Teacher Training, Mentoring or Performance Support Systems?

Roberto Araya(✉)

Centro de Investigación Avanzada en Educación, Universidad de Chile,
Periodista Mario Carrasco 75, Santiago, Chile
roberto.araya.schulz@gmail.com

Abstract. A major challenge in education is how to improve teaching. This means improving teaching so that all students effectively achieve the levels of performance stipulated in the curriculum and that they do so within the specified timeframes. This goal is particularly difficult to achieve in schools with students of low socioeconomic status. However, measuring the quality of instruction is not a straightforward task. This is partly due to a lack of rigorous and regular data on student performance gathered by independent third parties. On the other hand, there are several alternatives for improving teaching: teacher training, teacher mentoring programs and support systems to boost teacher performance. Our study looks at eight years of data on national standardized test scores for every school in a low SES district. We found that the effect size of a Performance Support System is larger than the benchmark effect sizes for teacher training and teacher mentoring programs.

Keywords: Performance Support Systems · Teacher training
Teacher mentoring programs · Effect sizes

1 Introduction

A major challenge in education is how to improve teaching. This challenge has two critical components. Firstly, it is not easy to measure the quality of teaching. Take for example quality assessments by principals and peers. In 2009 [1] found that, in most U.S. districts, less than 1% of teachers were rated as unsatisfactory. In other words, on paper, practically every teacher is satisfactory. However, 81% of administrators and 57% of teachers could identify a teacher in their school who they considered to be ineffective. The authors named this paradox as the Widget Effect. They describe this failure as the tendency of school districts to assume that classroom effectiveness is the same from teacher to teacher. This failure may be caused by inadequate evaluation systems. A reform in evaluation procedures could therefore help solve this problem. However, in a 2017 study, [2] analyzed teacher performance ratings across 24 U.S. states in which major reforms had been made to the teacher evaluation system. They also found that, in the vast majority of these states, the number of teachers rated as unsatisfactory was still less than 1%.

One significant alternative is to measure teacher knowledge and teaching practices. However, the focus should actually be on measuring student learning. In this sense,

S. Nazir et al. (Eds.): AHFE 2018, AISC 785, pp. 306–315, 2019.
https://doi.org/10.1007/978-3-319-93882-0_30

rigorously measuring student gains is a difficult and very expensive process. For example, according to [3] the 1966 Coleman Report, often described as "the largest and most important educational study ever conducted", cost approximately USD $1.5 million. To put this into perspective, the equivalent cost in 2016 would be USD $11 million. Furthermore, there are several factors at play when it comes to calculating the effect. This includes factors such as the type of assessment, the timing, critical contextual information, and the independence of the evaluation team from the program coordinators. For example, according to [4] there is a significant difference in the average effect sizes that are found for different kinds of measurements using achievement tests. In this sense, there are three types of measurements: (i) standardized tests on a broad subject matter, (ii) standardized tests that focus on a more specific topic, and (iii) specialized tests developed specifically for an intervention (typically developed by the researchers). Larger effect sizes have been found with specialized researcher-developed tests (median 0.34 and standard deviation 0.55 for elementary school), which are presumably more closely aligned with the intervention that is being evaluated. The effect size of interventions measured with standardized tests focusing on a specific topic tend to be smaller (median 0.17 and standard deviation 0.42 for elementary school). Finally, the effect size of interventions measured using more general standardized tests are much smaller (median 0.07, standard deviation 0.27 for elementary school). In this paper, we use results from a national standardized test in mathematics (SIMCE). This is a broad test that is designed to cover the contents of the national curriculum for mathematics.

Secondly, there are several strategies for improving teaching. The most common strategy is teacher training. However, measuring the impact of teacher training on the quality of education is not a straightforward task. Several studies show that teacher training can lead to significant changes in terms of teaching practice. Despite this, teacher training has not been shown to have any impact on student performance. For example, a recent study of a year-long teacher training program for math teachers revealed significant improvements in the teachers' content knowledge, as well as changes in their teaching practices [5]. However, the study also found that there was a negative effect on student performance on state-level assessments. In a 2010 study by the same authors [6], no improvement was found in student performance, despite there being a change in teaching practices. In this case, the authors analyzed an intervention at 77 moderately high and highly vulnerable schools from 12 districts. Each teacher received 68 h of training over a number of sessions throughout the year. The training focused on teaching fractions and involved three different teaching strategies: having students comment on the results or procedures of others, having students use representations, and having them justify their results. Furthermore, each teacher was accompanied in the classroom for a total of 10 days following the initial 5-day training session. These follow-up sessions lasted for two days and were spread out across the year. The results of the study revealed that the teachers did not improve their knowledge of fractions, did not use representations more frequently and did not request more justification from their students. In fact, the teachers only slightly improved in terms of having students comment of their peers' results. However, this change did not improve the students' performance in fractions. This is despite the fact that the teachers in the control group only received 12 h of training in mathematics for the year, and not specifically in fractions. Another large meta-study

of randomized controlled trials of teacher training programs [7], revealed that structured professional development programs (highly prescriptive programs with follow-up and support) have an average effect of 0.05 standard deviations on student gains. A recent summary of research by the U.S. Institute of Education [8] analyzed 910 studies on the effectiveness of different approached to professional development in math teaching. This comprehensive literature review concluded that "until more causal evidence becomes available, schools and districts must supplement the limited evidence of effectiveness with their best judgment. Schools and districts should be encouraged to rigorously evaluate professional development approaches themselves and, when possible, to report the findings publicly to build up the knowledge base on the topic".

Are there any other strategies that might have a more significant effect on improving the quality of teaching? One such strategy is mentoring programs, also called the "Third Way" [9]. This involves experienced full-time teachers, who are carefully-selected and trained to be mentors, giving support to newly-qualified teachers during their first one or two years of teaching. For example, (Schmidt et al. 2017) showed that after one mentor worked with 15 newly-qualified teachers for two years, the program obtained an average effect of 0.15 standard deviations in terms of student gains. This is a very large effect size when compared to the effects of teacher training. However, this effect was obtained with newly-qualified teachers, who are mostly likely to improve the most. On the other hand, other studies report no effect for mentoring programs, or even negative effects. This was the case with the Urban Teacher Residencies (UTRs) program in Boston [9]. This program covers a large number of newly-qualified teachers that are hired in the Boston area. Since 2008–2009, UTR accounts for about one third of all newly-qualified teachers in the district.

So, how about strategies for improving teaching among experienced teachers? One possible strategy is the use of Performance Support Systems. This kind of system supports teachers on the job and in real time. In this paper, we analyze the impact of ConectaIdeas, a cloud-based system that helps the teacher to teach more active learning classes. The system allows the teacher to pose closed and open-ended questions, teach using games, connect with other classes in real time, hold online, synchronous tournaments with several other schools, and share their experiences with other teachers. We report on 6 years of data from the system, which was implemented in every fourth-grade mathematics class at all 11 public schools in a low SES district of Santiago, Chile.

2 Methods

We study performance in mathematics at 11 schools in Lo Prado, a low socio-economic status (SES) district of Santiago, Chile. These 11 schools are all the public elementary schools that are run by the local district. One of the schools is classified as low SES by the Ministry of Education. This segment includes the lowest SES schools in the country and accounts for approximately 7% of schools in Chile. All of the schools in this segment are considered at-risk. The rest of the 11 schools are classified as medium-low SES. This segment accounts for another 20% of schools in the country. We analyze the fourth-grade students' performance on the National Standardized Math Test (SIMCE math test)

between 2009 and 2016. This information covers eight years of standardized measurements at each school.

Furthermore, teachers on Treatment classes have been using ConectaIdeas, a Performance Support System since 2011 [10, 11]. This is a web-based platform where the teacher selects from a list of exercises or can ask the students open-ended questions. The teacher can review the answers online and ask students to do peer review. The teachers normally access the system through a tablet or smartphone in order to monitor the class. The system also helps the teacher to analyze answers to multiple-choice and open-ended questions. Furthermore, the system also preselects students that can act as monitors and help the teacher provide support to the other students. Using the system, the teacher assigns monitors to help their peers. They then receive feedback from the monitor, as well as from the student who received the support. ConectaIdeas also provides a math game, where the teacher selects the area of the curriculum to be used as the focus of the game. Students can play the game alone, or compete against a classmate. There is also the option to host inter-class or inter-school tournaments [12].

Some of the schools in this study had two fourth-grade classes, while others only had one. Over the years, some of the schools also went from having two fourth-grade classes to having just one. This means that the data comes from a total of 122 classes. Of these classes, 80 of them were under treatment, 16 in 2011, 16 in 2012, 11 in 2013, 11 in 2014, 13 in 2015, and 13 in 2016. The remaining 42 classes had gaps at different stages during this period. Of these 42 classes, 32 of them did not work with the system during 2009 and 2010. Furthermore, five of these classes also failed to use the system in 2013, while five of them did not use it in 2014. These 10 classes belong to three schools that decided not to continue with the program during those years. However, all of the schools decided to return to the program in 2015. Previous analyses [13, 14] had considered only one or three years of experience with the Performance Support System. In this study, we review the impact after six years of using the system.

The government only publishes school-level data and does not release data on individual students' performance. We are therefore not able to perform a multilevel analysis. Furthermore, we cannot use the difference in difference methodology [15], since we only have one measurement per class. This measurement is done at the end of the school year. There is no pretest for each class. However, the number of classes is enough to estimate effects by comparing treated classes with untreated classes. In addition to this, the classes take the test in different years, with the average score on the national standardized test changing from year to year. We are therefore able to correct the estimate based on these yearly changes. In any case, the differences in the average are minimal, while there is practically no difference in the standard deviation of students' scores.

This study is not a Randomized Controlled Trial (RCT) as all of the schools in the district underwent treatment in 2011. Furthermore, the district was also not randomly selected. However, three schools left the program in 2013 and 2014. This is very interesting for the purposes of our assessment as it helps us to measure the impact of the treatment. However, the three schools were also not selected at random; their principals decided to leave the program. Nevertheless, they decided to return to the program in 2015 following two years of poor results. This provides another fantastic opportunity to measure the effect.

3 Results

The average scores on the fourth-grade SIMCE math test are shown for each school between 2009 and 2016 (Fig. 1). The government only publishes the average score for each school. Since some schools have two fourth-grade classes and others only have one, we calculate the average performance by weighting according to the number of classes in each school. The schools' overall average is shown as a bold line. We can see a significant jump in the average score between 2010 and 2011. This improved score is then maintained from 2011 onwards, despite a slight dip in 2013. In any case, the average in 2013 is still higher than in 2009 and 2010.

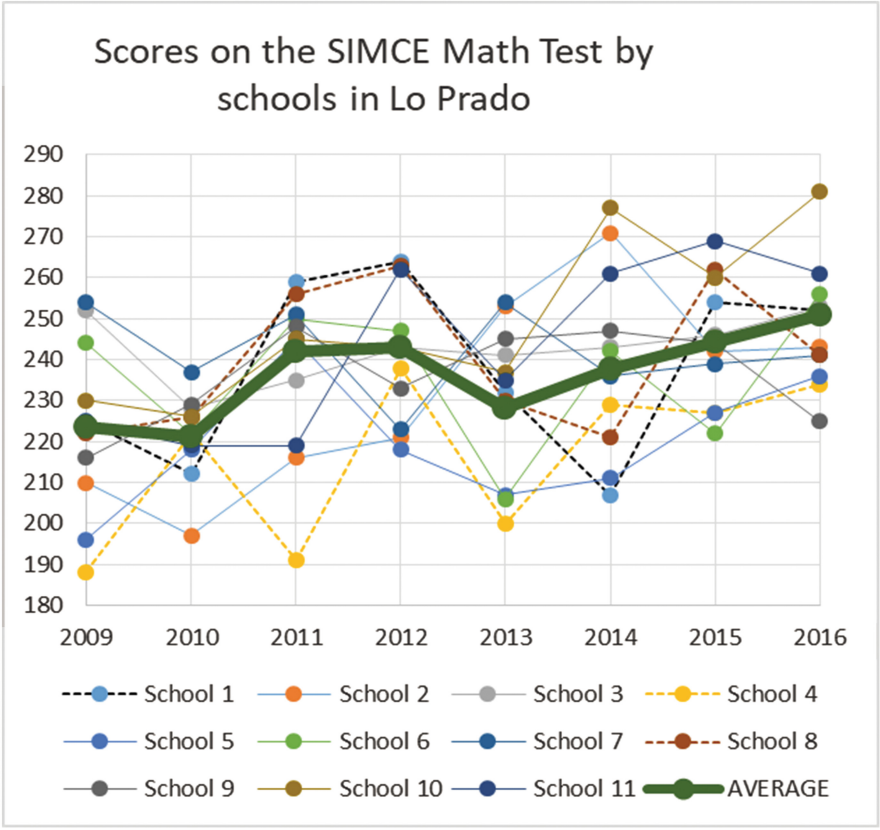

Fig. 1. Scores on the fourth-grade SIMCE math test between 2009 and 2016 for the 11 schools in Lo Prado. The overall average for all of the schools, weighted according to the number of classes in each school, is shown as a bold line. As the scores for each class are not publicly-available, we plot the overall scores for each school.

All of the schools began using the ConectaIdeas Performance Support System during the second semester of 2010. However, due to issues with their internet connection,

several schools were not able to connect and use the system. The System therefore started to be used properly in 2011. Given this, we consider that treatment started in 2011. The weighted average score on the SIMCE math test for the treatment schools is shown as a continuous bold line in Fig. 2. The average of the schools with no treatment is displayed as a dotted bold line (Fig. 2).

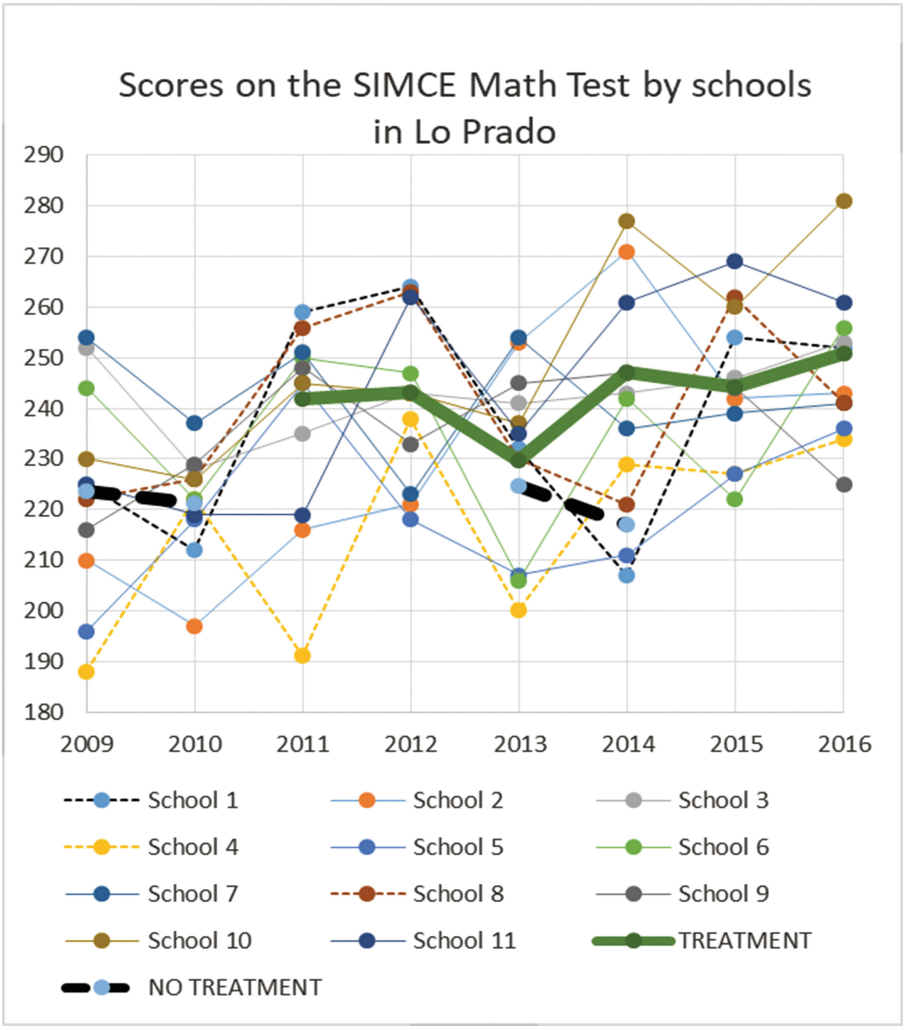

Fig. 2. Scores on the fourth-grade SIMCE math test between 2009 and 2016 for the 11 schools in Lo Prado. The weighted average score for the treatment schools is shown as a continuous bold line. The weighted average score for the schools without treatment is shown as a dotted bold lined.

Figure 2 reveals a significant improvement in the schools' performance in 2011 (i.e. the weighted average of all the schools). In this sense, there is a 20.6-point jump, which

represents 41.2 standard deviations of the students' scores on the SIMCE math test. This improvement in performance is probably the effect of the treatment. The average increase in the national average for that year was only 5.9 points, while for low-medium SES schools it was 8.44 points. In 2013 and 2014, three of the schools (School 1, School 4 and School 8, all depicted using dotted lines) were not included in the treatment program as their principals decided to leave the program. The weighted average of these three schools was less than the weighted average of the rest of the schools during 2013 and 2014. However, there was also a drop in the treatment schools' scores in 2013. The gap between the treatment and non-treatment schools was 5 points in 2013, which corresponds to 0.10 standard deviations. The overall drop was probably due to the test being more difficult, since the whole country dropped 5 points that year. This was also the case for low-medium SES schools. However, the drop-in score by students at the non-treatment schools may have been reduced by the fact that those students used the system the previous year as third graders. This is because the treatment was for both third and fourth graders. However, the scores for non-treatment students continued to drop in 2014. A probable cause of this is that these students did not use the system in 2013 and were therefore without the treatment for two consecutive years. Meanwhile, the treatment students recovered and improved their scores. Subsequently, the gap in 2014 increased to 30.1 points. This gap corresponds to 60.2 standard deviations.

Another possible explanation for the drop in the scores at these three schools in 2013 and 2014 could be the schools' management and leadership. However, as it involves three schools, the probability of mismanagement for two years at all three is quite low. Nevertheless, it may well be the case that the ineffectiveness of the principals in question led them to leave the program. However, the data on sixth-grade SIMCE Math scores for 2013 and 2014 also reveals that these three schools had reasonable good results, unlike the fourth-grade SIMCE Math scores The results on sixth grade of these schools were even better than their results in 2015 and 2016. Thus, the bad performance of the fourth graders on those schools in 2013 and 2014 was probably not due to management. The sixth-grade SIMCE tests only began in 2013 and therefore no data is available for previous years.

The average SIMCE math score for the 80 treatment classes was 242.8, whereas for the 42 non-treatment classes it was 222.2. Furthermore, the standard deviation among students in the country is 50, and in Lo Prado is 50 too. Over the years, there have been some minor variations in the national average score on the SIMCE test, as well as the average score for low SES schools. However, the national average between 2009 and 2016 has not changed by any more than 4 points, while the average for low-medium SES schools has not changed by any more than 10 points. When correcting for these yearly variations, the difference between the treatment and no treatment classes is 14.8 points. This difference corresponds to an effect size of 0.30 standard deviations.

It is interesting to observe that the average score on the SIMCE math tests for the 32 classes that did not use the system in 2009 and 2010 was 222.5. This is practically the same as the average for the 10 classes that did not use the system in 2013 and 2014 (220.9). In that sense, the classes that did not use the system in those years did not improve on their historic performance on the SIMCE math test.

4 Conclusions

In educational practice, it is not always possible or affordable to design experimental studies, such as randomized controlled trials. However, sometimes, new treatments are introduced to a large population. Then, it makes statistical sense to capitalize on the historical data that is generated whenever possible. According to [3] "Once a set of teachers or students are chosen for an intervention, the state databases could be used to match them with a group of students and teachers who have similar prior achievement and demographic characteristics and do not receive the intervention. By monitoring the subsequent achievement of the two groups, states and districts could gauge program impacts more quickly and at lower cost. The most promising interventions could later be confirmed with randomized field trials" p. 8.

In this paper, we have analyzed the data generated by a practical implementation of a program that introduces a Performance Support System. All schools in a low SES district received a new treatment in 2011. Furthermore, two additional events are very useful for estimating the effect of the treatment. Firstly, three schools (i.e. 10 classes) left the program for two years between 2013 and 2014. Secondly, these schools subsequently returned to the program in 2015 and 2016.

By comparing the performance of the treatment schools with the non-treatment schools over the years we are able to estimate an effect size. Under these special circumstances, we can estimate an effect size of 14.8 points on the National Standardized Math Test (SIMCE math test). This effect corresponds to 0.30 standard deviations.

How big is that effect? Rigorous studies [16] using randomized controlled trials and measuring gains in student performance reveal that the upper quartile of teachers in terms of added value (i.e. student gains on standardized test) is 0.33 standard deviations above the added value of the lower quartile. In other words, if we were to implement a radical (and politically impossible) strategy of firing all teachers in the lower quartile and replacing them with teachers of same quality as those in the upper quartile, we could expect to achieve an effect size of 0.08 standard deviations. Therefore, an effect size of 0.08 standard deviations seems to be the upper limit when it comes to teacher training. The effect size that is estimated for the ConectaIdeas Performance Support System is therefore much higher than the upper limit for teacher training. It is also higher than the reported effect size of mentoring programs [9, 17, 18].

Another critical issue is the need to understand this effect size in practical terms. Is the estimated effect large enough to be substantially important or relevant to policymakers? [19]. One strategy is to compare the effect with the typical increase in student learning over a whole year, as measured by standardized tests. This is another type of benchmark. Here, the effect of the intervention is compared to the natural growth in academic achievement that takes place over the course of a year for an average student. For example, according to [4, 19], the annual achievement gain on a National Standardized test in fourth-grade mathematics is 0.56. We do not have statistics on natural growth in Chile. These would be very difficult and expensive to obtain as the SIMCE math test is only sat by certain grade levels. Unlike in other countries, such as the U.S., in Chile there is not a standardized test for every grade level. The SIMCE test is a national test that is traditionally only sat by fourth graders, eighth graders and tenth graders, as

well as now by sixth graders. However, if we assume similar natural growth as in the U.S. then the intervention with the ConectaIdeas Performance Support System, which had an estimated effect size of 0.30, represents 54% of the annual natural achievement gain for fourth graders. This is an enormous gain. It is equivalent to the gain made in math through half a year of schooling.

The results obtained so far are very promising. However, this is not an RCT study and involves only one district. As [20] recommends "…more experimentation and evaluation is needed. The only way to know what works and what does not work is by innovating, piloting, evaluating, and learning".

Acknowledgments. Funding from PIA-CONICYT Basal Funds for Centers of Excellence Project FB0003 is gratefully acknowledged, as is the Fondef D15I10017 grant from CONICYT.

References

1. Weisberg, D., Sexton, S., Mulhern, J., Keeling, D.: The Widget Effect: Our National Failure to Acknowledge and Act on Differences in Teacher Effectiveness. New Teacher Project, Washington, DC (2009)
2. Kraft, M., Gilmour, A.: Revisiting the Widget Effect: Teacher Evaluation Reforms and the Distribution of Teacher Effectiveness. Educ. Res. **46**(5), 234–244 (2017)
3. Kane, T.: Connecting to practice. How we can we research to work. Educ. Next **16**(2) (2016)
4. Lipsey, M.W., Puzio, K., Yun, C., Hebert, M.A., Steinka-Fry, K., Cole, M.W., Roberts, M., Anthony, K.S., Busick, M.D.: Translating the statistical representation of the effects of education interventions into more readily interpretable forms. (NCSER 2013-3000). National Center for Special Education Research, Institute of Education Sciences (IES), U.S. Department of Education, Washington, DC (2012)
5. Garet, M.S., Heppen, J.B., Walters, K., Parkinson, J., Smith, T.M., Song, M., Garrett, R., Yang, R., Borman, G.D.: Focusing on mathematical knowledge: the impact of content-intensive teacher professional development (NCEE 2016-4010). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education, Washington, DC (2016)
6. Garet, M., Wayne, A., Stancavage, F., Taylor, J., Walters, K., Song, M., Brown, S., Hurlburt, S., Zhu, P., Sepanik, S., Doolitle, F., Warner, E.: Middle school mathematics professional development impact study findings after the first year of implementation. US Department of Education (2010)
7. Fryer, R.: The production of human capital in developed countries: evidence from 196 randomized field experiments. In: Handbook of Field Experiments, vol. 2, pp. 95–322, North-Holland, Amsterdam (2017)
8. Gersten, R., Taylor, M.J., Keys, T.D., Rolfhus, E., Newman-Gonchar, R.: Summary of research on the effectiveness of math professional development approaches. (REL 2014-010). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Southeast, Washington, DC (2014). http://ies.ed.gov/ncee/edlabs
9. Papay, J., West, M., Fullerton, J., Kane, T.: Does an urban teacher residency increase student achievement? Early evidence from Boston. Educ. Eval. Policy Anal. **34**(4), 413–434 (2012)
10. Reynolds, A., Araya, R.: Building Multimedia Performance Support Systems. McGraw Hill, New York (1995)

11. Araya, R.: Integrating classes from different schools using intelligent teacher support systems. In: Karwowski, W., Ahram, T. (eds.) Intelligent Human Systems Integration. IHSI 2018. Advances in Intelligent Systems and Computing, vol. 722, pp. 294–300. Springer, Cham (2018)
12. Araya, R., Aguirre, C., Bahamondez, M., Calfucura, P., Jaure, P.: Social Facilitation Due to Online Inter-classrooms Tournaments. LNCS, vol. 9891, pp. 16–29 (2016)
13. Araya, R., Van der Molen, J.: Impact of a blended ICT adoption model on Chilean vulnerable schools correlates with amount of on online practice. In: Proceedings of the Workshops at the 16th International Conference on Artificial Intelligence in Education AIED 2013, Memphis, 9–13 July 2013
14. Araya, R., Gormaz, R., Bahamondez, M., Aguirre, C., Calfucura, P., Jaure. P., Laborda, C.: ICT supported learning rises math achievement in low socio economic status schools. LNCS, vol. 9307, pp. 383–388 (2015)
15. Angrist, J., Pischke, J.: Mastering Metrics: The Path from Cause to Effect. Princeton University Press, Princeton (2015)
16. Kane, T., Rockoff, J., Staiger, D.: What does certification tell us about teacher effectiveness? Evidence from New York City. Econ. Educ. Rev. **27**, 615–631 (2008)
17. Schmidt, R., Young, V., Cassidy, L., Wang, H., Laguarda, K.: Impact of the New Teacher Center's New Teacher Induction Model on Teachers and Students. SRI International, Menlo Park (2017)
18. Young, V.M., Schmidt, R., Wang, H., Cassidy, L., Laguarda, K.: A comprehensive model of teacher induction: implementation and impact on teachers and students. Evaluation of the New Teacher Center's i3 Validation grant, final report. Prepared for the New Teacher Center. SRI International, Menlo Park (2017)
19. Bloom, H.S., Hill, C.J., Black, A.B., Lipsey, M.W.: Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. J. Res. Educ. Effectiveness **1**(4), 289–328 (2008)
20. Busso, M., Cristia, J., Hincapié, D., Messina, J., Ripani, L.: Learning Better. Public Policy for Skills Development. Inter-American Development Bank (2017)