Доц. др Хараламбос Хараламбус
Ермис Кириакидес
Одељење за образовање
Универзитет на Кипру
Никозија, Кипар

РАЗЛИЧИТЕ МЕТОДЕ МЕРЕЊА КВАЛИТЕТА НАСТАВЕ: ДА ЛИ СУ ЈЕДНАКО ЕФИКАСНЕ?

Увод и теоријске перспективе.

Емпиријске студије спроведене током претходне две деценије конзистентно потврђују да је улога наставника критична у процесу учења (Hattie, 2009; Strong, 2011). Наиме, пронађено је да ефекти наставе могу да објасне већи проценат варијансе у ученичким постигнућима у поређењу са другим школским или сиситемским ефектима. Имајући у виду ове налазе, није случајно што се у последње време све већи нагласак ставља на боље концептуализовање, операционализацију и мерење квалитета наставе.

Фокусирајући се на питање мерења, ова студија пореди и сучељава три различите методе којима се тежи да се испита квалитет наставе: посматрање часова (e.g., Douglas, 2009), наставничке процене (самопроцене) (Kaufman, Stein, & Junker, in press) и ученичке процене (De Jong & Westerhof, 2001; Fauth et al., 2004). Свака од ових метода има своје предности и ограничења. Често сматрано „златним стандардом" мерења квалитета наставе, посматрање часова избегава бројне пристрасности повезане са самоизвештеним подацима и као такво пружа поузданије податке (Strong, 2011). Додатно, употребом ове методе бележе се снажнији ефекти него помоћу наставничких самоизвештаја и ученичких упитника (Seidel & Shavelson, 2007). Ипак, посматрање часова је скупо за извођење, а процене прикупљене овом методом такође су подложне утицајима различитих фактора, укључујући регрутовање и инструктажу посматрача, као и број и дужину посматрања (Casabianca et al., 2013; Hill, Charalambous, & Kraft, 2012). Наставничке и ученичке процене, с друге стране, релативно је јефтино добити. Такође, пронађено је и да ученичке процене агрегиране на нивоу одељења имају вишу предиктивну валидност него посматрање часова (De Jong & Westerhof, 2001).

Када су у питању недостаци процена, наставници могу намерно (Blank, 2002) или ненамерно (Cohen, 1990) известити о свом раду на начин који

не кореспондира са праксом коју изводе, а слично и ученичке процене могу бити погођене утицајем различитих фактора, попут популарности наставника (Kunter& Baumert, 2006).

Циљеви истраживања / Питања

Са циљем да допринесемо актуелном дијалогу о ефикасним и валидним техникама мерења квалитета наставе, у овој студији постављамо следећа питања:

Који од поменутих метода има већу предиктивну валидност када је у питању предвиђање исхода учења?

Да ли постоје разлике у предиктивној валидности ових метода с обзиром на различите типове исхода учења (когнитивни и афективни)?

Методологија истраживања

Контекст истраживања, учесници и извори података. Спроведена на Кипру, ова студија је изведена у склопу већег пројекта усмереног на испитивање доприноса различитих наставних пракси учењу код ученика. За сврхе ове студије, фокусирали смо се на допринос специфичних пракси, односно на понашања наставника током подучавања специфичних области; у математици – предмету испитиваном у овој студији – таква понашања односе се, на пример, на повезивање различитих репрезентација (Mitchell, Charalambous, & Hill, 2014). Учешће у студији узело је укупно 948 ученика 3. и 6. разреда основне школе и њихових 50 наставника разредне наставе.

Прикупљени током школске 2014/2015. године, подаци у овој студији могу се поделити у четири категорије: ученичка постигнућа, посматрања часова, ученичке процене и наставничке самопроцене. Ученици су попунили тест који мери постигнуће из математике на почетку и на крају школске 2014/2015 године. Тест је валидиран у претходним студијама (Kyriakides & Creemers, 2008). Ученици су такође попунили и упитник који мери њихове ставове и уверења о математици и учењу математике. Овај упитник такође је задат на почетку, као и на крају школске године. Сваки од наставника из узорка посматран је три пута током школске 2014/2015. Године, а посматрања су спровела два независна процењивача помоћу инструмента *Квалитет подучавања математике* (видети Learning Mathematics for Teaching, 2011). На крају школске године, ученици и наставници су такође попунили и друге упитнике који испитују употребу различитих наставних пракси у одељењима у којима је спроведена ова студија.

Анализа података. Помоћу анализе која се базира на теорији ставског одговора (ИРТ), за потребе мерења ученичких постигнућа развијена

је скала са добрим психометријским карактеристикама. Експлоративна факторска анализа спроведена на подацима о ученичким ставовима и уверењима указала је на постојање два конзистентна фактора прихватљиве поузданости: позитивни ставови према математици и позитивна уверења о самоефикасности. Конформаторна факторска анализа спроведена на подацима добијеним посматрањем часова и ученичким проценама наставе дала је два конзистентна фактора (богатство математике и когнитивна активација), и два додатна фактора – један подржан ученичким упитником (рад са ученицима и математика) и други подржан подацима из посматрања часова (фокусирање на математичке процедуре). Када су у питању наставничке самопроцене, будући да је коришћен мали узорак наставника, експлоративне анализе нису могле бити спроведене. Да би се добила структура слична структури утврђеној помоћу других техника мерења, подаци добијени наставничким упитником структурисани су у складу са четири претходно поменута фактора. Како бисмо одговорили на постављена истраживачка питања, коришћена је хијерархијска анализа, са ученичким подацима угнежђеним на наставнички ниво. У једначинама 1 и 2, које су приказане у Прилогу, могу се видети варијабле које су разматране на оба нивоа.

Одабрани налази

Када су у питању когнитивни исходи, где се као зависна варијабла користи ученичко постигнуће на крају школске године, 28% варијансе смештено је на наставничком нивоу (у нултом моделу); само 3% варијансе остало је необјашњено када се као контрола уведе ученичко постигнуће на предтесту (на почетку школске године). Стога је коришћење *учења* као зависне варијабле довело до тога да је на наставничком нивоу остало необјашњено 9.69% варијансе. Након контролисања ученичких карактеристика и карактеристика које се тичу састава одељења, посматрања часова објаснила су 17.65% необјашњене варијансе, док ни ученичке ни наставничке процене нису објасниле део ове варијансе. Слика је веома различита када се разматрају афективни исходи (ставови и уверења). За позитивне ставове, од 8.76% необјашњене варијансе на наставничком нивоу у нултом моделу, само су ученичке процене могле да објасне део варијансе (37.63%). За позитивна уверења о самоефикасности, када се учење користи као зависна варијабла, од 4.70% необјашњене варијансе на наставничком нивоу у нултом моделу, ученичке процене опет објашњавају највећи део варијансе (25.71%). Следе наставничке самопроцене (22.86%), док посматрања часова не објашњавају ниједан део необјашњене варијансе.

Закључак и импликације

Имајући у виду да посматрања часова објашњавају већи проценат необјашњене варијансе за когнитивне исходе на наставничком нивоу, док ученичке процене конзистентно објашњавају већи проценат варијансе за оба афективна исхода, одговор на питање која је метода мерења најбоља не може бити једнозначан: ефикасност различитих метода мерења изгледа да зависи од тога који тип исхода учења се разматра. Уколико се овај налаз потврди и у другим студијама, могао би имати важне импликације за мерење квалитета наставе, будући да сугерише да се већа пажња мора обратити на исход који је предмет испитивања: одређене методе мерења могу бити предиктивније за одређене исходе учења. Дакле, чини се да је за разумевање везе између квалитета наставе и учења, у свој њеној комплексности, потребан систем, а не једна одређена евалуациона метода.

Литература

Blank, R. K. (2002). Using surveys of enacted curriculum to advance evaluation of instruction in relation to standards. Peabody Journal of Education, 77 (4), 86–121. doi:10.1207/S15327930PJE7704_5

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. Educational and Psychological Measurement, 73 (5), 757–783. doi:10.1177/0013164413486987

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. Educational Evaluation and Policy Analysis, 12 (3), 311–329.

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

De Jong, R., & Westerhof, K. J. (2001). The quality of student ratings of teacher behaviour. Learning Environments Research, 4, 51–85.

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. Educational Researcher, 38 (7), 518–521.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. Learning and Instruction, 29, 1–9. doi:10.1016/j.learninstruc.2013.07.001Hattie, J. (2009). Visible learning: a synthesis of over 800 meta-analyses relating to achievement. New York: Routledge.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. Educational

Researcher, 41 (2), 56–64. / DOI: 10.3102/0013189X12437203

Kaufman, J. H., Stein, M. K., & Junker, B. (in press). Factors associated with alignment between teacher survey reports and classroom observation ratings of mathematics instruction. Elementary School Journal, 116 (3).

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. Learning Environment Research, 9, 231–251. doi: 10.1007/s10984-006-9015-7

Kyriakides, L., & Creemers, B. P. M. (2008). Using a multidimensional approach to measure the impact of classroom level factors upon student achievement: a study testing the validity of the dynamic model. School Effectiveness and School Improvement, 19 (2), 183–205

Learning Mathematics for Teaching. (2011). Measuring the mathematical quality of mathematics instruction.Journal of Mathematics Teacher Education, 14, 25–47.

Mitchell, R., Charalambous, C. Y., & Hill, C.H. (2014). Examining the task and knowledge demands needed to teach with representations. Journal of Mathematics Teacher Education, 17, 37–60. / DOI: 10.1007/s10857-013-9253-4

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. Review of Educational Research, 77 (4), 454–499. doi:10.3102/0034654307310317

Strong, M. (2011). The highly qualified teacher: What is teacher quality and how do we measure it? New York, NY: Teachers College Press.

Прилог

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

- Y_{ij} исход на крају школске године (когнитивни или афективни) ученика и подучаваног од стране наставника j;
- X_{1ij} афективни или когнитивни исход ученика на предтесту, [центрирано] (унето у Модел 1);
- π_{0j} ученичке карактеристике (унесене у Модел 2);
- π_{0j} просечно постигнуће ученика наставника j после контролисања постигнућа ученика на предтесту и ученичких карактеристика;
- π_1 фиксни ефекат ученичког иницијалног постигнућа (постигнућа на почетку школске године);
- π_s фиксни ефекат ученичких карактеристика;
- e_{ij} рандом „ученички ефекат", односно одступање ученика и наставника j од групног просека (група дефинисана наставником који предаје одређеној групи ученика).

где је: $$\pi_{0j} = \beta_{00} + \sum_{l=1}^{L} \beta_{0l} W_{lj} + \sum_{m=1}^{M} \beta_{0m} W_{mj} + \sum_{n=1}^{N} \beta_{0n} W_{nj} + \sum_{p=1}^{P} \beta_{0p} W_{pj} + u_{0j} \qquad \text{(Eq. 2)}$$

- β_{00} велики просек (Grand mean);
- W_{lj} варијабле које описују састав одељења (просечно иницијално постигнуће и проценат девојчица у одељењу, центрирано на просек; унето у Модел 3);
- W_{mj} скорови за садржај специфичне наставне праксе наставника j, добијени посматрањем часова (центриране на просек, унете у Модел 4);
- W_{nj} скорови за садржај специфичне наставне праксе наставника j, добијени ученичким проценама (центрирано на просек; унето у Модел 5 без варијабли из Модела 4);
- W_{pj} скорови за садржај специфичне наставне праксе наставника j, добијени наставничким самопроценама (центрирано на просек; унето у Модел 6 без варијабли из Модела 4 и 5);
- β_{0l} ефекти састава одељења;
- β_{0m} ефекти специфичних наставних пракси, мерених преко посматрања часова;
- β_{0n} ефекти специфичних наставних пракси, мерених преко ученичких посматрања;
- β_{0p} ефекти специфичних наставних пракси, мерених преко наставничких самопроцена;
- u_{0j} рандом „наставнички ефекти", односно одступање наставника j од великог просека

Кључне речи: посматрање часова, математика, ученичке процене, квалитет наставе, наставничке процене.

Charalambos Y. Charalambous
Ermis Kyriakides
Department of Education,
University of Cyprus

DIFFERENT METHODS OF MEASURING TEACHING QUALITY: ARE THEY EQUALLY EFFECTIVE?

Introduction and Theoretical Perspectives

Empirical studies over the past two decades have repeatedly and consistently documented the critical role that teachers have for student learning (Hattie, 2009; Strong, 2011). Specifically, teacher effects have been found to explain a higher percentage of variance in student achievement compared to school- or system-level effects. Given these results, the increased emphasis recently placed on better conceptualizing, operationalizing, and measuring teaching quality is not coincidental.

Focusing on issues of measurement, this study compares and contrasts three different methods pursued to investigate teaching quality: classroom observations (e.g., Douglas, 2009), teacher ratings (i.e., self-reports) (e.g., Kaufman, Stein, & Junker, in press), and student ratings (e.g. De Jong & Westerhof, 2001; Fauth et al., 2004). Each of these methods has its strengths and limitations. Often considered the "gold standard" of measuring teaching quality, classroom observations can avoid many of the biases associated with self-reported data; as such, they can yield more reliable data (Strong, 2011). They can also produce stronger effects than those obtained through teacher self-reports or student surveys (Seidel & Shavelson, 2007). On the other hand, they are expensive to perform; the estimates obtained from classroom observations are also influenced by many factors, including the recruitment and training of classroom observers, and the number and the length of observations to be conducted (Casabianca et al., 2013; Hill, Charalambous, & Kraft, 2012). Teacher and student ratings, on the other hand, are relatively inexpensive to obtain; student ratings have also been found to have higher predictive validity than classroom observations when aggregated at the classroom level (De Jong & Westerhof, 2001). Yet, both these latter methoods are not without their own limitations. Teachers might deliberately (Blank, 2002) or unwittingly (Cohen, 1990) delineate their work in ways that depart from their actual practice; likewise, student ratings are affected by factors such as teacher popularity (Kunter& Baumert, 2006).

Research Aims/Questions

Aiming to contribute to the ongoing dialogue about measuring teaching quality effectively and accurately, in this study we asked:

Which of the abovementioned methods has more predictive power in determining student learning outcomes?

Are these approaches differentially effective in predicting student learning when it comes to different types of learning outcomes (cognitive and affective)?

Research Methods

Setting, participants and data sources. Conducted in Cyprus, this study was based on a larger project aimed at examining the contribution of different teaching practices to student learning. For the purposes of this study, we focused on the contribution of content-specific practices, that is teaching behaviors that are more pertinent to teaching specific content areas; in mathematics—which is the subject matter of interest in this study—such behaviors pertain, for instance, to linking and connecting different representations (Mitchell, Charalambous, & Hill, 2014). A total of 948 3rd to 6th grade elementary school students and their 50 generalist teachers participated in the study.

Collected during the academic year 2014-2015, the study data comprised four categories: student learning outcomes; classroom observations; student ratings; and teacher ratings. In particular, students completed a test measuring their performance in mathematics at the beginning and end of the academic year (AY) 2014-2015; the test administered to students was validated in prior studies (Kyriakides & Creemers, 2008). Students also completed a questionnaire measuring their attitudes and beliefs toward doing and learning mathematics; informed by the TIMSS student survey questions, this questionnaire was also administered at the beginning and end of the AY. Each of the teachers participating in the study was observed three times during the AY 2014-2015; observations were conducted by two independent raters using the *Mathematical Quality of Instruction* Instrument (see Learning Mathematics for Teaching, 2011). At the end of the AY 2014-2015, students and teachers alike were also asked to complete different surveys gauging the use of certain teaching practices in the classes under consideration.

Data analysis. An Item-Response-Theory (IRT) scale with good psychometric properties was developed to capture student achievement in mathematics. Exploratory factor analyses of student affective responses yielded two consistent factors that had acceptable reliabilities: positive attitudes toward mathematics and positive self-efficacy beliefs. Confirmatory factor analyses applied to the classroom observations and student ratings yielded two consistent factors (richness of the

mathematics and cognitive activation), and two additional factors, one supported by the student survey (working with students and mathematics) and the other supported by the classroom-observation data (focusing on mathematical procedures). For teacher ratings, because of the small teacher sample, we could not run any exploratory analyses. To have a similar structure to that obtained from the other two measurement methods, we imposed the structure of the aforementioned four factors on the teacher-survey data. To then answer the two research questions, we used multi-level analysis, with students nested within teachers. Equations 1 and 2 that appear in the Appendix outline the variables considered in each of the two levels, respectively.

Selected Findings

For the cognitive outcome, when using student final performance as the dependent variable, 28% of the variance was situated at the teacher level (for the null model); only 3% of this variance remained unexplained once controlling for student pre-test performance. Hence, we used student *learning* as the dependent variable; this led to 9.69% of the unexplained variance lying at the teacher level. Once controlling for student background and classroom compositional effects, classroom observations explained 17.65% of the unexplained variance, whereas both student and teacher ratings did not explain any portion of this variance. The picture was quite different when considering student affective outcomes. For *positive attitudes*, out of the 8.76% of the unexplained teacher-level variance in the null model, only student ratings explained a portion (37.63%). For *positive self-efficacy beliefs*, when using student learning as the dependent variable, of the 4.70% of the unexplained teacher-level variance in the null model, student ratings again explained the biggest portion (25.71%), followed by teacher ratings (22.86%); no variance was explained by classroom observations.

Conclusions and Implications

Given that classroom observations were found to explain a bigger portion of the unexplained teacher-level variance for the cognitive outcome, whereas student ratings consistently explained a bigger portion of the variance for both affective outcomes, the answer to which measurement method is best cannot be straightforward: the effectiveness of different measurement methods seems to depend on the type of learning outcome considered. Provided that this finding is replicated in other studies, it can have important implications for measuring teaching quality, since it would suggest that greater attention needs to be paid to the outcome examined: certain measurement methods might be more predictive of particular

learning outcomes than others. Hence, a system of teacher/teaching evaluation methods—as opposed to a single evaluation method—seems to be needed, if we are to better understand the link between teaching quality and student learning in all its complexity.

References

Blank, R. K. (2002). Using surveys of enacted curriculum to advance evaluation of instruction in relation to standards. *Peabody Journal of Education*, *77*(4), 86–121. doi:10.1207/S15327930PJE7704_5

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *73*(5), 757-783. doi:10.1177/0013164413486987

Cohen, D. K. (1990). A revolution in one classroom: The case of Mrs. Oublier. *Educational Evaluation and Policy Analysis, 12* (3), 311-329.

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

De Jong, R., & Westerhof, K.J. (2001). The quality of student ratings of teacher behaviour. *Learning Environments Research 4,* 51-85.

Douglas, K. (2009). Sharpening our focus in measuring classroom instruction. *Educational Researcher, 38* (7), 518-521.

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1-9. doi:10.1016/j.learninstruc.2013.07.001

Hattie, J. (2009). *Visible learning: a synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Hill, H. C., Charalambous, C. Y., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the G-study. *Educational Researcher, 41*(2), 56-64. / DOI: 10.3102/0013189X12437203

Kaufman, J. H., Stein, M. K., & Junker, B. (in press). Factors associated with alignment between teacher survey reports and classroom observation ratings of mathematics instruction. *Elementary School Journal, 116* (3).

Kunter, M., & Baumert, J. (2006). Who is the expert? Construct and criteria validity of student and teacher ratings of instruction. *Learning Environment Research, 9,* 231-251. doi: 10.1007/s10984-006-9015-7

Kyriakides, L., & Creemers, B.P.M. (2008). Using a multidimensional approach to measure the impact of classroom level factors upon student achievement: astudy testing the validity of the dynamic model. *School Effectiveness and School Improvement, 19*(2), 183-205

Learning Mathematics for Teaching. (2011). Measuring the mathematical quality of mathematics instruction.*Journal of Mathematics Teacher Education, 14,* 25-47.

Mitchell, R., Charalambous, C. Y., & Hill, C.H. (2014). Examining the task and knowledge demands needed to teach with representations. *Journal of Mathematics Teacher Education, 17, 37–60. / DOI: 10.1007/s10857-013-9253-4*

Seidel, T., & Shavelson, R. J. (2007). Teaching effectiveness research in the past decade: The role of theory and research design in disentangling meta-analysis results. Review of Educational Research, 77 (4), 454–499. doi:10.3102/0034654307310317

Strong, M. (2011). The highly qualified teacher: What is teacher quality and how do we measure it? New York, NY: Teachers College Press.

Appendix

where

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

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

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

π_{0j} are the student background characteristics (entered in Model 2);

π_{0_j} 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_{ij} is the random "student effect," that is the deviation of student *i* of teacher from the teacher-group mean.

where:

$$\pi_{0j} = \beta_{00} + \sum_{l=1}^{L} \beta_{0l} W_{lj} + \sum_{m=1}^{M} \beta_{0m} W_{mj} + \sum_{n=1}^{N} \beta_{0n} W_{nj} + \sum_{p=1}^{P} \beta_{0p} W_{pj} + u_{0j} \qquad \text{(Eq. 2)}$$

β_{00_l} is the grand mean;

W_{lj} are classroom composition variables (average initial performance and percentage of girls within a class, grand-mean centered; entered in Model 3);

W_{mj} are the content-specific teaching practice scores from classroom observations of teacher j (grand-mean centered; entered in Model 4);

W_{nj} are the content-specific teaching practice scores from student ratings for teacher j (grand-mean centered; entered in Model 5 without the variables of Models 4);

W_{pj} are the content-specific teaching practice scores from teacher ratings for teacher j (grand-mean centered; entered in Model 6 without the variables of Models 4 and 5);

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

β_{0n} are the effects of content-specific practices for the observational scores;

β_{0n} are the effects of content-specific practices for the student ratings;

β_{0p} are the effects of content-specific practices for the teacher ratings;

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

Keywords: classroom observations, mathematics, student ratings, teaching quality, teacher ratings.