Nils BUCHHOLTZ, Oslo, Kirsti KLETTE, Oslo, & Astrid ROE, Oslo

Students' ratings of instructional quality and achievement in mathematics

Introduction

The assessment of instructional quality through student ratings is pursued in many studies and is regarded as an economic instrument for assessing teaching quality (Kunter & Baumert, 2006; Praetorius, 2014, Kane & Staiger, 2012, Ferguson, 2010). However, the assessment of instructional quality through student ratings is still criticised for not being sufficiently valid. Praetorius (2014) for example points out that when measuring different dimensions of instructional quality, the measurement does not have to be equally valid for all sub-dimensions. Furthermore, the measurement of instructional quality through student ratings only seems appropriate for those facets that concern the perceived teaching climate and do not require methodical or didactical knowledge that students generally do not possess. Despite this critique, several studies have however argued for the value of student ratings of instructional quality when aggregated to class means in order to predict outcomes such as students' learning gains or motivational development (Kunter & Baumert, 2006). For example, Kuhfeld (2017) investigates the predictive validity of student ratings from the Tripod Survey (Ferguson, 2010; Wallace, Kelcey & Ruzek, 2016) with student survey data collected by the Measures of Effective Teaching (MET) Project. The Tripod instrument is used in the USA for the evaluation of teachers. The findings of the study showed significant relationships between the dimension classroom management assessed by the students and end-of-year student test scores in mathematics. However, the findings of Kuhfeld (2017) also suggest that the theoretical dimensionality of the Tripod instrument with its differentiation into seven different facets of instructional quality is empirically not guaranteed: “The exploratory analyses suggested that Control (classroom management) dimension was a separate factor from the other Seven Cs“ (ibid. S. 271). No significant correlations were found for the other dimensions examined.

The LISA study

The study LISA (Linking Instruction and Student Achievement) also examines, among other things, the relationship between the quality of instruction in Norwegian language and Mathematics and the achievement development of Norwegian students (Klette, Blikstad-Balas & Roe, 2017). The quality of

Siller, H.-S., Weigel, W. & Wörler, J. F. (Hrsg.). *Beiträge zum Mathematikunterricht 2020.* Münster: WTM-Verlag, 2020.
doi: 10.37626/GA9783959871402.0

instruction in the classroom was measured by video recordings of four mathematics lessons each and by questionnaires at student level. The ratings of instructional quality of N = 991 students from a total of 47 Norwegian mathematics classes were collected. Aggregated at class level these ratings can be related to student outcome data. Students' achievements and their learning gains were measured by two national tests on numeracy which allow comparison of student achievements at the end of grade 8 and grade 9.

Dimensions of the Tripod Survey instrument

The LISA study surveys instructional quality with the Tripod Survey (Ferguson, 2010) which focuses on different aspects of what teachers do in the classroom and is operationalized in seven dimensions of instructional quality as the so-called Tripod Seven Cs framework of teacher effectiveness (*Care, Control, Clarify, Challenge, Captivate, Confer and Consolidate*, cf. Wallace et al., 2016). Examples of items can be found in Table 1. Students assess the quality of teaching on a five-level Likert scale (1 = never, 2 = seldom, 3 = sometimes, 4 = often, and 5 = always). Negatively formulated items were recoded for the analyses. Confirmatory factor analyses at the individual level confirmed the theoretically assumed seven-factor dimensioning of the Norwegian empirical data. The model showed an acceptable to good quality of fit ($\chi^2/df = 3.55$; RMSEA = 0.05; CFI = 0.95). What is noticeable in Table 1 is the clearly high average approval which the students give to their teachers and which do not differ particularly across the dimensions. All dimensions except the *Control* dimension also have low Inter-Class Correlations (ICC), suggesting a high degree of variance within classes but not between classes. The classes therefore differ relatively little with regard to the assessment of instructional quality. Nevertheless, the hierarchical structure of the data was taken into account for further analyses, since the students' achievement data were due to data restrictions only available at class level.

Relationships between instructional quality and achievement development

The LISA study can use student achievement data from the Norwegian national tests ("nasjonale prøver") in numeracy for each class participating in the study. All students in eighth and ninth grade take the same national tests in reading and numeracy each year in order to track students' learning gains over a school year (Klette et al., 2017). At class level, aggregated scores for 47 classes were available from the official scaling of the test results of the Norwegian Directorate for Education and Training from 2012 and 2013. The average performance of the 47 classes (Mean_8: 50.7; Mean_9: 54.3) roughly corresponds to the national average of all classes (Mean_8: 50; Mean_9: 54).

Dimen-sion	**Items**	**Mean**	**SD**	**α**	**Example**	**ICC**
Care	5	3.88	0.75	.83	My teacher in this class makes me feel that s/he really cares about me.	.12
Control	7	3.97	0.57	.84	Student behavior in this class is under control.	.26
Clarify	6	3.92	0.69	.84	My teacher explains difficult things clearly.	.13
Challenge	7	3.90	0.58	.74	My teacher asks students to explain more about answers they give.	.13
Captivate	4	3.52	0.92	.88	My teacher makes lessons interesting.	.14
Confer	5	3.38	0.72	.76	My teacher gives us time to explain our ideas.	.12
Consoli-date	4	3.87	0.78	.81	We get helpful comments to let us know what we did wrong on assignments.	.11

Tab.: Factor structure with descriptives

Multiple regression analyses were carried out at class level to predict students' achievements and their learning gains from aggregated factor scores on class level based on the Seven Cs model. In the first analysis *Control* and *Captivate* statistically significantly predicted students' achievements at the end of grade 9 (standardized regression coefficients *Control*: .37, $p < 0.07$; *Captivate*: .51; $p < 0.05$). The two variables added statistically significantly to the prediction (total $R^2 = 0.66$; $p < .001$).

When looking at learning gains between the two tests, differences in aggregated scores were taken as dependent variable. The gains of the classes are in the range between 0.52 and 9.15 points, which corresponds to effect sizes between 0.05 and 1.15. *Captivate* statistically significantly predicted students' learning gains (standardized regression coefficients *Captivate*: .54; $p < 0.05$). Furthermore, *Clarify* became a predictor of students' learning gains on the 10% level of significance albeit with a negative standardized regression coefficient (*Clarify*: -.51; $p < 0.1$). The 7Cs together explained 74% of the variance on class level ($R^2 = 0.74$; $p < .05$).

Discussion

The influence of classroom management on student performance has already been demonstrated in numerous studies. The influence of motivational and interest-promoting actions of the teacher on student achievement gains also appears plausible. Interestingly, no further relationships could be found. As the findings of Kuhfeld (2017) suggest, however, multicollinearities of dimensions could be responsible for this. The dimension of the teacher's actions that promote motivation and interest, which is valued by the students, also has an effect on the increase in learning at class level. In this context, it is interesting to note the negative influence of the teacher's assessed explanatory activities on learning gains, which initially appears to be counterintuitive. This could be due to the fact that explanatory activities are more pronounced in performance-heterogeneous and weak performing classes than in performance-homogeneous and high performing classes and that students perceive these teacher activities differently either as promoting learning or as hindering a fast pace in learning. Further analyses appear appropriate here for clarification purposes.

References

Ferguson, R. (2010). *Student perceptions of teaching effectiveness*. Discussion brief. Cambridge, MA: National Center for Teacher Effectiveness and the Achievement Gap Initiative, Harvard University.

Kane, T. J. & Staiger, D.O (2012) *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains*. Research paper, MET Project, Bill & Melinda Gates Foundation.

Klette, K., Blikstad-Balas, M. & Roe, A. (2017). Linking Instruction and Student Achievement – Research design for a new generation of classroom studies. *Acta Didactica Norge,* 11(3), article 10, 19 pages.

Kuhfeld, M. (2017). When students grade their teachers: A validity analysis of the Tripod Student Survey. *Educational Assessment,* 22(4), 253–274.

Kunter, M. & Baumert, J. (2006). Linking TIMSS to research on learning and instruction: A re-analysis of the German TIMSS and TIMSS video data. In S. J. Howie & T. Plomp (Hrsg.), *Learning mathematics and science: Lessons learned from TIMSS* (S. 335–351). London: Routledge.

Praetorius, A.-K. (2014). *Messung von Unterrichtsqualität durch Ratings*. Münster: Waxmann.

Wallace, T. L., Kelcey, B. & Ruzek, E. (2016). What can student perception surveys tell us about teaching? Empirically testing the underlying structure of the tripod student perception survey. *American Educational Research Journal,* 53(6), 1834–1868.