Journal for Educational Research Online
Journal für Bildungsforschung Online
Volume 9 (2017), No. 1, 99–117

Eyvind Elstad, Eli Lejonberg & Knut-Andreas Christophersen

Student evaluation of high-school teaching: Which factors are associated with teachers' perception of the usefulness of being evaluated?

Abstract

This study builds on and contributes to work about the use of student evaluation of teacher performance. Although many studies have examined multiple forms of teacher evaluation, not much has been written about high school teachers' perception of the usefulness of evaluations performed anonymously by students. This article provides additional insight by exploring factors contributing to Norwegian high school teachers' perceptions of the usefulness of evaluations. Structural equation modelling indicates that perceptions of the developmental purposes of the evaluation process and of clear communication from the school leadership, as well as acknowledgement of the students' ability to evaluate, are associated with teachers' perceptions of the usefulness of the evaluation. Student ratings are often used for administrative purposes and tend to be underutilized for developmental purposes. Our findings suggest that feedback from student ratings can be useful in improving teaching practices by providing high school teachers with constructive feedback with which to improve the quality of their teaching.

Keywords

Teacher evaluation; Student ratings of teacher performance; Teaching evaluation; Teacher appraisal; Professional development

Prof. Dr. Eyvind Elstad (corresponding author) · PhD Eli Lejonberg, Department of Teacher Education and School Research, University of Oslo, P.O. Box 1099 Blindern, 0317 Oslo, Norway
e-mail: eyvind.elstad@ils.uio.no
eli.lejonberg@ils.uio.no

Associate Prof. Knut-Andreas Christophersen, Department of Political Science, University of Oslo, P.O. Box 1097 Blindern, 0317 Oslo, Norway
e-mail: k.a.christophersen@stv.uio.no

Schülerbewertungen von Unterricht an weiterführenden Schulen: Welche Faktoren hängen mit der lehrerseitigen Nützlichkeitswahrnehmung von Unterrichtsevaluationen zusammen?

Zusammenfassung

Die vorliegende Studie leistet einen Beitrag zur Forschung über den Einsatz von Schülerbefragungen zur Unterrichtsevaluation. Wenngleich zahlreiche Studien vielfältige Formen der Lehrerevaluation untersucht haben, ist über die lehrerseitige Nützlichkeitswahrnehmung von anonymen Unterrichtsbewertungen durch Schülerinnen und Schüler wenig bekannt. Der vorliegende Artikel beleuchtet diese Thematik, indem Faktoren für den wahrgenommenen Nutzen von Evaluationen bei Lehrkräften weiterführender Schulen in Norwegen untersucht werden. Strukturgleichungsmodelle deuten darauf hin, dass die Wahrnehmung eines Entwicklungsziels beim Evaluationsprozess, eine klare Kommunikation der Schulleitung sowie die Anerkennung der Beurteilungsfähigkeit von Schülerinnen und Schülern mit der von Lehrkräften wahrgenommenen Nützlichkeit der Evaluation zusammenhängen. Schülerbewertungen dienen oftmals zu administrativen Zwecken und werden hingegen für entwicklungsorientierte Ziele bislang zu wenig genutzt. Unsere Befunde legen nahe, dass Schülerbewertungen einen Nutzen für die Unterrichtsentwicklung an weiterführenden Schulen haben können, indem sie Lehrkräften konstruktive Rückmeldungen zur Verbesserung ihrer Lehrpraxis bereitstellen.

Schlagwörter

Lehrerevaluation; Schülerbewertung von Lehrerleistung; Unterrichtsevaluation; Berufliche Fortbildung

1. Introduction

Teacher evaluation[1] is at the core of current education policies in many countries (OECD, 2009). The usefulness of teacher evaluations for the improvement of teaching depends upon the extent to which teachers respond to and use them. This study focuses on student ratings of teacher performance in Norwegian high schools. Student evaluation of teaching serves several outcomes: Student ratings of teacher performance have the potential to provide both administrative quality assurance of teaching as well as information needed to promote self-evaluation and

1 Teacher evaluation, teacher assessment, teacher appraisal or student ratings of teacher performance have often been used among scholars as interchangeable terms. Teacher evaluation is, however, also used as a broader concept comprising student ratings of teacher performance as well as observation of teachers' educational performance by school administrators and measurement of added value in the students' learning outcomes.

reflection upon one's teaching practice, which promotes personal growth and learning. These dual aims may create an unresolved tension (Penny, 2003). However, the usefulness of student ratings of teacher performance is linked to teachers' attitudes towards the evaluation scheme. The purpose of this article is to explore the strength of the statistical associations between how teachers perceive the usefulness of student ratings of teacher performance (teaching evaluation) as the dependent variable and the following independent variables: perceived purposes of evaluation, communication with leaders and acknowledgement of the feedback provided by students.

The actual use of student ratings for formative purposes often falls short of its potential. We presume that the effectiveness of students' evaluations may be related to teachers' perceptions of usefulness: If teachers deny the importance of student feedback, we cannot expect that student feedback to contribute significantly to their personal or professional development. However, if teachers take the feedback from students seriously, we can expect that they may also find it useful. Therefore, usefulness is linked to teachers' attitudes towards student ratings of teacher performance, and the perception of usefulness is a possible precursor for improving educational practices and for professional development. This study attempts to answer the following research question: Which variables are associated with teachers' perception of the usefulness of the evaluations?

2. Background

Teacher evaluation is part of an international trend in which different means to evaluate teachers' educational practices have been implemented in schools in a number of countries since the start of the millennium (Isoré, 2009).[2] Systems in which students provide ratings of teacher performance are ultimately aimed at improving the educational performance of teachers and thereby furthering student learning. Anonymous surveys of students are an uncommon method to evaluate teachers' educational practice (Stronge, 2010). It is therefore essential to investigate what factors are statistically associated with teachers' perceptions of the usefulness of this form of student feedback.

Some local and national education authorities emphasize the summative component of ratings: The results of student ratings of teacher performance may include administrative monitoring of the teachers' educational performance, and when the results measured at the school level are collated in internet portals (and/or reported in the media), student satisfaction with teaching is turned into a quality indicator that can be used as an argument in support of the quality of schools, as well as to monitor quality development. Quality measures can be thus used to market particular schools to present and future students and their parents as well

2 In Norway, in 2003, a student association, launched initiatives for systematic anonymous teaching evaluation.

as providing evidence for institutional accountability for local politicians. From a perspective that emphasizes the professional development of teachers, student ratings of teacher performance can contribute to self-evaluation and reflection upon one's teaching practice, promoting personal growth and learning (Darling-Hammond, Wise, & Pease, 1983; Darling-Hammond, 2013; Day, Flores, & Viana, 2007). However, the proposal to introduce student ratings of teacher performance in schools has met with resistance (Avalos & Assael, 2006; Flores, 2010; Elstad, Lejonberg, & Christophersen, 2015), and the claim that a teacher's teaching evaluation data, value-added data and observation data could promote professional development remains controversial (Isoré, 2009; Smylie, 2014).

The *Organisation for Economic Co-operation and Development* (often abbreviated OECD) has advised Norway to implement a national framework for teacher evaluation schemes (OECD, 2011), and the central Norwegian Ministry of Education and Research has launched teacher evaluation as a key measure (The Government, 2013). Some high schools have implemented an approach involving an anonymous survey among students. The results of this survey are used in conversations between the teachers and their leaders as a starting point for each teacher's professional development work. It is interesting to investigate what factors are associated empirically with how teachers perceived the usefulness of an evaluation process in which students evaluate their teachers' educational practices using anonymous surveys. For student ratings of teacher performance to contribute positively to the teachers' reflections on their own potential for professional improvement (what they are doing successfully and less successfully), it is essential to understand the antecedents of the usefulness of student ratings of teacher performance as perceived by teachers. Since teachers' perceptions are assumed to depend on the context, we focus here on exploring factors that are statistically associated with the measured usefulness through systematically analyzing teacher perceptions in a particular context.

3. Theoretical framework

Student evaluation of teachers' performance (teaching evaluation) has been an active field of study for more than 80 years (Clayson, 2009). However, student ratings of teacher performance have been and are still a contested topic in educational practice and among educational researchers (Jong & Westerhof, 2001; Millman & Darling-Hammond, 1989; Stronge, 2010; Clayson, 2009; Aleamoni, 1987; Feldman, 2007; Kulik, 2001; Svinicki & McKeachie, 2011; Theall & Franklin, 2007; Tuytens & Devos, 2014). There is a huge body of research concerning student ratings of teaching quality, examining their validity and their popularity (for instance, Costin, Greenough, & Menges, 1971; Abrami, d'Apollonia, & Cohen, 1990; Benton & Cashin, 2014). Wright and Jenkins-Guarnieri (2012) find that student ratings of teacher performance appear to be valid, while other scholars express doubts (e.g.,

Greenwald & Gillmore, 1997). Whether ratings of teaching quality by students are reliable and valid is still an open question, although some progress has been shown in these measures; our investigation focuses on the potential of student ratings of teacher performance to inform professional development, rather than on their validity. Benton and Cashin (2014) and Spooren, Brockx and Mortelmans (2013) have summarized research on student ratings in higher education; they have found that the utility and validity of these measures should continue to be called into question. Nevertheless, student ratings are used as a measure of teaching performance in almost every institution of higher education, throughout the world (Zabaleta, 2007; Marsh, 2007). Therefore, improvement of evaluation practices is important.

Many factors are found to affect student ratings of teachers (Eagly, Ashmore, Makhijani, & Longo, 1991; Patzer, 2012). Several unintended consequences have been identified; for instance, teachers may seek popularity by raising grades (Anderson, 2002; Murray, 1997), while student ratings also depend on the subject taught (Feldman, 1978), physical attractiveness (Eagly et al., 1991), and gender (Boring, Ottoboni, & Stark 2016a). Even the ability of students to evaluate teaching practices has been called into question (Weems & Rogers, 2010). Although the validity of these ratings has been challenged, nevertheless student evaluations of individual teachers do provide student perspectives on educational practice. Meanwhile, little has been written about teachers' perceptions of the usefulness of evaluations done anonymously by high school students (Peterson, 2000; Ellett & Teddlie, 2003; Marzano & Toth, 2013). Therefore, this study focuses on how teachers perceive high school students' ratings of their teaching.

Teachers' perceptions of the usefulness of student ratings of teacher performance is the dependent variable in our theoretical model because we presume that student ratings of teacher performance may foster professional development and educational improvement through feedback and reflection in, on and about practice. The underlying assumption is that the usefulness of student ratings of teacher performance relates to the perceived purposes of the evaluation, characteristics of the school administrators and teachers' regard for the students as appraisers. The following sections identify relevant literature supporting this framework.

3.1 Purpose of the ratings

In the Norwegian context, the declared premise is that the student ratings should serve formative purposes and are not in themselves a way to monitor performance (GNIST, 2013). This means, for example, that the results from student surveys could be analyzed in a performance review, where a middle manager or the school principal discusses the mean class scores with the teacher in question. It is recommended that such performance reviews focus on the potential for improvement and on issues to work on until the next round of evaluations and performance reviews (Fullan, 2014). However, although the declared purpose is development-oriented, the evaluations can be used informally for staffing decisions or local merit-pay de-

cisions. This means that the teachers' perception of the way leadership exercises its role is related to whether the teachers regard the scheme as useful for developing their own efforts as teachers, or whether it is used as a means to control teachers. Indeed, such purposes can be combined (Isoré, 2009; Katsuno, 2010). Others have found that the perception of developmental purposes is important for student ratings of teacher performance to have beneficial outcomes (Deneire, Vanhoof, Faddar, Gijbels, & Van Petegem, 2014; Smylie, 2014). On the contrary, however, Delvaux et al. (2013) found that perceived summative purposes had a small, significant and positive effect on professional development, while formative purposes did not. Recent research has found that teachers' perceptions of the purposes of evaluation are related to the ways they understand the evaluation (Flores, 2012; Katsuno, 2010).

The evaluation context matters (Symposium, 2013). Therefore, this study explores the statistical associations between the perceived usefulness of evaluations, by systematically analyzing teacher perceptions of this issue on the one hand and, on the other, whether teachers perceive the purposes of the evaluation to be to control them or to contribute to their professional development. This gives the following hypothesis (H1): Perceived purpose of student ratings of teacher performance is positively associated with their perceived usefulness of being evaluated.

3.2 Leadership characteristics

Student ratings of teacher performance are followed up through conversations with the leadership; therefore, school leaders assume overall responsibility for this process. Their role is to communicate the administrative goals and expectations for the school in a way that helps teachers understand what is expected of them (Heck, 1992). Transformational leadership assumes that an understanding of the school's goals will help teachers develop and refine their instructional practices (Leithwood, 1994). Delvaux et al. (2013) found that transformational leadership had no significant effect on professional development, but meanwhile, Tuytens and Devos (2010) found that the principal's vision related significantly to teachers' perceptions of the need for classroom observations to factor into teacher evaluation. Based on this evidence, our second hypothesis (H2) is that clear communication of what is expected of the teachers will predict whether student ratings of teacher performance are perceived as useful.

3.3 Acknowledgement of students as evaluators

One major concern in this area involves the validity and the reliability of student opinion on teaching (Greenwald, 1997). In general, teachers tend to agree that student ratings are an acceptable means of assessing institutional integrity, and that they may be useful for administrative decision-making. The extent to which stu-

dents are able to provide appropriate teacher evaluations may depend on their maturity. Students in high schools may report a mixture of mature and immature assessments of their teachers' educational practices. Teachers who have a good relationship with their students may more easily have confidence in students' judgements of their teaching quality than teachers who have a less positive relationship with their students. Good relationships between teachers and students are regarded as a significant feature of quality schools (Hughes, Cavell, & Wilson, 2001; Bryk & Schneider, 2002). Others have found that students' perceptions of teaching differ with age (Bakx, Koopman, de Kruijf, & den Brok, 2015). In our context, we expect that the level of acknowledgement of the students' feedback will predict the level of perceived usefulness of the student ratings of teacher performance (H3).

4. Design and sample

The educational authorities in Norway have proposed a trial arrangement of teacher evaluations. This approach, which includes anonymous surveys of students and, typically, follow-up through individual conversations and group sessions for each discipline, is widely regarded as representing best practices and therefore provides an interesting model for further trials and a subsequent pilot study. Teachers' unions and representatives have endorsed the implementation of this scheme in the county under investigation, which is identical for the various schools in the county. For this study, we selected five high schools with general studies programs spread out over the entire county. A comparison by gender (56 % female teachers, 44 % male teachers) and age (47 years on average) shows that this sample from this county is well-aligned with the characteristics of the general population of high school teachers in Norway. Students in these schools perform at approximately the national average in key disciplines, and to some extent better than the national average. The average age among the teachers is somewhat higher than the average of all teachers in general studies programs, while the gender distribution is approximately equal to that of the reference population (teachers in general studies programs in this county). All five of the schools included in the purposive sample are regarded as well-established and well-managed, with a workforce that has remained relatively stable over several years. The empirical patterns that emerge from our study may thus provide some insight into statistical associations between the dependent variable and the independent variables in these types of schools. We also assume that these patterns are valid beyond our actual sample.

The study was implemented by one of the authors of this article using the occasion of joint mandatory meetings with various teachers. The main features of the study were explained, and the teachers were informed that participation was voluntary. All the respondents (N = 268) are teachers of students at high school (age 16–18) who had been through the student ratings of teacher performance process at least once. None of the teachers present at the time of the data gathering exercised

their right to decline to participate, and the response rate was therefore 89.33 % (n = 255) in four of the five schools, based on the teachers' availability.[3] The teachers completed a paper-based questionnaire, in which they ticked off pre-determined response alternatives. The teachers entered no information that could reveal their identity, and the survey is thus fully anonymized. We may therefore assume that the teachers have entered truthful and well-considered judgments in completing their responses, and that the sample is representative of teachers in the county.

4.1 Measurement instruments

A questionnaire was constructed based on measurement instruments previously adopted in the literature. In this survey, teachers responded to items on a 7-point Likert scale, where the alternative 4 represented a neutral midpoint: "How strongly do you agree with the statements below? Rank your responses from 1 (*fully disagree*) to 7 (*fully agree*)." Each concept was measured using three or four items. The analysis reported in the following is based on eight measurement instruments. The internal consistency (Cronbach's alpha) for each concept is satisfactory; Cronbach's alpha varies between .88 and .92. The concepts, items and Cronbach's alpha (α_c) for the concepts are represented in Table 1, below.

4.2 Analysis

Structural equation modelling (SEM) was used to analyze the relationships between the variables. SEM is suitable for confirmatory factor analysis and path analysis (Kline, 2005). Assessments of fit between the model and the data are based on the following indices: root mean square error of approximation (RMSEA), normed fit index (NFI), goodness-of-fit index (GFI) and comparative fit index (CFI). Values of RMSEA < .05 and NFI, GFI and CFI > .95 indicate good fit and RMSEA < .08 and NFI, GFI and CFI > .90 indicate acceptable fit (Kline, 2005).

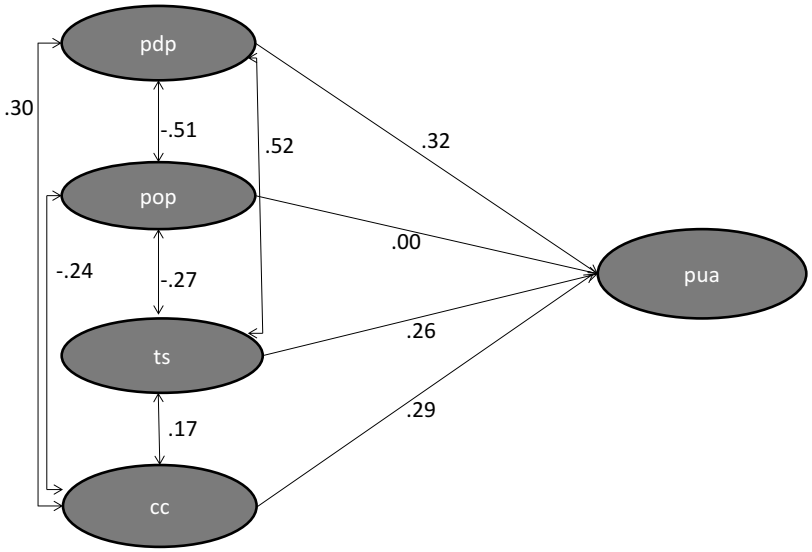
The measurement and structural models were estimated with IBM SPSS Amos 22. The values of RMSEA, NFI, GFI and CFI indicate that the structural model in Figure 1 has an acceptable degree of fit. Ellipses represent the latent variables, circles represent measurement errors and rectangles represent the observed measured variables. The structural model consists of terms with paths (arrows) between them. The path arrows indicate theoretical common causes, while the figures (standardized regression coefficients) reflect the measured strength of the connections. The strength increases together with the numerical value.

3 In one school, an extraordinary situation occurred that obligated some of the personnel to be elsewhere on the day that the survey was scheduled to take place. As a result, only 13 of the 43 teachers that made up the permanent teaching staff of this school participated. However, we have no reason to believe that this situation has given rise to a sampling bias because of the approximately random process of selection.

Table 1: Concepts, Cronbach's alpha, item wording, item means and standard deviations

Perceived usefulness of student ratings of teacher performance, α_c = .90 *From Mo, Conners, & McCormick (1998); Heneman III, & Milanowski (2003)*	*M*	*SD*
I learn a lot from student ratings of teacher performance.	3.78	1.46
Student ratings of teacher performance help provide insight into aspects of my teaching that I need to develop.	4.31	1.45
Student ratings of teacher performance help me learn more about what my strengths are in terms of teaching.	4.4	1.46
I have a clear idea of what is expected of me, thanks to the feedback I receive from the student ratings of teacher performance.	3.79	1.54
Perceived developmental purposes of student ratings of teacher performance, α_c = .92 *From Kelly, Ang, Chong, & Hu (2008)*	*M*	*SD*
The purpose of student ratings of teacher performance is to help teachers in their professional development.	4.33	1.67
The purpose of student ratings of teacher performance is to help teachers in the execution of their profession.	4.34	1.60
The purpose of student ratings of teacher performance is to provide teachers with better insight into their own teaching.	4.88	1.53
Perceived control purposes of student ratings of teacher performance, α_c = .88 *From Kelly et al. (2008)*	*M*	*SD*
The purpose of student ratings of teacher performance is to obtain an overview of which teachers are good and which are poor.	3.47	1.80
The purpose of student ratings of teacher performance is to establish competition among the teachers.	2.34	1.62
The purpose of student ratings of teacher performance is to monitor the teachers' classroom work.	3.96	1.92
Acknowledgement of the feedback from students, α_c = .88 *From Kelly et al. (2008)*	*M*	*SD*
I trust the judgement of those students who contribute to the evaluation of my teaching.	4.28	1.56
Those students who contribute to the evaluation of my teaching are competent to evaluate my teaching.	4.08	1.60
Those students who contribute to the evaluation of my teaching have sufficient insight into the teaching profession to evaluate the work of a teacher.	3.06	1.54
Perceived clear communication with the school leadership, α_c = .90 *From Hulpia, Devos, & Rosseel (2009)*	*M*	*SD*
The communication by the leadership of this school is generally clear and understandable.	4.64	1.50
The communication with the leadership helps me understand what is expected of me.	4.68	1.37
The communication with the leadership helps me understand the goals that the school seeks to achieve.	5.01	1.30

Figure 1: Estimated model



Notes. N = 206. PUA = Perceived usefulness of being evaluated; CC = Perceived clear communication with the leadership; PDP = Perceived developmental purposes of teacher evaluation; PCP = Perceived control purposes of teacher evaluation; TS = Acknowledgement students as relevant evaluator. Standardized estimates are RMSEA = 0.035, NFI = .945, CFI = .989, and GFI = .931.

5. Results

The structural equation models show which factors are statistically associated with the concept of perceived usefulness (PUA). First, perceived usefulness is clearly associated with the teachers' acknowledgement of the students' feedback ($b_{(TS \rightarrow PUA)}$ = .26). Moreover, the analysis shows that teachers' perception of clear communication with the school leadership is strongly associated with perceived usefulness of the student ratings of teacher performance ($b_{(CC \rightarrow PUA)}$ = .29). Furthermore, the model shows strong and positive associations between perceived developmental purposes of student ratings of teacher performance and acknowledgement of students' feedback ($r_{(PDP \leftrightarrow TS)}$ = .52), and also between perceived developmental purposes of student ratings of teacher performance and perception of clear communication of goals ($r_{(PDP \leftrightarrow CC)}$ = 0.30). Perceived developmental purposes of student ratings of teacher performance are positively associated with perceived usefulness of student ratings ($b_{(PDP \rightarrow PUA)}$ = .32) and negatively associated with perceived control purposes ($r_{(PDP \leftrightarrow PCP)}$ = -.51). However, those teachers who gave lower assessments of these aspects tended to give higher ratings of the perceived control purposes of student ratings of teacher performance ($r_{(PCP \leftrightarrow CC)}$: = -.24 and $b_{(PCP \rightarrow TS)}$ = -.27). In other words, at least two groups of teachers can be identified: (a) those who appreciate student ratings of teacher performance, have positive atti-

tudes towards the developmental purposes of the evaluations and the school leaderships' communication of goals, and also acknowledgement of students' feedback on the one hand; and (b) those who rate the usefulness of student ratings of teacher performance lower, consider student ratings of teacher performance as an instrument to obtain an overview of which teachers are good and which are poor, and do not experience clear communication from school managers.

6. Discussion

The purpose of this study was to explore the factors that are statistically associated with the perceived usefulness of having students evaluate teaching anonymously. First, these results provide an empirical basis for suggesting that the exercise of leadership by school administrators can have a positive effect on teachers' perceptions of the usefulness of this approach. In a Norwegian context, positive effects of student ratings on teacher performance could only be seen if teachers themselves accept a scheme in which someone is keeping tabs on them, that is, if the school leadership has access to information about how the students perceive the quality of their teaching (in the form of the teachers' average scores on some quality indicators). The beneficial effects of this procedure therefore depend on teachers' attitudes towards it. However, institutionalized feedback may bring unintended consequences, such as for instance grade inflation (Greenwald & Gillmore, 1997) and gender bias (Boring et al., 2016a). Further, student evaluations of teaching (mostly) do not measure teaching effectiveness and other behavior aimed at improving ratings rather than educational outcomes. The perceived usefulness of student ratings of teacher performance is not automatically related to improvements in teaching effectiveness (Boring et al., 2016b). How is perceived usefulness related to learning and under which conditions? How can student ratings of teacher performance improve educational practices and their professional development programs and under what conditions might they impede real quality improvements? All of these issues are avenues for further research.

Indeed, tensions related to student ratings of teacher performance are evident within Norway as well. The ability of students to evaluate teaching practices has been questioned, asking in particular whether students evaluate the teaching or the teacher. In the extension of such arguments, it follows that teachers might seek popularity – by raising grades, for instance – to get better evaluations. However, the relevant teachers' unions in Norway seem to have accepted student ratings of teacher performance as part of a system of national student ratings of teacher performance. It is relevant to ask under what conditions the unions have accepted this.

As things stand in Norway, schools that have teachers whose educational practices are perceived as very poor by the students will face an intractable problem, unless a teacher has broken laws or regulations of their profession. While the prin-

cipal has certain leadership prerogatives, a permanently appointed teacher can hardly be dismissed on the basis of nothing more than low scores in student ratings of teacher performance (Eriksen, 2012). In our study, poor scores could be interpreted in terms of cognitive dissonance (Festinger, 1957). Some teachers may think: "I don't care what the students think about my teaching, since they are not academically competent to assess its quality."[4] The potential beneficial effects of the evaluation system may therefore be weakened if teachers perceive feedback from students as irrelevant. This means that the acceptance the scheme receives will be a key factor in any beneficial effects. In a substantial sense, confidence in the students depends on whether they are seen as providing adequate assessments of the teacher's educational practice. There are studies showing that some groups of students may disparage absolutely everything associated with their schooling (Willis, 1977). In our study, however, there is a moderately strong relationship between the teachers' confidence in their students and their perception of the usefulness of evaluations. This provides grounds to assume that these medium-to-high achieving students at the upper secondary level are able to play a constructive role in the professional development of teachers. It also shows that relational confidence in the students is a key precondition for student ratings of teacher performance to be seen as valuable for the self-evaluation and reflection upon one's teaching practice. How to encourage teachers to recognize their students' ability to evaluate their educational experience is also a matter of interest. Teachers' confidence in their students' evaluations is negatively related to perceived control purposes of student feedback and positively related to its perceived developmental purposes. This indicates that clear communication about the developmental purposes of student feedback, and consistently using the results in this way, could contribute to teachers' perception that their students' evaluations are useful.

7. Limitations

There are several limitations to this study. This type of analysis has limitations from a conceptual perspective (parsimonious modelling) and in terms of its methodological (cross-sectional) approach. We acknowledge these limitations and argue that they can serve as points of departure for future research. One limitation of this study is the use of self-reported questionnaire data. The subjective component of such data is undeniable, and only a limited number of concepts have been examined. A final limitation is the limited sample of teachers. Overall, these shortcomings provide several directions for future research.

Multiple factors are related to more detailed aspects of teachers' behavior. Longitudinal and quasi-experimental studies are needed in order to find causal effects. Cross-sectional studies can only provide a momentary glimpse of a phenomenon, and they do not allow for the testing of causal relationships between the de-

4 Interview quote from the study Garmannslund et al. (2008).

pendent and independent variables. Reversed causation may play a role, which implies that the direction of the arrows in the models could go in the opposite direction. Omitted variables may have influenced the overall situation, and variables that are missing from the model could be important. For instance, such factors as whether the school culture is characterized by openness to different approaches to teaching, as well as levels of trust between teachers and school leaders or between students and teachers, could affect the perceptions of an evaluation system. More longitudinal research is needed in order to address the complexity of the interactional dynamics between leaders and teachers and how they are related to teacher motivation (Firestone, 2014).

8. Implications for practice and further studies

Despite its limitations, this study contributes to our understanding of the factors that are related to teachers' perceptions of the usefulness of student ratings of teacher performance. The results from this study contribute to knowledge about the conditions related to teachers' perceptions of the usefulness of student ratings of teacher performance schemes. One such limitation is that we have not distinguished between teachers in vocational and general studies programs; these groups may perceive student ratings of teacher performance differently (Tarrou, 1997), and more research is required in this area. The recruitment of students also varies between these two types of programs. For example, the teachers' judgements of the students' level of maturity, with regard to their ability to provide well-considered feedback, may vary depending on whether the students are in vocational or general studies programs. In addition, the professional culture among teachers in vocational programs is quite different from that characteristic of teachers in general studies programs (Tarrou, 2003). We thus need more research on student ratings of teacher performance in different school contexts.

In its election program, the governing Norwegian Conservative Party has committed itself to permitting students in both lower secondary schools and high schools to evaluate their experience of teaching practices (The Conservative Party, 2013). As a consequence, national initiatives to implement student ratings of teacher performance in lower secondary schools may become relevant as well. Students at the lower secondary level are younger than those in high schools, and some issues remain to be clarified regarding how age is related to the assessment of various aspects of teaching quality, as well as how teachers relate to such feedback from students in younger age groups. However, this is not necessarily a problem; for instance, Fauth, Decristan, Rieser, Klieme and Büttner (2014) showed that student ratings can be useful measures of teaching quality even in a primary school context. Another question is: In what situations will student ratings of teacher performance be deemed meaningless by teachers (Tornero & Taut, 2010)? Future research may help us better understand the potential that student ratings of teach-

er performance may have in contexts other than general studies programs, which have been the focus of this study. We are of the opinion that the context within which student ratings of teacher performance is undertaken may have a significant effect on the way in which those subjects to evaluation respond to their feedback.

There are expectations in policy documents (GNIST, 2013) as well as research literature (Symposium, 2013) that student ratings of teacher performance will help improve educational practice, which in turn will improve students' learning outcomes. To clarify whether a system of student ratings of teacher performance will produce better learning, in a definitive way, would require a research design involving the comparison of a test group and a control group. Today, there is insufficient evidence to conclude that anonymous student ratings of teacher performance of the kind presented in this study will actually produce better learning outcomes, and this remains an issue for future empirical investigation.

We know less about how student ratings, in a teacher performance system, could help develop schools as so-called learning organizations. This study provides a limited empirical basis for assessing how schools can use student ratings of teacher performance, for example, in the subject group's efforts to improve teaching practices or the planning and follow-up of teaching (such as feedback to students on their academic performance). Further research should investigate how student ratings of teacher performance can be followed up through collective processes, of which collaboration among teachers would be a key aspect (Smylie, 2014). Since we suspect that the culture for collaboration among teachers may vary between subject groups, we recommend that a study be composed of several types of academic and vocational subjects to provide a better understanding of the collective or social dimension in the follow-up of student ratings of teacher performance.

Student ratings of teacher performance are only one source of data about teaching and could be used in combination with other sources of information (Hill & Grossman, 2013), for instance, alongside observations and value-added measures. Papay (2012) emphasizes that evaluations must provide teachers with a clear understanding not only of their current success or failure, but also of the practices they need to develop to become more successful with their students. If so, the system should promote continued teacher development to raise the instructional quality of existing teachers. More studies are needed to investigate multiple sources of feedback.

9. Conclusion

Student evaluation of teacher performance will not be effective if the feedback teachers receive misses important components that could contribute to their professional growth. Students see the teachers' teaching practices from a point of view that is quite different from that of colleagues or school administrators. This study has shown that student responses have the potential to provide teachers with use-

ful feedback about their educational practices. However, student ratings of teacher performance are not a magic bullet to enhance teachers' professional development. Using student ratings of teacher performance could become an exercise in navigating through demanding waters. Additional research is needed to clarify the extent of this potential, and how student evaluations can be used together with other data sources to give teachers optimal feedback (Peterson, Wahlquist, & Bone, 2000). How the schools and educational authorities will relate to these challenges is an empirical question that must be answered through more research.

Acknowledgements

This work was supported by the Research Council of Norway under Grant no. 237863. We thank all participating teachers for their engagement. We also thank two anonymous reviewers for comments made on an earlier draft. Any remaining errors, mistakes and omissions are solely our own.

References

Abrami, P. C., d'Apollonia, S., & Cohen, P. A. (1990). Validity of student ratings of instruction: What we know and what we do not. *Journal of Educational Psychology, 82*(2), 219–232.

Aleamoni, L. M. (1987). Typical faculty concerns about student evaluation of teaching. In L. M. Aleamoni (Ed.), *Techniques for evaluating and improving instruction: New directions for teaching and learning* (pp. 25–31). San Francisco, CA: Jossey-Bass.

Anderson, M. (2002). *Evaluating student performance in university level course work: The certification of academic accomplishment reveals a hidden conflict*. Retrieved from http://commons.erau.edu/bollinger-rosado/2002/10th/8/

Avalos, B., & Assael, J. (2006). Moving from resistance to agreement: The case of the Chilean teacher performance evaluation. *International Journal of Educational Research, 45*(4–5), 254–266.

Bakx, A., Koopman, M., de Kruijf, J., & den Brok, P. (2015). Primary school pupils' views of characteristics of good primary school teachers: An exploratory, open approach for investigating pupils' perceptions. *Teachers and Teaching, 21*(5), 543–564.

Benton, S. L., & Cashin, W. E. (2014). Student ratings of instruction in college and university courses. In M. Paulsen (Ed.), *Higher education: Handbook of theory and research* (pp. 279–326). Dordrecht, Netherlands: Springer.

Boring, A., Ottoboni, K., & Stark, P. B. (2016a). *Student evaluations of teaching are not only unreliable, they are significantly biased against female instructors*. Retrieved from http://blogs.lse.ac.uk/impactofsocialsciences/2016/02/04/student-evaluations-of-teaching-gender-bias/

Boring, A., Ottoboni, K., & Stark, P. B. (2016b). *Student evaluations of teaching (mostly) do not measure teaching effectiveness*. Retrieved from https://www.math.upenn.edu/~pemantle/active-papers/Evals/stark2016.pdf

Bryk, A., & Schneider, B. (2002). *Trust in schools: A core resource for school improvement*. New York, NY: Russel Sage Foundation.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31*(1), 16–30.

Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 41*(5), 511–535.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York, NY: Teachers College Press.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53*(3), 285–328.

Day, C., Flores, M. A., & Viana, I. (2007). Effects of national policies on teachers' sense of professionalism: Findings from an empirical study in Portugal and in England. *European Journal of Teacher Education, 30*(3), 249–265.

Delvaux, E., Vanhoof, J., Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching and Teacher Education, 36*, 1–11.

Deneire, A., Vanhoof, J., Faddar, J., Gijbels, D., & Van Petegem, P. (2014). Characteristics of appraisal systems that promote job satisfaction of teachers. *Education Research and Perspectives, 41*, 94–114.

Eagly, A. H., Ashmore, R. D., Makhijani, M. G., & Longo, L. C. (1991). What is beautiful is good, but …: A meta-analytic review of research on the physical attractiveness stereotype. *Psychological Bulletin, 110*(1), 109.

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17*(1), 101–128.

Elstad, E., Lejonberg, E., & Christophersen, K. A. (2015). *Teaching evaluation as a contested practice: Teacher resistance to teaching evaluation schemes in Norway*. Retrieved from http://dx.doi.org/10.3402/edui.v6.27850

Eriksen, B. (2012). *Rektors styringsrett* [School principals' right to govern]. Oslo, Norway: Gyldendal.

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.

Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers: What we know and what we don't. *Research in Higher Education, 9*(3), 199–242.

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 93–143). Dordrecht, Netherlands: Springer.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford, CA: Stanford University Press.

Firestone, W. (2014). Teacher evaluation policy and conflicting theories of motivation. *Educational Researcher, 43*(2), 100–107.

Flores, M. A. (2010). Teacher performance appraisal in Portugal: The (im)possibilities of a contested model. *Mediterranean Journal of Educational Studies, 15*(1), 41–60.

Flores, M. A. (2012). The implementation of a new policy on teacher appraisal in Portugal: How do teachers experience it at school? *Educational Assessment, Evaluation and Accountability, 24*(4), 351–368.

Fullan, M. (2014). *Teacher development and educational change*. New York, NY: Routledge.

Garmannslund, P. E., Elstad, E., & Langfeldt, G. (2008). Lærernes opplevelse av måling og rangering av kvalitetsaspekter ved undervisning og læringsprosesser [Teachers' perception of measures and ranking of qualities concerning teaching and learning processes]. In G. Langfeldt, E. Elstad, & S. T. Hopmann (Eds.), *Ansvarlighet i skolen* [School accountability] (pp. 250–270). Oslo, Norway: Cappelen Forlag.

GNIST. (2013). *Oppdragsbrev* [Assignment Letter] (Unpublished manuscript).

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Heck, R. H. (1992). Principals' educational leadership and school performance: Implications for policy development. *Educational Evaluation and Policy Analysis, 14*(1), 21–34.

Heneman III, H. G., & Milanowski, A. T. (2003). Continuing assessment of teacher reactions to a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(2), 173–195.

Hill, H., & Grossman, P. (2013). Learning from teacher observations: Challenges and opportunities posed by new teacher evaluation systems. *Harvard Educational Review, 83*(2), 371–384.

Hughes, J. N., Cavell, T. A., & Wilson, V. (2001). Further support for the developmental significance of the quality of the teacher-student relationship. *Journal of School Psychology, 39*(4), 289–301.

Hulpia, H., Devos, G., & Rosseel, Y. (2009). Development and validation of scores on the distributed leadership inventory. *Educational and Psychological Measurement, 69*(6), 1013–1034.

Isoré, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review* (OECD Education Working Papers, No. 23). Paris, France: OECD Publishing.

Jong, R. de, & Westerhof, K. J. (2001). The quality of student ratings of teacher behavior. *Learning Environments Research, 4*(1), 51–85.

Katsuno, M. (2010). Teacher evaluation in Japanese schools: An examination from a micro-political or relational viewpoint. *Journal of Education Policy, 25*(3), 293–307.

Kelly, K. O., Ang, S. Y. A., Chong, W. L., & Hu, W. S. (2008). Teacher appraisal and its outcomes in Singapore primary schools. *Journal of Educational Administration, 46*(1), 39–54.

Kline, R. B. (2005). *Principle and practice of structural equation modeling*. New York, NY: The Guildford Press.

Kulik, J. A. (2001). Student ratings: Validity, utility, and controversy. *New Directions for Institutional Research, 109*, 9–25.

Leithwood, K. (1994). Leadership for school restructuring. *Educational Administration Quarterly, 30*(4), 498–518.

Marsh, H. W. (2007). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry, & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence-based perspective* (pp. 319–383). Dordrecht, Netherlands: Springer.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement*. Alexandra, VA: ASCD.

Millman, J., & Darling-Hammond, L. (1989). *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. New York, NY: Corwin Press.

Mo, K. W., Conners, R., & McCormick, J. (1998). Teacher appraisal in Hong-Kong self-managing secondary schools: Factors for effective practices. *Journal of Personnel Evaluation in Education, 12*(1), 19–42.

Murray, H. G. (1997). Does evaluation of teaching lead to improvement of teaching? *The International Journal for Academic Development, 2*(1), 8–23.

OECD – Organisation for Economic Co-operation and Development. (2009). *Teacher evaluation. A conceptual framework and examples of country practices*. Paris, France: OECD.

OECD – Organisation for Economic Co-operation and Development. (2011). *OECD reviews of evaluation and assessment in education. Norway 2011*. Paris, France: OECD.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123–141.

Patzer, G. L. (2012). *The physical attractiveness phenomena*. Dordrecht, Netherlands: Springer Science & Business Media.

Penny, A. R. (2003). Changing the agenda for research into student's views about university teaching: Four shortcomings of SRT research. *Teaching in Higher Education, 8*(3), 399–411.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices*. New York, NY: Corwin Press.

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*(2), 135–153.

Smylie, M. A. (2014). Teacher evaluation and the problem of professional development. *Mid-Western Educational Researcher, 26*(2), 97–111.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research, 83*(4), 598–642.

Stronge, J. H. (2010). *Evaluating what good teachers do: Eight research-based standards for assessing teacher excellence*. New York, NY: Routledge.

Svinicki, M., & McKeachie, W. J. (2011). *McKeachie's teaching tips: Strategies, research, and theory for college and university teachers* (13th ed.). Belmont, CA: Wadsworth.

Symposium: Teacher effectiveness. The role of context. (2013). *Harvard Educational Review, 83*(2), 346–348.

Tarrou, A. L. H. (1997). *Yrkespedagogikk og yrkesfaglærerutdanning: en studie av utdanningskulturer* [Vocational education and vocational teacher education: A study of educational cultures]. Oslo, Norway: Universitetsforlaget.

Tarrou, A. L. H. (2003). La formation professionnelle initiale en Norvège: Entre valeurs démocratiques et partenariats économiques: La formation professionnelle initiale: Une question de société [Initial vocational training in Norway: Between democratic values and economic partnerships: Initial vocational training: A question of society]. *Revue Internationale d'*Éducation*, 34*(83), 127–136.

Theall, M., & Feldman, K. A. (2007). Commentary and update on Feldman's (1997) "Identifying exemplary teachers and teaching: Evidence from student ratings." In R. P. Perry & J. C. Smart (Eds.), *The teaching and learning in higher education: An evidence-based perspective* (pp. 130–143). Dordrecht, Netherlands: Springer.

The Conservative Party. (2013). *New ways, better solutions. The conservative party's program 2013–2017*. Oslo, Norway: The Conservative Party.

The Government. (2013). *Political platform for the government*. Retrieved from https://www.regjeringen.no/nb/dokumenter/politisk-plattform/id743014/

Tornero, B., & Taut, S. (2010). A mandatory, high-stakes national teacher evaluation system: Perceptions and attributions of teachers who actively refuse to participate. *Studies in Educational Evaluation, 36*(4), 132–142.

Tuytens, M., & Devos, G. (2010). The influence of school leadership on teachers' perception of teacher evaluation policy. *Educational Studies, 36*(5), 521–536.

Tuytens, M., & Devos, G. (2014). The problematic implementation of teacher evaluation policy: School failure or governmental pitfall? *Educational Management Administration & Leadership, 42*(4, Supplement), 155–174.

Weems, D. M., & Rogers, C. B. H. (2010). Are US teachers making the grade? A proposed framework for teacher evaluation and professional growth. *Management in Education, 24*(1), 19–24.

Willis, P. E. (1977). *Learning to labor: How working class kids get working class jobs*. Aldershot, England: Gower.

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699.

Zabaleta, F. (2007). The use and misuse of student evaluations of teaching. *Teaching in Higher Education, 12*(1), 55–76.