"Does Professor Quality Matter?": A Critical Review of the Study of Carrell and West (2010)

Economic Theory I (ECO-M005)

CARSTEN JÜRGEN CREDE*

Universities' heavy usage of evaluation scores in staff pay and promotion decisions has recently come under attack by researchers who believe student evaluations to be an inadequate indicator of teaching quality. In this essay, a recent contribution of Carrell and West (2010) in support of this view is critically analysed by comparison with other studies. Their principal research question is whether students reward lecturers in course evaluations for focussing lectures on exam-relevant topics allowing students to gain higher marks in the exam or, alternatively, for imparting deep learning - a profound comprehension of the curriculum allowing them to excel in subsequent courses.[1] In order to answer this question, Carrell and West conduct an econometric analysis using data of student achievement as well as on student and lecturer characteristics collected at the United States Air Force Academy (USAFA). An outstanding feature of the analysis is the validity of the data set for the analysis as due to the way students are assigned to teaching personnel and courses, an unbiased analysis free of problems of self-selection and attrition that undermine the validity of most econometric analyses in this research area is possible.[2]

* MSc Industrial Economics Student

[1] The "teaching to the test" behaviour requires a lecturer to know about the contents of an exam in advance such that an adaption of the teaching putting emphasis on exam-relevant topics is possible. Such knowledge was available to the teaching personnel in the analysed data set.

[2] The USAFA randomly assigns students to different lecturers teaching the same course and curriculum (preventing self-selection of students) and standardised exams are marked jointly by all course lecturers (which ensures the validity of their outcome measure, see below). Furthermore, all students are obliged to take subsequent modules that build upon the core modules (therefore attrition is not a problem). However, note that the sample is not random, as students are chosen in a complicated selection process. Therefore, there is no certainty about the external validity of the results, as the sample population is not a random draw representing the average population (although the university considers ethnic diversity and equal treatment of women in the selection process). Using econometric terminology, it can be said that one should interpret these results as the effect of treatment on the treated rather than thinking about them as the average treatment effect. See, e.g. Schlotter et al. (2011) for evaluation problems of non-experimental data and strategies to construct counterfactual situations for valid statistical inference.

© Norwich Economic Papers

The authors use random effects and normalised Bayesian shrinkage estimators to estimate the introductory module teacher's value-added to a student influencing his follow-on performance in subsequent courses with the normalised student's average percentage points in each course in combination with the student-level records of all courses serving as the primary outcome measure.

Their results indicate that lecturers producing higher achievements of students in the contemporaneous courses (the introductory modules) tend to impart less deep knowledge than those with lower contemporaneous course results. Moreover, in course evaluations students are inclined to give better evaluations to lecturers teaching to the test improving student achievement in the contemporaneous course than to lecturers, who teach deeper knowledge allowing students to excel in subsequent courses. High rank and experience of the lecturer are negatively correlated with student evaluations and contemporaneous course results, but positively related to follow-on course performance. Opposite effects apply to lecturers with low experience or rank. In light of these results, Carrell and West put the validity of student evaluations for academic promotions and payments into question and allude to the fact that this incentive scheme favours lecturers producing inferior long-term learning outcomes, i.e. lower grades in follow-on courses.

The authors provide several explanations for the findings: first, teachers with less experience might follow the curriculum more strictly, whereas more experienced personnel such as professors provide a broader teaching of topics facilitating deep learning. Second, the "teaching to the test" could negatively affect student effort for follow-on exam preparations. Third, students exposed to lecturers imparting more deep learning and producing lower than expected marks in the contemporaneous course could increase effort in follow-on classes to compensate their perceived (but in fact non-existent) lack of knowledge.

These results are in sharp contrast with other papers analysing the validity of student evaluations, which remain the subject of heated debate among researchers. Older studies have predominantly found student evaluations to be a mostly unbiased indicator of teaching performance (Marsh 1987:369, Cashin 1995:7, d'Apollonia and Abrami 1997:1203-1204). Not only do factors such as expected grade, course difficulty and workload do not negatively affect evaluations (Centra 2003:514, Remedios and Lieberman 2008:112), perceived low workload and unchallenging course difficulty even get punished by students (Marsh and Roche 2000:226). These results contradict the findings of Carrell and West by leading to the conclusion that student evaluations are a good indicator of lecturer performance.

However, the majority of recent contributions tend to support the view that student evaluations are biased by numerous factors. This shift in view is a result of a change of focus in econometric analysis away from using the perceived learning of students as measure and relying on actual student outcome measures instead, as was done by Carrell and West (see, e.g. Galbraith and Merrill 2011:9). Then, in line with Carrell and West, Jacob et al. (2010:940) find contemporaneous course results to be a poor measure for long-term value-added learning outcomes. Moreover, their hypothesis that teaching to the test results in higher scores than imparting deep learning is supported by findings of Braga et al. (2011:30). Furthermore, one of the main findings of the older literature defending the validity of student evaluations has been subject to criticism: evidence suggests that lecturers indeed can influence their evaluation scores by inflating grade expectations (McPherson et al. 2009:48, Phipps et al. 2006:242). With other papers proposing that students might purposely submit false evaluations (Clayson and Haley 2011:107), provide their end-of-year evaluation partly on the basis of pre-course attitudes or unfavourable reputation (Francis 2011:151, McNatt 2010:225) and generally have an insufficient ability of self-assessment (Lew et al. 2010:152, Walker and Palmer 2009:166), West and Carrell's critical view on student evaluations is shared by many researchers. In accordance with them, widespread support can thus be found for the call for caution using student evaluation in institutional assessment of teaching quality for decisions regarding job tenure, promotions and bonus payments (McPherson et al. 2009:48, McNatt 2010:238, Clayson and Haley 2011:109, Galbraith et al. 2011:17).

In light of the advances of research in this area, the paper of Carrell and West can be seen as a valuable contribution providing new insights into the nature of biases affecting the validity of student evaluation. Notable strengths of their study are the reliance on actual teaching outcomes as measures of teaching quality and the integrity of data leaving few doubts about problems of self-selection and attrition biases, which other papers have only been able to address with strong assumptions about counterfactual situations casting doubt on the validity of corresponding results.[3] The results of the paper conform with recent findings of other studies, giving credibility to their teaching to the test-hypothesis. Nonetheless, their other explanations, while conceivable, remain untested and require further analyses. In addition, the authors only allude to the idea that the unfavourable link between high scores in student evaluations (the lecturer performance indicator) and teachers producing lower long-term educational outcomes might be the result of a wrong institutional design of incentive schemes negatively affecting teaching quality. This question has so far not been addressed by researchers in econometric analyses with the exception of McPherson et al. (2009:48) and remains an aspect of student evaluation with unknown consequences.[4]

[3] Despite the high internal validity of the data, it remains the task of future research to conduct analyses in random samples (i.e. universities with less strict selection processes) to assess the external validity of the results of Carrell and West.

[4] However, note that game-theoretic approaches to this problem indicate that due to wrong incentive schemes, teachers might not provide the socially optimal quality of teaching, see, e.g. Holmstrom and Milgrom (1991) or Baker (1992).

References

BAKER, G. P. (1992). Incentive Contracts and Performance Measurement. *Journal of Political Economy*, **100** (3), 598–614.

BRAGA, M., PACCAGNELLA, M. and PELLIZZARI, M. (2011). Evaluating students' evaluation of professors. *Innocenzo Gasparini Institute for Economic Research*, (IGIERWorking Papers No. 384).

CARRELL, S. E. and WEST, J. E. (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, **118** (3), 409–432.

CASHIN, W. E. (1995). Student Ratings of Teaching: The Research Revisited. *Kansas State Univ., Manhattan. Center for Faculty Evaluation and Development in Higher Education*, (IDEA Paper No. 32).

CENTRA, J. A. (2003). Will Teachers Receive Higher Student Evaluations by Giving Higher Grades and Less Course Work? *Research in Higher Education*, **44** (5), 495–518.

CLAYSON, D. E. and HALEY, D. A. (2011). Are Students Telling Us the Truth? A Critical Look at the Student Evaluation of Teaching. *Marketing Education Review*, **21** (2), 101–112.

D'APOLLONIA, S. and ABRAMI, P. C. (1997). Navigating Student Ratings of Instruction. *American Psychologist*, **52** (11), 1198–1208.

FRANCIS, C. A. (2011). Student Course Evaluations: Association with Pre-course Attitudes and Comparison of Business Courses in Social Science and Quantitative Topics. *North American Journal of Psychology*, **13** (1), 141–154.

GALBRAITH, C. S. and MERRILL, G. B. (2011). Faculty Research Productivity and Standardized Student Learning Outcomes in a University Teaching Environment: A Bayesian Analysis of Relationships. *Studies in Higher Education*, forthcoming.

—, — and KLINE, D. (2011). Are Student Evaluations of Teaching Effectiveness Valid for Measuring Student Learning Outcomes in Business Related Classes? A Neural Network and Bayesian Analyses. *Research in Higher Education*, forthcoming.

HOLMSTROM, B. and MILGROM, P. (1991). Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design. *Journal of Law, Economics, & Organization*, **7**, 24–52.

JACOB, B. A., LEFGREN, L. and SIMS, D. P. (2010). The Persistence of Teacher-Induced Learning. *Journal of Human Resources*, **45** (4), 915–943.

LEW, M. D., ALWIS, W. and SCHMIDT, H. G. (2010). Accuracy of students' self–assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, **35** (2), 135–156.

MARSH, H. W. (1987). Students' evaluations of University teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, **11** (3), 253–388.

— and ROCHE, L. A. (2000). Effects of Grading Leniency and Low Workload on Students' Evaluations of Teaching: Popular Myth, Bias Validity, or Innocent Bystanders? *Journal of Educational Psychology*, **92** (1), 202–228.

MCNATT, D. B. (2010). Negative Reputation and Biased Student Evaluations of Teaching: Longitudial Results From a Naturally Ocurring Experiment. *Academy of Management Learning & Education*, **9** (2), 224–242.

MCPHERSON, M. A., JEWELL R. TODD and KIM, M. (2009). What Determines Student Evaluation Scores? A Random Effects Analysis of Undergraduate Economics Classes. *Eastern Economic Journal*, **35** (1), 37–51.

PHIPPS, S., KIDD, R. and LATIF, D. (2006). Relationships among student evaluations, instructor effectiveness, and academic performance. *Pharmacy Education*, **6** (4), 237–243.

REMEDIOS, R. and LIEBERMAN, D. A. (2008). I liked your course because you taught me well: the influence of grades, workload, expectations and goals on students' evaluations of teaching. *British Educational Research Journal*, **34** (1), 91–115.

SCHLOTTER, M., SCHWERDT, G. and WÖSSMANN, L. (2011). Econometric methods for causal evaluation of education policies and practices: a non-technical guide. *Education Economics*, **19** (2), 109–137.

WALKER, D. J. and PALMER, E. (2009). The relationship between student understanding, satisfaction and performance in an Australian engineering programme. *Assessment & Evaluation in Higher Education*, **36** (2), 157–170.