April 2016
High-Quality Curricula: A Cost-Effective Way to Increase Student Learning
Alanna Bjorklund-Young, Research Fellow

Curricula, or the instructional materials used to teach including teacher's guides and textbooks (Boser et al., 2015), inevitably influence a teacher's lesson content and instructional approach (Reys et al., 2003). Despite such obvious importance, however, the academic curriculum is often overlooked as a factor in student outcomes. Indeed, what is striking is the paucity of research on the subject.

There have been a few scholarly forays into this domain. For instance, Grover ("Russ") Whitehurst found that using higher quality curricula increases student learning more than other, more well-known, interventions such as expanding preschool programs, giving merit pay to successful teachers, decreasing class sizes, and increasing the number of charter schools in a district (Whitehurst, 2009; Chingos & Whitehurst, 2012). Morgan Polikoff argued recently that using better curricula is a relatively inexpensive, yet impactful, intervention, since school districts regularly change their curricula, and the cost difference between different curricular programs is small (Polikoff, 2014).

What else does the research record tell us about curricula and student outcomes? Are some curricula really better than others? What important questions about curricula remained unanswered? Two recent studies advance our understanding of the curriculum effect.

Recent Research

In the first study, "Large-Scale Evaluations of Curricular Effectiveness: The Case of Elementary Mathematics in Indiana," economists Rachana Bhatt and Cory Koedel use school-level data to measure the differences in student learning outcomes *between* three different elementary mathematics curricula (2012). In the second study, "Is Curriculum Quality Uniform? Evidence from Florida", authors Bhatt, Koedel, and Douglas Lehmann use school-level data to compare whether mathematics subtopics are taught equally well *within* a particular curriculum compared to subtopics in other curricula (2013).

Both studies used a similar methodology of matching schools. The logic behind school-level matching stems from the fact that in an ideal setting, the research team would run a randomized control trial in which a group of schools would be randomly selected to participate. Some of the

schools within the group would then be randomly selected to use a new curriculum (the treatment group), and the other schools would continue to use the existing curricular materials (the control group). Researchers would then look at the average differences in the outcomes between the two groups to see if the new curriculum were more effective. Randomized control trials are ideal because *randomization* ensures that unseen factors do not drive the actual changes in student outcomes. This is what allows researchers to say that the average difference in outcome across the two groups is *caused* by the curricular change. Otherwise, unseen factors could actually be the real cause of the difference in outcomes. There are numerous unseen factors that might potentially cause a school both to choose a certain curriculum and also to increase student learning outcomes: school quality, strong administrative leadership, or better teachers, for example.

Since neither of the data sets is from randomized control trials, the research teams matched schools that chose a specific curriculum with similar schools that chose a different curriculum, in some ways mimicking the randomized control trial setting. How do researchers match schools accurately? They estimate a propensity score, designed to indicate the probability of choosing a specific curriculum, for each school. The propensity score takes into account school level information (e.g. enrollment size, student demographics, free lunch eligibility, student language, student test scores), district level information (e.g. enrollment, per-pupil revenue), and socio-economic information about the neighborhood surrounding the school (e.g. median household income, adult education levels). The team can then calculate the difference in student outcomes between schools who choose curriculum A and a weighted average of similar schools who choose curriculum B; the difference between schools who choose curriculum B and a weighted average of similar schools who choose curriculum A; and take the average difference between these two quantities to estimate the average treatment effect.

One of the most important aspects of this matching strategy is that it relies on the assumption that student test scores are independent of the selection of curricula once school and district characteristics in the data are taken into account. This assumption would be valid if, for example, curriculum is chosen because of student characteristics (e.g. the number of students, if many students are English Language Learners), as the data include these measures. This assumption would not be valid if curriculum is chosen because district leadership is more knowledgeable as more knowledgeable leadership also increases student learning in other ways (such as hiring more effective teachers). The authors provide several arguments for why this assumption might be true. For example, they describe how curricular decisions are made, which is a complex process and takes into account the opinions of many different parties. The authors further conduct falsification tests, which follow the logic of "if the curriculum is not really producing better test scores but instead there is another factor (like strong district leadership) that drives both curricular decisions and student test scores, then we should find similar kinds of results in the years before or after the curriculum is no longer in use. We should also see increased student learning in other subjects." The falsification tests do not show student learning gains in other subjects or in the years before or after the curriculum was used. Thus, the original assumption seems to hold. However, it should

be noted that the main assumption cannot be directly tested, and this assumption is critical for the validity of the results. (Bhatt & Koedel, 2012; Bhatt, Koedel, & Lehmann, 2013)

Results

In the first study, the authors compare three elementary mathematics curricula: Saxon Math, Silver Burdett Ginn (SBG) Mathematics, and Scott Foresman-Addison Wesley (SFAW). Saxon Math and SBG are considered to follow a more "traditional" pedagogy in which learning is primarily teacher-led and students receive explicit, step-by-step instruction on how to solve problems and then use worksheets and other drills to practice. SFAW follows a more "constructivist" pedagogy in which learning is more student-led, and students are encouraged to develop their own methods for solving problems, which are often based on real-world examples. These differing pedagogies reflect a longstanding debate in education over the best ways to teach students (Ravitch, 2000).

One interesting finding is that neither pedagogical approach was shown to be superior: no statistical difference was found between the constructivist curriculum (SFAW) and one of the traditional curriculum (SBG), and both of these curricula produced better student learning outcomes than the other traditional curricula (Saxon). More specifically, the authors found no statistical difference between SBG and SFAW. The largest differences in student outcomes were found between the two pedagogically traditional programs, Saxon and SBG. SBG produced effect sizes of around 0.13 standard deviations of the Indiana Statewide Testing for Educational Progress (ISTEP) exam when compared to Saxon. Effect sizes of 0.10 translate into three additional months of learning on nationally normed tests (Hill et al., 2008). The authors also found evidence that SFAW, the more constructivist program, also produced statistically significant higher outcomes than Saxon; these effect sizes were around 0.06 standard deviations when compared to Saxon. (Bhatt & Koedel, 2012).

Another important finding of this project is that Saxon remained as popular with the districts after the study, despite being the least effective program. Approximately 45% of the schools in their study used Saxon during the time period of interest (1998-2003), and 48% of the schools chose Saxon during the next procurement opportunity. Although we might suspect that cost was at play, this cannot be: only $2.26 per student separated the difference in cost between the three programs. This suggests that schools, and perhaps even districts and states, do not routinely evaluate the efficacy of their curricular decisions (Bhatt & Koedel, 2012). Other researchers have concluded this based on national audits (Chingos et al., 2015).

The second study investigated whether all elementary mathematics content material is taught equally well within a single curriculum. To do this, the researchers compare student subdomain-test scores from students who had been taught using Harcourt Math, the most commonly used curriculum in Florida at the time of the study, to other students' subdomain-test scores who were taught using other commonly used curricula in Florida (SFAW, SRA/McGraw Hill, Cambuim,

Scott Foresman Investigations, MacMillan/McGraw Hill, and Houghton Mifflin). In contrast to the first study, this research compares Harcourt to all of the other curricula collectively (as opposed to comparing two curricula at a time). Also in contrast to the first study, which looks at student outcomes using one overall math test score, the authors investigated whether students who are taught using Harcourt score equally well across the five subdomains on the Florida Comprehensive Assessment Test (FCAT): number sense, measurement, algebraic thinking, data analysis, and geometry. The authors found that in data analysis and geometry, Harcourt produced statistically significant gains over the other curricula; the estimated effect size for data analysis was between 0.092 and 0.115 and the estimated effect size for geometry was 0.108 and 0.126 for third-grade students who were taught using Harcourt in first, second, and third grades. In contrast, the authors found that Harcourt did not produce statistically higher results than the other curricula in number sense, measurement, or algebraic thinking. Therefore, the authors concluded that Harcourt does not teach all subdomains equally well, although it is a superior curriculum for teaching data analysis and geometry. (Bhatt, Koedel, & Lehmann, 2013)

Policy Implications

Together, these two papers provide evidence that curriculum does matter: some curricula produce better learning outcomes than others. Furthermore, switching to a more effective curriculum seems to be a cost-effective way to improve student outcomes. Because the districts in the studies bought new curricula every few years, their budgets already included this cost. The cost differences between the programs were small, and expense did not equate to high quality or higher test scores. From a policy perspective, therefore, choosing stronger instructional programs makes sense – provided districts and states are incentivized to evaluate on this basis.

The findings in these papers also emphasize the need for more research in this area. Bhatt, Koedel & Lehmann found that all concepts within a specific curriculum are not all taught equally well. This highlights an additional challenge: should school districts adopt a curriculum that produces the best overall scores, or use different curricula to teach different subdomains?

Most importantly, neither study answers why or under which circumstances certain curricula are better than others. For example, while Bhatt and Koedel found that SBG and SFAW were more effective than Saxon, an earlier randomized control trial found that Saxon was more effective than SFAW (Agodini et al., 2009). What accounts for the disparate findings? Is it because different types of students respond to curricula differently? Bhatt and Koedel examined schools across Indiana, while the Agodini report specifically studied disadvantaged schools. Or do the findings conflict because the time teachers spend actually teaching math differs? Agodini's report, for example, found that teachers who use Saxon reported spending one hour more per week on mathematics than the teachers using other curricula, while Bhatt and Koedel reported no time use

information. Even here, does the curriculum itself generate different amounts of time, or are other factors at play?

It would also be important to study the specific implementation plans associated with each curriculum. This might be important if, for example, one curriculum were easier to implement than another, or received more generous support within school districts in general; either of these factors could influence associated test score gains. It would also be good to know how faithfully schools and districts followed the various curricula, and whether some were conducive to partnering with additional materials. Thus, the capacity of a school or district to implement effectively might be an important consideration in which curriculum would work best.

In sum, both of these studies provide evidence that different curricula produce important differences in student learning. At the same time, they highlight the near-universal lack of attention to curricular efficacy, which is reflected in the scarcity of research on the subject.

Works Cited:

Agodini, R., Harris, B., Atkins-Burnett, S., Heaviside, S., Novak, T., & Murphy, R. (2009). Achievement Effects of Four Early Elementary School Math Curricula: Findings from First Graders in 39 Schools. NCEE 2009-4052.*National Center for Education Evaluation and Regional Assistance*.

Bhatt, R., & Koedel, C. (2012). Large-Scale Evaluations of Curricular Effectiveness The Case of Elementary Mathematics in Indiana. *Educational Evaluation and Policy Analysis*, *34*(4), 391-412.

Bhatt, R., Koedel, C., & Lehmann, D. (2013). Is curriculum quality uniform? Evidence from Florida. *Economics of Education Review*, *34*, 107-121.

Boser, U., Chingos, M., and Straus, C. (2015). "The Hidden Value of Curriculum Reform" (Washington, D.C.: Center for American Progress), retrieved at: https://cdn.americanprogress.org/wp-content/uploads/2015/10/06111518/CurriculumMatters-report.pdf.

Chingos, M. and Whitehurst, R. (2012). Choosing Blindly: Instructional Materials, Teacher Effectiveness, and the Common Core. Brown Center on Education Policy at Brookings.

Hill, C.J., Bloom, H.S., Black, A.R., and Lipsey, M.W. (2008). Empirical Benchmarks for Interpreting Effect Sizes in Research. Child Development Perspectives, 2(3):172-177.

Polikoff, M. (2014) At last, accountability for textbook publishers? Thomas Fordham Institute. Retrieved from: http://edexcellence.net/articles/at-last-accountability-for-textbook-publishers

Ravitch, D. (2000). Left back. *A century of failed school reforms. Simon Schuster, New York.*

Reys, R., Reys, B., Lapan, R., Holliday, G., & Wasman, D. (2003). Assessing the impact of" standards"-based middle grades mathematics curriculum materials on student achievement. *Journal for Research in Mathematics Education*, 74-95.

Whitehurst, G.J. (2009). Don't Forget Curriculum. Brookings Institute: Brown Center Letters on Education. Retrieved from: http://www.brookings.edu/research/papers/2009/10/14-curriculum-whitehurst