0.0126) | (0.0121) | (0.0181) | (0.0185) | (0.0198) | (0.0173) | | (0.0168) |
| Free/Reduce Lunch | -0.165*** | -0.164*** | -0.111*** | -0.236*** | -0.246*** | -0.168*** | -0.188*** | | -0.137*** |
| | (0.0204) | (0.0201) | (0.0192) | (0.0232) | (0.0233) | (0.0199) | (0.0346) | | (0.0348) |
| Female | 0.046*** | 0.050*** | 0.031** | 0.032 | 0.035 | 0.018 | 0.034 | | -0.012 |
| | (0.0151) | (0.0154) | (0.0144) | (0.0205) | (0.0210) | (0.0164) | (0.0210) | | (0.0233) |
| Asian | 0.423*** | 0.423*** | 0.226** | 0.704*** | 0.705*** | 0.735*** | 0.446*** | | 0.447** |
| | (0.0754) | (0.0756) | (0.1191) | (0.1897) | (0.1897) | (0.2550) | (0.1474) | | (0.2179) |
| Black | -0.089* | -0.087 | -0.066 | -0.173** | -0.168** | -0.107 | -0.326*** | | -0.184* |
| | (0.0524) | (0.0529) | (0.0558) | (0.0800) | (0.0816) | (0.0731) | (0.0909) | | (0.1077) |
| Hispanic | -0.086 | -0.100 | -0.007 | 0.032 | 0.000 | 0.192 | -0.106 | | -0.113 |
| | (0.1178) | (0.1189) | (0.1354) | (0.1901) | (0.1912) | (0.1565) | (0.1688) | | (0.2420) |
| Native | 0.074 | 0.072 | 0.992*** | 0.306 | 0.299 | 0.065 | 0.120 | | 0.240 |
| | (0.3674) | (0.3626) | (0.0682) | (0.1920) | (0.1897) | (0.1050) | (0.1967) | | (0.3392) |
| Other | -0.106 | -0.107 | -0.010 | 0.112 | 0.111 | 0.203 | 0.081 | | 0.342* |
| | (0.0924) | (0.0925) | (0.0943) | (0.1618) | (0.1618) | (0.2262) | (0.1425) | | (0.1905) |
| School Controls? | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Teacher Controls? | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ |
| Constant | -0.279 | -0.476 | 0.021 | -0.356 | -0.201 | -1.291 | 0.020 | | 0.414 |
| | (0.6953) | (0.7382) | (0.7054) | (1.1639) | (1.1646) | (1.6648) | (1.0049) | | (1.3339) |
| Observations | 13,805 | 12,920 | 11,351 | 7,961 | 7,622 | 5,735 | 7,949 | | 5,855 |
| R-squared | 0.341 | 0.340 | 0.503 | 0.337 | 0.335 | 0.477 | 0.369 | | 0.480 |
| Number of schools | 78 | 67 | 68 | 71 | 67 | 62 | 67 | | 56 |

*Note.* Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1. Models include controls for school characteristics (size, mean teacher experience, proportion of teachers with master's degrees, proportion of free/reduced lunch students, students/teacher ratios, and average math index scores) as well as teacher degree and experience). Also included are indicators for missing student prior scores by subject as described in text.

each state. Analysis of the assessment data consistently reveals lower performance at all K-12 levels for central Appalachian students when compared to the state averages and/or students from more affluent regions of the states. Enrollment in higher-level mathematics courses, including algebra II and calculus, is less than one-third the enrollment in lower level mathematics. Introductory science courses have more than three times the enrollment of higher-level courses such as chemistry and physics. Although 25 of the central Appalachian school districts report offering some type of AP or dual credit program, enrollment in these programs is non-existent in many schools.

The AMSP program was funded and developed on the assumption that achievement gaps may exist, in part, because teachers in central Appalachia are less prepared to teach math and science than teachers in other areas. Demand for teachers often exceeds supply, especially in low-performing schools such as those in central Appalachia. These school districts report a major problem with attracting and maintaining a staff of highly qualified mathematics and science teachers.

In these policy and scholarly contexts, our results provide some direction for investing both future resources and new lines of research. We have shown that targeted, intensive training of teachers in Appalachian schools can have positive impacts on student outcomes not only in the year each child's teacher received new professional development, but in at least an additional year beyond that training as well. That our study also indicates that the lowest performing teachers in some of the least advantaged schools were especially likely to opt into these new trainings—and thus to benefit from them—suggests professional development in these areas can indeed reach those who need it the most. This finding implies that schools in remote, rural locales may be well served by investing resources in improving the skills of their existing workforce rather than making wholesale changes to their teaching staff.

There are broader lessons for research and policy, both within and outside of rural schools systems. Kentucky, like many other states, has invested substantial resources in the creation of education databases since 2004. Although these new data provide useful information that the public can use to assess raw scores for particular schools, the true value goes far beyond average school scores. Systematic analysis of any number of major initiatives in educational policy depends on the quality and availability of these data. One explanation for the relative scarcity of large-sample statistical work on rural schools is that until the development of statewide data systems, only large urban areas such as New York City, Chicago, and metropolitan areas of Texas and Florida had developed such databases on their own (Miller, 2012). In this article we have focused on the sustained impact of teacher training on student outcomes,

relying directly on ongoing access to several sources of data to complete the project. Absent such access, researchers and policymakers alike would be unable to monitor or confirm improvements in rural districts like those we detect here.

Teacher participation in professional development is not a new phenomenon. Only in recent years, however, can the results of such participation be directly tied to student outcomes. Establishing the relationship between teacher quality and student achievement requires accurate information on teacher and student characteristics, which includes a reliable indicator connecting students to teachers to whom they are assigned. Also critical is the ability to track students and teachers between classrooms and schools over time. Constructing a system to organize that information is costly, but it is inherently necessary to answer research questions that establish links between teacher preparation and student achievement.

Regular and systematic oversight of professional development may be needed to ensure long-term improvement not only in rural schools but in other locales as well. States typically require professional development or continuing education for their teachers as part of ongoing licensure procedures. In most instances, there has been an implicit assumption that the training improved teacher content knowledge, making them more effective in the subject and ultimately improving the achievement of their students in the area. As state budgets for education grow at lesser rates or even decline, it becomes ever more important to examine whether particular training programs improve student learning and how long the impact of effective programs might last. Of special relevance to this article, it is important to consider whether gains in student achievement, if they exist, persist in their own right or need renewed reinforcement over time. We have shown that relative to other students, those with teachers trained in the AMSP program not only realized substantial achievement gains in the year they were assigned to AMSP teachers, but maintained that advantage in at least one follow-up year.

This study confirms, in summary, the importance of the partnerships that develop between the professional development providers in rural locales and the data agencies in outside jurisdictions. In the case of the AMSP, effectiveness could be evaluated only because of cooperation between local school districts and state agencies that provided the data and those who administered the teacher professional development. The fact that the program also depended on faculty at state universities to administer training in the subject area only broadens the scope of this collaboration. As a result of the partnership between the service providers, the agencies that collected the data and the scholarly community, both the AMSP program and state and local policymakers have learned that a targeted, content-based teacher development program can have sustained

impacts on student outcomes in rural communities. This finding holds policy implications for the design of future professional development as well as implications for the methods by which these programs are evaluated.

### References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago Public Schools. *Journal of Labor Economics*, *25*, 95-135. doi:10.1086/508733

Arnold, M. L., Newman, J. H., Gaddy, B. B., & Dean, C. B. (2005). A look at the condition of rural education research: Setting a direction for future research. *Journal of Research in Rural Education*, *20*(6). Retrieved from http://jrre.vmhost.psu.edu/wp-content/uploads/2014/02/20-6.pdf

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a practice-based theory of professional education. *Teaching as the learning profession: Handbook of policy and practice*, *1*, 3-22.

Ballou, D., & Podgursky, M. (1995). Recruiting smarter teachers. *Journal of Human Resources*, *30*, 326-338. doi:10.2307/146122

Barrett, N., Butler, J. S., & Toma, E. F. (2012). Do less effective teachers choose professional devlopment does it matter? *Evaluation Review*, *36*, 346-374. doi:10.1177/0193841X12473304

Borman, G. D., Gamoran, A., & Bowdon, J. (2008) A randomized trial of teacher development in elementary science: First-year achievement effects. *Journal of Research on Educational Effectiveness*, *1*, 237-264. doi:10.1080/19345740802328273

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2005). Explaining the short careers of high-achieving teachers in schools with low-performing students. *American Economic Review*, *95*(2), 166-171. doi:10.1257/000282805774669628

Carnoy, M., & Loeb, S. (2002). Does external accountability affect student outcomes? A cross-state analysis. *Educational Evaluation and Policy Analysis*, *24*, 305-331. doi:10.3102/01623737024004305

Chester, M. D., & Beaudin, B. Q. (1996). Efficacy beliefs of newly hired teachers in urban schools. *American Educational Research Journal*, *33*, 233-257. doi:10.3102/00028312033001233

Cowen, J. M., Butler, J. S., Fowles, J., Streams, M. E., & Toma, E. F. (2012). Teacher retention in Appalachian schools: Evidence from Kentucky. *Economics of Education Review*, *31*, 431-441. doi:10.1016/j.econedurev.2011.12.005

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management*, *30*, 418-446. doi:10.1002/pam.20586

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualiztions and measures. *Educational Researcher*, *38*, 181-199. doi:10.3102/0013189X08331140

Desimone, L. M., Porter, A. C., Garet, M. S., Yoon, K. S., & Birman, B. F. (2002). Effects of professional development on teachers' instruction: Results from a three-year longitudinal study. *Educational Evaluation and Policy Analysis*, *24*, 81-112. doi:10.3102/01623737024002081

Foster, J. M., Toma, E. F., & Troske, S. P. (2013). Does teacher professional develiopment improve math and science outcomes and is it cost effective? *Journal of Education Finance*, *38*(3), 255-275.

Fowles, J., Butler, J. S., Cowen, J. M., Streams, M. E., & Toma, E. F. (2014). Public employee quality in a geographic context a study of rural teachers. *The American Review of Public Administration*, *44*, 503-521. doi:10.1177/0275074012474714

Garet, M. S., Porter, A. C., Desimone, L., Birman, B. F., & Yoon, K. S. (2001). What makes professional development effective? Results from a national sample of teachers. *American Educational Research Journal*, *38*, 915-945. doi:10.3102/00028312038004915

Goldhaber, D., & Hansen, M. (2009). National board certification and teachers' career paths: Does NBPTS certification influence how long teachers remain in the profession and where they teach? *Education Finance and Policy*, *4*, 229-262. doi:10.1162/edfp.2009.4.3.229

Grigg, J., Kelly, K. A., Gamoran, A., & Borman, G. D. (2013). Effects of two scientific inquiry professional development interventions on teaching practice. *Educational Evaluation and Policy Analysis*, *35*, 38-56. doi:10.3102/0162373712461851

Guarino, C. M., Reckase, M., & Wooldridge, J. M. (2011). *Can value-added measures of teacher education performance be trusted?* (Working Paper No. 18). Retrieved from Education Policy Center, Michigan State University website: http://education.msu.edu/epc/publications/documents/Guarino-Reckase-Wooldridge-2011-Value-Added-Measures-v2.pdf

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *Journal of Human Resources*, *39*, 326-354. doi:10.2307/3559017

Hanushek, E. A., & Raymond, M. E. (2005). Does school accountability lead to improved student performance? *Journal of Policy Analysis and Management*, *24*, 297-327. doi:10.1002/pam.20091

Hill, H. C., & Ball, D. L. (2004). Learning mathematics for teaching: Results from California's mathematics professional development institutes. *Journal for Research in Mathematics Education*, *35*, 330-351. doi:10.2307/30034819

Hill, H. C., & Ball, D. L. (2009). The curious—

and crucial—case of mathematical knowledge for teaching. *Phi Delta Kappan*, *91*(2), 68-71. doi:10.1177/003172170909100215

Hill, H. C., Rowan, B., & Ball, D. L. (2005). Effects of teachers' mathematical knowledge for teaching on student achievement. *American Educational Research Journal*, *42*, 371-406. doi:10.3102/000283120420 02371Ingersoll, R., & Rossi, R. (1995). *Which types of schools have the highest teacher turnover? Issue brief* (NCES-95-778). Retrieved from http://files.eric.ed.gov/fulltext/ED388663.pdf

Jacob, B. A. (2005). Accountability, incentives and behavior: The impact of high stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*, 761-796. doi:10.1016/j.jpubeco.2004.08.004v

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (NBER Working Paper No. 14607). Retrieved from National Bureau of Economic Research website: http://www.nber.org/papers/w14607

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy*, *6*, 18-42. doi:10.1162/EDFP_a_00027

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis*, *24*, 37-62. doi:10.3102/01623737024001037

Loeb, S., Darling-Hammond, L., & Luczak, J. (2005). How teaching conditions predict teacher turnover in California schools. *Peabody Journal of Education*, *80*(3), 44-70. doi:10.1207/s15327930pje8003_4

Loucks-Horsley, S., Stiles, K. E., Mundry, S., Love, N., & Hewson, P. W. (2009). *Designing professional development for teachers of science and mathematics* (3rd ed.). Thousand Oaks, CA: Sage.

Miller, L. (2012). Situating the rural teacher labor market in the broader context: A descriptive analysis of the market dynamics in New York state. *Journal of Research in Rural Education*, *27*(13), 1-31. Retrieved from http://jrre.vmhost.psu.edu/wp-content/uploads/2014/02/27-13.pdf

Penuel, W. R., Fishman, B. J., Yamaguchi, R., & Gallagher, L. P. (2007). What makes professional development effective? Strategies that foster curriculum intervention. *American Educational Research Journal*, *44*, 921-958. doi:10.3102/0002831207308221

Penuel, W. R., Gallagher, L. P., & Moorthy, S. (2011). Preparing teachers to design sequences of instruction in earth systems science: A comparison of three professional development programs. *American Educational Research Journal*, *48*, 996-1025.

doi:10.3102/0002831211410864

Reininger, M. (2012). Hometown disadvantage? It depends on where you're from: Teachers' location preferences and the implications for staffing schools. *Educational Evaluation and Policy Analysis*, *34*, 127-145. doi:10.3102/0162373711420864

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica*, *73*, 417-458. doi:10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *94*(2), 247-252. doi:10.1257/0002828041302244

Rockoff, J. E., & Turner, L. J. (2010). Short-run impacts of accountability on school quality. *American Economic Journal: Economic Policy*, *2*(4), 119-147. doi:10.1257/pol.2.4.119

Rothstein, J. (2012). *Teacher quality policy when supply matters* (IRLE Working Paper No. 129-12). Retrieved from Institute for Research on Labor and Employment website: http://www.irle.berkeley.edu/workingpapers/129-12.pdf

Sherwood, T. (2000). Where has all the "rural" gone? Rural education research and federal reform. *Journal of Research in Rural Education*, *16*(3), 159-167. Retrieved from http://jrre.vmhost.psu.edu/wp-content/uploads/2014/02/16-3_1.pdf

Streams, M., Butler, J. S., Cowen, J., Fowles, J., & Toma, E. F. (2011). School finance reform: Do equalized expenditures imply equalized teacher salaries? *Education Finance and Policy*, *6*, 508-536. doi:10.1162/EDFP_a_00046

Wayne, A. J., Yoon, K. S., Zhu, P., Cronen, S., & Garet, M. S. (2008). Experimenting with teacher professional development: Motive and methods. *Educational Researcher*, *37*, 469-479. doi:10.3102/0013189X08327154

Winters, M. A., & Cowen, J. M. (2013a). Would a value-added system of retention improve the distribution of teacher quality? A simulation of alternative policies. *Journal of Policy Analysis and Management*, *32*, 634-654. doi:10.1002/pam.21705

Winters, M. A., & Cowen, J. M. (2013b). Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies. *Educational Researcher*, 42, 330-337.

doi:10.3102/0013189X13496145

**Appendix A**

*Balance Statistics of Unmatched and Matched Samples*

| Independent Variables | Before Match | After Match |
|---|:---:|:---:|
| **Teacher Characteristics** | | |
| Estimated Previous Effectiveness | **31.35** | 3.14 |
| Max Degree Held | 11.43 | 6.53 |
| Experience | 1.88 | 3.21 |
| Experience Squared | 6.39 | 1.80 |
| | | |
| **School Characteristics** | | |
| Average Experience | **44.39** | 13.41 |
| Enrollment (000's) | 6.49 | 1.84 |
| Percent Master's | **26.32** | 8.34 |
| Percent of FRP Students | **24.02** | 4.73 |
| Expenditure per Student (000's) | 9.16 | 2.32 |
| Student-Teacher Ratio | **62.59** | **41.14** |
| Student-Computer Ratio | **23.86** | 12.01 |
| Math Index | **32.74** | 12.15 |
| n | 5830 | 460 |

Standardized differences are presented as absolute values. Values in boldface represent large standardized differences as suggested by Rosenbaum and Rubin (1985). See also Barrett, Butler and Toma (2012) for these estimates and additional computational details.