Teaching Mathematics and Physical Education: Does Effectiveness Vary by Subject Matter?

Charalambos Y. Charalambous, Ermis Kyriakides, Leonidas Kyriakides, & Niki Tsangaridou
Department of Education, University of Cyprus

Rationale and Research Question

Responding to calls to consider the subject-specificity of teacher/teaching effectiveness, researchers have recently started comparing teaching quality and student learning in different subject areas (e.g., Goldhaber, Cowan, & Walch, 2013; Graeber, Newton, & Chambliss, 2012; Loeb et al., 2012; Praetorius et al., 2016). A close examination of this recent line of inquiry leads to three important observations. First, criterion-consistency findings of teaching effectiveness across different subject matters seem to be mixed. Second, all studies have attended to what can be considered "core" subject matters and cognitive outcomes. Third, two different approaches have been pursued to examine effectiveness, one attending to *student learning* through value-added models (VAM) and another considering *teaching quality*. Drawing on both these approaches, we considered teaching effectiveness in two remarkably different subject matters with respect to their targeted learning outcomes and the context in which they take place: Mathematics, which typically attends to cognitive outcomes and is taught indoors, and Physical Education (PE), which usually aims at promoting psychomotor outcomes and is taught outdoors. Seeking to extend and complement the aforementioned literature, in this study, we asked:

- *Are teachers equally effective in two subject matters that exhibit substantial differences in targeted learning outcomes and the context in which they take place?*

Methods

Setting and participants. Conducted in Cyprus, this study was based on a larger project aimed at examining teaching effectiveness in Mathematics and PE. For the purpose of the study, we focused on 24 generalist elementary school teachers teaching both subject matters in any of Grades 3 to 6 and their 548 students; 18 of these teachers were teaching both subject matters to exactly the same students. The student sample was representative of the population in terms of gender (x^2=0.01, p=0.95). The teacher sample included more males (75%) than females; this is typical of the population of generalist teachers teaching PE in Cyprus, where PE is taught for only two 40-min periods weekly and is considered a peripheral subject-matter.

Data sources. Collected during the academic year 2014-2015, the study data comprised student cognitive (for Mathematics) and psychomotor (for PE) scores obtained through initial and end-of-year tests. Additionally, each teacher was observed six times during the academic year, three in Mathematics and three in PE. Each lesson was observed by an independent trained rater who was using both a low- and a high-inference rubric of the *Dynamic Model of Educational Effectiveness* (*DMEE*, Creemers & Kyriakides, 2008)—one of the most up-to-date models of educational effectiveness focusing on generic teaching skills (cf. Scheerens, 2013).

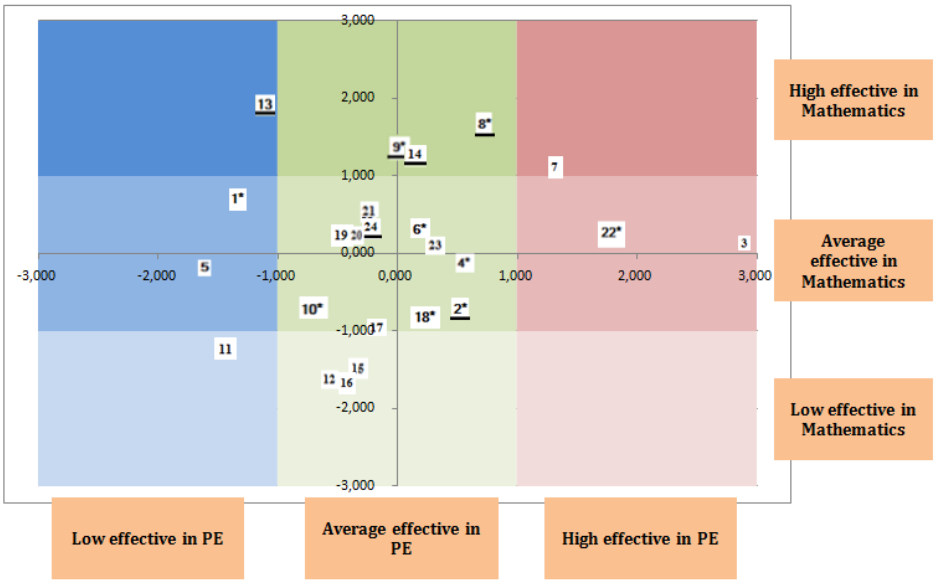
Data analyses. Item-Response-Theory (IRT), and especially the Extended Model of Rasch (Andrich, 1988), was applied to the student data to develop psychometric scales for both subject matters (pre and post). Based on the case-ability estimates of these scales, an initial and an end-of year score in Mathematics and PE was then obtained for each student; next, these scores were entered in VAM (see Appendix). The standardized residuals at the teacher level resulting from these models were utilized as indicators of teaching effectiveness in each subject matter (Goldstein, 2003); the association between these residuals was explored using Spearman's *rho*. Using these standardized residuals, we then classified teachers into categories, three for Mathematics and three for PE (low/average/high effective).[1] This resulted into a nine-category grid which allowed clustering teachers according to their effectiveness in both subject matters.

The lesson observation data were also subjected to Rasch analyses that resulted in two scales—one for each subject matter—with good psychometric properties. The scales were then equated using the mean/sigma transformation method (see Kolen & Brennan, 2014, p. 183). Using the equated case estimates for each lesson, we then explored whether the findings could be generalized at the teacher level, which was indeed the case (Mathematics: $F_{(49, 144)}$=1.55, p<.05; PE: $F_{(48,141)}$=3,93, p<.05). Hence, the mean estimate for each teacher was calculated, and Spearman's *rho* was run to explore the association in teaching quality for the two subject matters. Following a similar approach to that pursued above (mean estimate ±1SD), teachers were then clustered into one of nine categories resulting by classifying each teacher into low/average/high quality for each subject matter.

[1] Low: standardized residual ≤ -1SD; average: -1SD ≤ standardized residual ≤1SD; high: standardized residual ≥ 1SD. We did not use ±2SD since that resulted in clustering nearly all teachers in the "average" category.

Summary of Findings

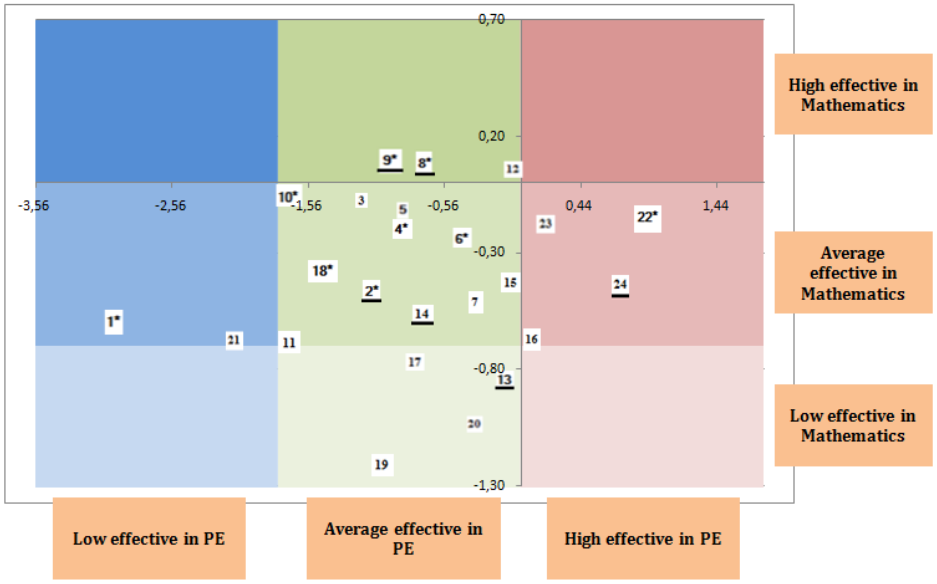
Consistency with regards to student learning. A low non-significant correlation (rho=0.23, p=0.28) among the residuals of teachers' VAM scores in both subject matters was obtained. In addition, as can be seen in Figure 1, 13 of the teachers (54%) were clustered into the same level of effectiveness for both subject matters. Eleven teachers, however, exhibited different levels of effectiveness, with one of them even clustered as (marginally) high effective in Mathematics and low effective in PE.



Note: Underlined teachers: teaching both subject matters to different students.
Teachers with asterisk: consistently clustered in the same category when using both approaches (VAM and teaching quality)

Figure 1. The classification of 24 teachers based on student learning (VAM) in Mathematics and PE.

Consistency with regards to teaching quality. A negligible correlation (rho=0.14, p=0.52) was obtained between teachers' instructional quality in the two lessons.[2] Moreover, Figure 2 shows that only 11 (46%) of the teachers were classified in the same level for both subject matters. The remaining 13 teachers occupied four of the six off-diagonal categories, but unlike for the VAM classification, no teacher was clustered as high effective in one lesson and low effective in the other. Interestingly, only nine teachers (38%, presented with an asterisk) were clustered at the same levels for both subject matters when considering both approaches outlined above.



Note: Underlined teachers: teaching both subject matters to different students.
Teachers with asterisk: consistently clustered in the same category when using both approaches (VAM and teaching quality)

Figure 2. The classification of 24 teachers based on teaching quality (*DMEE* scores).

[2] Even when considering only the 18 teachers who were teaching both subject matters to exactly the same students (teachers *not* underlined in Figures 1 and 2), results did not change noticeably (VAM residuals: rho=0.33, p=0.19; teaching quality mean estimates: rho=0.12, p=0.63).

Conclusions

This study has certain limitations, including the fact the raters observing Mathematics and PE lessons were not identical.[3] However, even if we assume that the raters of one of these subject matters were consistently stricter than those of the other, the correlations obtained for the teaching-quality approach would not necessarily be higher, since they are based on the *relative* standing of teachers. Without underestimating that the non-significant correlations obtained in the study could partly be due to rater differences—which we could not control for since the rater and the lesson effects were confounded—the present study illustrates considerable variation in teaching quality across the two subject matters. This variation held regardless of the approach pursued (VAM residuals or teaching quality estimates). Although replication studies are needed to corroborate these findings, this study suggests that when it comes to teaching remarkably different subject matters, effectiveness might differ not only across but also within teachers, thus lending credence to the idea of differential effectiveness (Campbell et al., 2004). Such findings have important implications for teacher evaluation systems, especially in elementary grades, where teachers are often evaluated mainly in "core" subject matters regardless of being expected to teach several different subject matters.

Selected references

Andrich, D. (1988). A general form of Rasch's extended logistic model for partial credit scoring. *Applied Measurement in Education, 1*(4), 363-378, doi: 10.1207/s15324818ame0104_7.

Campbell, J., Kyriakides, L., Muijs, D., & Robinson, W. (2004). *Assessing teacher effectiveness: Developing a differentiated model*. New York, NY: Routledge Falmer.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools.* London: Routledge.

Goldhaber, D., Cowan, J., & Walch, J. (2013). Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review, 36*, 216-228, doi: 10.1016/j.econedurev.2013.06.010.

Goldstein, H. (2003). *Multilevel statistical models*. London: Edward Arnold.

Graeber, A. O., Newton, K. J., & Chambliss, M. J. (2012). Crossing the borders again: Challenges in comparing the quality instruction in mathematics and reading. *Teachers College Record, 114*(4), 1-30.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking* (3rd ed.). New York, NY: Springer.

Loeb, S., Kalogrides, D., & Beteille, T. (2012). Effective schools: Teacher hiring, assignment, development, and retention. *Education Finance and Policy, 7*(3), 269–304, doi: 10.3386/w17177.

Praetorius, A. K., Vieluf, S., Saß, S., Bernholt, A., & Klieme, E. (2016). The same in German as in English? Investigating the subject-specificity of teaching quality. *Zeitschrift für Erziehungswissenschaft, 19*(1), 191-209, doi: 10.1007/s11618-015-0660-4.

Scheerens, J. (2013). The use of theory in school effectiveness research revisited. *School Effectiveness and School Improvement, 24*(1), 1-38, doi: 10.1080/09243453.2012.691100.

Corresponding Author:
Charalambos Y. Charalambous, Department of Education, University of Cyprus (cycharal@ucy.ac.cy)

[3] We had only two raters who coded both Mathematics and PE lessons of the same teachers. Even in their case, the correlations in the lesson estimates they coded were not significant (Rater $1_{(6\ lessons)}$: Spearman's rho=-0.50, p=0.31, Rater $2_{(5\ lessons)}$: Spearman's rho=-0.67, p=0.22).

Appendix

Equations used in the multi-level analyses

$$Y_{ij} = \pi_{0j} + \pi_1 X_{1ij} + \sum_{s=2}^{S} \pi_s X_{sij} + e_{ijk} \qquad \text{(Eq. 1)}$$

where:

Y_{ij} is the end-of-year outcome (cognitive or psychomotor) of student *i* taught by teacher *j*;

X_{1ij} is the variable corresponding to students' initial cognitive or psychomotor performance, [grand-mean centered]) (entered in Model 1);

X_{sij} are the student background characteristics (gender [dummy variable], and family SES) (entered in Model 2);

π_{0j} is the adjusted mean performance for students of teacher *j* after controlling for student initial performance and background characteristics;

π_1 is the fixed effect of student beginning-of-year performance;

π_s are the fixed effects of student background characteristics;

e_{ijk} is the random "student effect," that is the deviation of student *i* of teacher j from the teacher-group mean.

$$\pi_{0j} = \beta_{00} + \sum_{l=1}^{L} \beta_{0l} W_{lj} + u_{0j} \qquad \text{(Eq. 2)}$$

where:

β_{00} is the grand mean;

W_{lj} are classroom composition variables (aggregated pre-test performance at the classroom level, percentage of girls in classroom, aggregated family SES at the classroom level; entered in Model 3);

β_{0l} are the classroom-composition effects;

u_{0j} is the random "teacher effect," that is the deviation of teacher *j's* mean from the grand mean.