Pairwise comparison in teacher evaluation: feedback instead of competition

Paweł Wołoszyn
Department of Computational Systems
Cracow University of Economics
Cracow, Poland
pawel.woloszyn@uek.krakow.pl

Abstract—**A typical student evaluation questionnaire consisting of several questions with ordinal-scaled answers is commonly used as a tool of gathering students' feedback in teaching quality assurance systems in higher education in Poland and other countries. The numeric output of such survey often prompts usual methods of basic statistical analysis like calculating mean scores and arranging teachers in sorted rankings. However there are reasons for doubt whether it is both mathematically correct and beneficial for maintaining good teaching quality. We postulate that entirely different approach to gathering student feedback, based on pairwise comparison of teachers, could solve at least some of the problems noticeable in the traditional survey. We present arguments that adopting the proposed approach could provide more informative feedback while reducing competitive attitude to teaching.**

Keywords- Course evaluation, Elo rating system, intransitive preferences

I. INTRODUCTION

Quality assurance became an important concept in contemporary university practice where teaching is regarded more as a commercial service than art and tradition. Countries participating in the Bologna Process aim to maintain the quality of teaching on a standardized level in order to achieve comparability of education effects. Therefore higher education institutions adopt specific standards of quality assurance recommended by organizations such as ENQUA (European Association for Quality Assurance in Higher Education) [1]. Those standards recognize two levels of quality assurance, the internal and external level. Both levels provide different kinds of feedback loops allowing universities to diagnose and solve problems and to control further development of their educational offer.

The internal level feedback loop is the shorter one and therefore internal quality assurance measures are more willingly adopted in many universities, Polish higher schools among them. The shorter the feedback loop, the sooner the results are available without the need for waiting until students graduate and start their careers allowing to observe the effects of education in an external context. The most commonly implemented internal quality measurements are quantitative performance indicators [2], namely student evaluations conducted with the use of questionnaires. In typical scenario students are rating teachers and courses using quantitative scales, optionally providing additional comments and opinions. Such surveys are the basic element of quality assurance systems in Polish higher education [3] as well as in other countries.

Student evaluations of teachers and courses have been long criticized for being unreliable and misleading as they fail to adequately reflect teachers' engagement and provide hints on potential improvements [4]. Satisfaction surveys are cheap and easy way of collecting data, however they are strongly associated with customer-oriented model of education where students are regarded customers of educational services. Such a model is also criticized [5] and there are known examples showing that good teachers receive bad evaluations because their students prefer short term utility (less challenging and narrower teaching but better grades) over long term benefits (more work but better postgraduate career) [6]. Similarly, student evaluations of teachers are highly correlated with those students' current grades but not with their future achievements in follow-on courses [7].

The aim of this paper is however not to criticize nor defend student questionnaires as teacher evaluation tool. Instead we want to focus on quantitative nature of those surveys. It is possible to propose different mechanics of gathering evaluation data based on pairwise comparisons performed by students. It requires resigning from using numeric scales in surveys but it also provokes discussion whether quantitative scales are needed in processing and analyzing survey results as well as in providing conclusive feedback for teachers and faculties.

II. THE PROBLEM OF QUANTITATIVE SCALES

Typical student satisfaction survey consists of several closed questions asking student for rating a selected aspect of course and teaching quality with the use of ordinal scale. In Polish universities the scale commonly resembles traditional academic scale ranging from 2 (unsatisfactory) to 5 (very good) in case of questions which require students to provide their overall judgment. The other widely used type of questions offers some statement about course or teacher and requires the student to express his or her agreement or disagreement with the statement, usually in five-level Likert scale, for example from 1 (strongly disagree) to 5 (strongly agree). There are many possible variations of these scales obtained by transformation of mentioned numeric intervals, however it is important to emphasize that all such scales used in questionnaires are ordinal.

The exact formulation of questions and scale labels is itself a nontrivial problem lying in the overlapping area of pedagogy,

eISSN: 2453-7144, cdISSN: 1339-5572
ISBN: 978-80-554-1170-5

psychology and sociology. However a new class of problems arises when survey responses are processed and analyzed which often involves one or another form of data aggregation. The common use of computers and online surveys makes it very easy and straightforward to treat students' responses as raw numbers and process them accordingly. This leads to widely adopted but still controversial practice of treating Likert scale as interval measurement scale [8].

Ordinal scales indeed use numbers, but only to encode ordering relations and not to convey any information about intervals between items. The difference between 'strongly agree' and 'agree' cannot be compared to that between 'agree' and 'no opinion', yet still these values are treated as evenly spaced points on real numbers scale. This leads to contra-intuitive results when student ratings are unjustifiably aggregated by calculating arithmetic mean: when half of students strongly agree and the other half strongly disagree, the averaged response is 'no opinion'.

Ordinal scales are not necessarily linear and therefore linear combination of items such as their mean may have no meaningful interpretation. To avoid the problem a different central tendency measure should be used, for example median or mode value, which in turn can be problematic in case of even number of items or small number of observations. The best solution however is to avoid aggregating responses at all.

The need for aggregation arises, in our opinion, mainly because of the need for ordering courses and teachers according to their quality and performance. This is the impact of ubiquitous information technology. Computers are primarily designed to accept numbers, process numbers and output numbers. Therefore more and more aspects of everyday life become digitized and transformed to the realm of numbers. The ease of aggregating and sorting numbers with the help of computer causes widespread adoption of number-oriented conceptual frameworks in many disciplines, including education.

In the case of teaching quality assurance such number-centric approach leads to a paradox: quality becomes a quantity itself. The quality of teaching, by the means of specific measurement process (numerically encoded survey responses), becomes digitized and enters the typical chain of data processing aimed at reducing complexity and generating simple and comprehensible results. From the managerial point of view the most simplified result is a single number representing teacher performance and course quality. Without aggregation the results would be much more complex in interpretation as they should contain entire distributions of responses to each question in a survey. Not aggregated results would be the most faithful and undistorted output from quality measuring procedure but of course they would also have little or no use for constructing a ranking of teachers.

The notion of a teacher or course being 'better' or 'worse' than another immediately brings out an imagination of total linear ordering. Teachers become comparable entities and their positions on linear scale are given by single numbers. It creates the opportunity to construct entire rankings of teachers – an opportunity eagerly taken by universities and their quality assurance divisions which often publish top portion of teacher ranking as didactic 'hall of fame' in order to motivate teaching staff to improving their skills and performance.

Such rankings are however inherently flawed because teacher ratings obtained with a questionnaire have impenetrable upper bound determined by the highest numeric value assigned to the most appreciative responses in the survey. Teachers are evaluated independently, without the context of their peers and it is not impossible for more than one good teacher to get the highest possible rating. It does not mean that they are equally good, it only means that they cannot be distinguished within the numeric scale used and they cannot further improve their ratings thus reaching the end of their measurable development as teachers. It also means that there is such implicitly defined concept as a perfect teacher and moreover the perfection is really achievable, not asymptotically but linearly. The perfect teacher is the one who gets the highest possible grades in quality assurance survey, therefore the characteristics of the perfect teacher can be reversely deducted from survey questions even if university authorities and survey authors have never considered such a concept.

The problem with closed ordinal scales is caused by the very mechanics of survey which essentially asks the student a question: 'how close is the teacher to the ideal role model?' The way of aggregating the data (for example by calculating means) and sorting the results proves even further that measured quantities are treated as linear distances, which is otherwise quite natural interpretation of rational numbers. It is impossible to avoid this problem as long as quality measurement relies basically on assigning each teacher a couple of numbers drawn from a small set of available values. Then it seems interesting to consider different quality assessment mechanics while still maintaining the ability to construct a rating.

III. PAIRWISE COMPARISONS

Sport is a well known area of human activity where there is no concept of perfection and instead all judgments are based on comparisons. Sport competition is always relative and uses open rankings without upper or lower bounds imposed, at least in sports involving physical measurements of time, distance or other magnitude. It guarantees that it is always possible to improve a result and beat a previously established record. It also allows for running a competition regardless of participant skills, even if all of them are on the same novice or master level. In fact it depends primarily on the precision of measurement how close two results can be to each other in order to remain still distinguishable. The measurement in such sports produces rational numbers in truly continuous scale without any need for aggregation.

Therefore it seems that adopting similar comparison-based mechanics in teaching quality assessment could solve at least some of the problems discussed above. Teachers could be ranked and ordered not by their absolute performance but by relative achievements in comparison to their peers. It would still allow for maintaining the 'hall of fame' tradition, which managers and university authorities are so much used to. It would also match the contemporary pervasive competition paradigm which manifests itself in countless rankings and

eISSN: 2453-7144, cdISSN: 1339-5572
ISBN: 978-80-554-1170-5

scores in games, entertainment, social media, work culture, education or even science itself.

However, comparing teachers does not resemble the sport of athletics as there is no objective measure of educational proficiency analogous to physical time or distance. If students were to grade their teachers on a purely rational numeric scale just like in such sports, it would force them to internally aggregate their opinions into a single number. Fortunately there are other kinds of sports which do not utilize continuous measurements and instead use binary distinction of 'winner' versus 'defeated'. A good example is chess. There is no such concept as the best possible chess player and the only way of determining who is the World Champion is to confront players in pairs in as many matches as possible.

Pairwise comparison is rarely used in education for several reasons [9], including practical issue of monotony accompanying large number of pairs being compared. However we believe that the most important obstacle is the lack of immediate numeric output from comparison procedure. The mere fact of winning a chess game is insufficient to determine to what extend the winner's skills surpass those of the defeated player. Similarly if a student decides that one teacher is better than another it does not help in assigning those teachers precise positions in a ranking. It is then necessary to implement additional ranking system for converting binary match outputs to a numeric scale.

There is one well established and widely adopted ranking system used in chess competitions, namely Elo rating system [10]. Primarily designed as a specialized system tailored to the specific needs of chess players, it found numerous other applications, for example in online gaming or social media. Elo system assigns each player a number representing his or her ranking. Traditionally it is an integer value ranging from zero to a few thousands (currently the highest FIDE rating ever recorded does not exceed 3000). Any single rating itself is meaningless, but difference between ratings of two players predicts the expected outcome if those players would play a tournament against each other. If the tournament indeed takes place, the actual result may be different than predicted and ratings of both players should be adjusted to counterbalance the error: underestimated player gains points and his rating increases and vice versa.

It is important feature of Elo ratings that they have no theoretical limits on both ends of scale which is purely relative as there is no fixed reference point. It is known phenomenon that distribution of rating values among worldwide population of chess players changes over time and contemporary ratings are difficult to compare with ratings from past decades. But on the other hand, there is little sense in comparing skills of two chess masters from different historic periods if they never played a tournament against each other.

These features make Elo rating system an interesting alternative to traditional measurement of teacher performance. Teacher's knowledge, attitude, experience, communication and social skills also have no reference point or fixed scale and so they should be rated only relatively. The teacher with highest Elo rating is not a living example of perfection but only a teacher who has the highest so far recorded probability of being perceived better than other teachers. Moreover the highest rating still can be improved or beaten by another teacher.

It seems that problems with ordinal scales discussed in the first part of this paper could be solved by adopting pairwise comparison followed by Elo rating (or any similar rating scheme, for example the more sophisticated Glicko system [11]). In order to use such system the very mechanics of teaching quality assessment must be changed with students no longer rating individual teachers in questionnaires but instead comparing pairs of teachers with respect to overall perceived quality of teaching. The essential question students should answer becomes then: 'which one of these two teachers is preferred more?'

Answering thus posed question should be in fact easier for students than assigning numerical score in several categories. The latter calls for 'objective' rating which, at least implicitly, motivates respondents to precisely justify their reasons for a given score. Pairwise comparison on the other hand is entirely subjective which is more natural in the feedback loop between students and teachers because students often like or dislike their teachers according to their own intrinsic value system and not an official ideal model of professional educator. If quality assessment is meant to improve students experience then it should rely more on subjective feelings than objective measurements. If those subjective feelings are insufficient for a student to decide which one of a pair of teachers is preferred, the option of drawing a game is also provided in Elo system. Indeed, many chess tournaments end in draws.

IV. Partial Ordering in Ratings

Although Elo rating system seems promising, it has also some important drawbacks. One minor issue is related to continuous nature of rating. The system was originally devised to track player rating throughout his or her entire career, with small adjustment made every time the player is engaged in a tournament with another player. Adjustments must be small relatively to rating scores to avoid violent oscillations and to ensure that ratings stabilize over time. The precise magnitude of adjustment depends on previous ratings of both players and therefore players have to be already rated before entering the tournament (or at least some provisional ratings must be assumed if players are not rated yet). One can observe that the process has strong memory of past states.

Teaching quality assessment systems, on the contrary, usually have little or no memory of past ratings. Teachers are rated afresh in each semester and students are not asked to consider their previous opinions. The ratings are established not by the means of continuous adjustment but by periodic independent measurements. Such situation is not fully compatible with Elo system because ratings are not given a chance to stabilize before they are reset to default initial value. Adopting the Elo scheme would require changing the measurement model and tracking each teacher's rating since the beginning of teaching career until retirement.

A more important issue concerns the number and diversity of games. In pairwise teacher comparison confronting two teachers can be thought of as a tournament, each choice made by a single student being a game withing that tournament. In

eISSN: 2453-7144, cdISSN: 1339-5572
ISBN: 978-80-554-1170-5

chess it is possible, at least theoretically, for any player to match with whichever opponent he or she chooses. Since rating adjustments are symmetric and one player gains the same amount of points the other one looses, the diversity of player matching creates an entire ecosystem of ratings with points flowing freely between players on truly global scale. The total pool of points remains constant, at least until new players enter or old players leave the system. This ensures that Elo ratings are consistent and comparable worldwide.

Comparing teachers is much more constrained because the opponents in a match cannot be freely chosen. Students are able to compare only those teachers who have taught them. It never happens in a university that every student knows every teacher, instead both populations are usually partitioned into several isolated groups according to faculty structure and teaching programs. Such topology of academic society is bound to produce tightly knit clusters in the rating procedure with teachers compared frequently within clusters but almost never across them. Each cluster would have a separate pool of points and it would be meaningless to confront rating values from different clusters. It would be no longer possible to maintain a single 'top 10' list of best teachers in the entire university, instead each faculty and study curriculum would generate its own 'hall of fame' with disclaimer that the worst teacher in one faculty could still be compared superior to the best teacher in another, if only they had shared the same group of students.

The lack of comparability between teacher ratings could be interpreted as a serious flaw in potential applications of Elo system in teaching quality assessment. However we would like to present the entirely opposite point of view: rating incomparability exposes an important weakness in traditional 'hall of fame' approach. Unlike sportsmen, teachers are not rivals and their profession does not call for competition. The ultimate task for every teacher is to help students explore the world of science and not to prove superiority over other teachers. Then it should be normal and expected situation that teachers are incomparable except those rare occasions when two or more teachers share the same group of students.

Competitive approach to teaching in our opinion is caused artificially by imposing total ordering on the set of university educational staff through assigning a number score to every teacher. The more realistic approach would correspond to mathematical concept of partially ordered set where not all pairs of elements are given a precedence relationship. Such concept is used only to a little extent in university practice where language instructors and physical education teachers are usually rated independently of the rest of teaching staff. Nevertheless adopting the pairwise comparison scheme would require a complete revision of quality assurance system and admitting the fact that only those teachers who work with the same students could be compared to each other.

On the other hand even in the traditional approach it is disputable whether teachers with scores given in an ordinal scale are indeed comparable across study programs and faculties. Students inherently judge their teachers relatively to other educators they know. If they credit some teacher with the best score on a Likert-like scale, it does not guarantee that they would choose still the same score if they had known another yet better teacher.

V. INTRANSITIVE PREFERENCES

Comparability of teacher ratings causes at least one more trouble associated with general assumption that quality assessments are transitive. Transitivity of preferences, or the lack of it, is a long known research motif in science concerned with making decisions and rational choices [12]. It seems reasonable that if A is a better teacher than B and B is still better than C, then A is definitely better teacher than C. In other words, if students prefer A over B and B over C, then if given only A and C for choice they would choose A. Although both statements seem equivalent, it may be no longer true if the second version is based on more sophisticated preferences than simple comparison of magnitudes. Research findings suggest that intransitive preferences are not only a theoretical possibility but they are also intrinsic feature of human judgment reflected in the very structure of brain [13].

Intransitivity can occur regardless of rating model used for comparing teachers, both with ordinal scales as well as with pairwise comparison. In the case of ordinal scales it can happen if teachers are rated in several aspects independently, which is a common practice in universities. For example if three teachers are rated in three categories it is easy to assign scores (by cyclically shifting values) in such a way, that every teacher in two categories has higher ratings than one of the other two teachers, and the relation is cyclic. It is impossible to pick a best one of them because regardless of the choice there would always be another candidate who is superior in majority of rated categories.

To deal with such intransitivity one can try to convert several rating categories into a single scalar score, either by aggregating categories (typically by averaging) or by introducing additional 'overall' score. However it can only conceal the problem, as the single value would no longer reflect real preferences. In the example of cyclic shifted scores the average overall rating of each teacher is the same suggesting that all three teachers are performing equally good, which is obviously not true if only two of them are considered.

As a side note, the example also illustrates why it would be wrong to generalize comparison procedure and ask students to choose preferred teachers from triples or higher length tuples: although faster and less repetitive than judging each pair separately, it could make it impossible for students to decide who is the most preferred teacher according to their authentic feelings. Instead it would force them to use some artificial, fully transitive criterion.

Due to intransitivity of preferences the principal question asked in teaching quality survey performed by pairwise comparison could be stated more precisely: 'if there were only these two teachers employed in the faculty, which one of them would you prefer more?' It expresses clearly that any other candidates must not be considered in current comparison, although they will appear in other pairs.

Intransitive preferences are not compatible with Elo rating system for the obvious reason: the main goal of the system is to convert binary preference pairs to scalar ratings which are

eISSN: 2453-7144, cdISSN: 1339-5572
ISBN: 978-80-554-1170-5

transitive. This makes it more difficult to process and analyze the results of pair-comparing survey. A different approach is therefore needed for modeling the entire network of students' preferences. Perhaps the most straightforward way to organize survey responses is to construct a directed graph with nodes corresponding to teachers and edges denoting preferences. If many responses for each pair of nodes are collected, the graph becomes weighted and each edge weight is determined by the fraction of students who prefer one teacher more than the other. Such a graph essentially matches the model for intransitive preferences proposed in [14].

Adopting a graph-based model opens new possibilities of analyzing survey results although it implies abandoning traditional rankings, scores and statistics. For example it is no longer meaningful to consider an average teacher performance in an academic faculty or to compare any teacher with that hypothetical average. Nonetheless there are many interesting new features that can be observed and measured in a graph model such as connectedness of students' preferences, possible partitioning or the presence of cycles and cliques. The exact meaning of these features and their value as predictors of teacher performance are yet to be learned, as we do not know in the moment of writing this paper about any university adopting similar model for internal teaching quality evaluation.

VI. Conclusion

Undeniably the best method of obtaining higher education teaching quality feedback on the internal level would be to thoroughly interview each student and ask for his or her opinion about each teacher. It would provide the most complete information and allowed for all kinds of analysis and reflection. However such an idealistic approach is too expensive and time-consuming. For the sake of commonly used information technology student opinions are usually gathered in a most succinct and computer-friendly form of numbers, occasionally accompanied by textual remarks.

This ubiquity of numbers leads to overly enthusiastic use of scalar grades and sequential rankings, which in turn create an impression that teaching is a highly competitive profession. But in fact student evaluation of academic courses is not a voting for the champion teacher. It is inevitable, if not desirable, that students are taught by diversified staff with different personalities, attitudes, experience and style. Too much emphasis on continuous improving teacher ratings could be even detrimental to the quality of education [15].

This suggests that the original purpose of course evaluation should be reconsidered. The essential goal of teaching quality assurance is to find out what should be done to raise or at least maintain the level of quality. It is important to note that this is a matter of 'what' instead of 'how much'. Every education facility is not a simple physical system like a water boiler and there are no simple feedback loops like in a thermostat. Quantitative measures obtained in questionnaires can fail to provide an advice on how to improve teaching quality, as exemplified in [4], and in our experience optional comments written by students only sporadically contain anything more than generic approval or critique.

We believe that the quality feedback loop must have, unsurprisingly, more qualitative nature. From this perspective the method of pairwise comparison considered in this paper can be seen as a compromise between qualitative interviews and quantitative questionnaires. Instead of measuring how close are the teachers to some abstract archetype, pair matches can give more insight into students' subjective sentiments and preferences. The insight could be even more reliable because the results of pair comparisons are harder to manipulate for the respondent in case of emotional bias or intentional revenge. Still, the data collected in pairwise comparisons are computer-friendly and can be easily processed with the use of well known graph algorithms.

The task of comparing teachers, due to its truly subjective nature, is addressed directly to students who become a little more active party in the entire quality assurance system. In contrast, traditional questionnaires typically treat students as mere observers asking them for example about punctuality of teachers or the usage of multimedia and teaching aids. Such questions could be answered by anyone who attends the courses or even, as in case of punctuality, by dedicated computerized system. This kind of survey depreciates the role of students [16] reducing it to an alternative to employing school inspectors.

It should not be expected that the pairwise comparison approach would produce immediate recipe for improving teaching quality. Nevertheless it opens possibilities of finding and understanding new factors influencing the quality of teaching perceived by students. The most natural question to be asked during analysis of student responses is why do they prefer some teachers more than other. The approach proposed here gives much flexibility in answering such posed question. First, compared teachers can discuss the reasons between themselves sharing their observations, experience and teaching methods which prove most successful. Second, faculty authorities can analyze the entire graph of preferences searching for higher level patterns and consulting them with teachers in order to establish best practice standards and recommendations.

Third, students can explain their choices in additional comments recorded in the same survey, possibly augmented with a computer system dynamically deciding which choices need explanation thus avoiding too much comments and information overload. While in most situations it is awkward to ask about the exact reason behind choosing one or another item on Likert scale, it is quite natural to ask about reason for preferring one teacher over another.

All of the above three options share an important trait: they put emphasis on exploratory approach to teaching quality assessment. In our opinion implementing pairwise comparison scheme in course evaluation could shift the accent from competitive 'who is better' contest to more informative 'how to improve' debate thus providing useful feedback for teachers and educational institutions.

References

[1] Standards and Guidelines for Quality Assurance in the European Higher Education Area (ESG), Brussels, Belgium, 2015.

[2] H. Fabrice, "Learning Our Lesson: Review of Quality Teaching in Higher Education," OECD Publishing, vol. 2010, no. 2., pp. 79-98, 2010.

[3] M. Cieciora, "An attempt to analyse the implementation of higher education quality assurance standards in Poland: A case study," Zeszyty Naukowe Małopolskiej Wyższej Szkoły Ekonomicznej w Tarnowie, vol. 20, no. 1, pp. 49-62, 2012.

[4] C.R. Emery, T. R. Kramer, and R. G. Tian, "Return to academic standards: A critique of student evaluations of teaching effectiveness," Quality assurance in Education, vol. 11, no. 1, pp. 37-46, 2003.

[5] D. Bay, and H. Daniel, "The student is not the customer—An alternative perspective," Journal of Marketing for Higher Education, vol. 11, no. 1, pp. 1-19, 2001.

[6] M. Braga, M. Paccagnella, and M. Pellizzari, "Evaluating students' evaluations of professors," Economics of Education Review, vol. 41, pp. 71-88, 2014.

[7] S. E. Carrell, and J. E. West. "Does professor quality matter? Evidence from random assignment of students to professors," Journal of Political Economy, vol. 118, no. 3, pp. 409-432, 2010.

[8] S. Jamieson, "Likert scales: how to (ab)use them," Medical education, vol. 38, no. 12, pp. 1217-1218, 2004.

[9] S. Heldsinger, and S. Humphry, "Using the method of pairwise comparison to obtain reliable teacher assessments," The Australian Educational Researcher, vol. 37, no. 2, pp. 1-19, 2010.

[10] A. E. Elo, The Rating of Chessplayers, Past and Present, Ishi Press International, 2008 (original edition: Arco Pub., 1978).

[11] M. E. Glickman, Example of the Glicko-2 system, Boston University, 2012.

[12] M. Regenwetter, J. Dana, and C. P. Davis-Stober, "Transitivity of preferences," Psychological Review, vol. 118, no. 1, pp. 42-56, 2011.

[13] T. Kalenscher, P. N. Tobler, W. Huijbers, S. M. Daselaar, and C. M. Pennartz, "Neural signatures of intransitive preferences," Frontiers in Human Neuroscience, vol. 4, 2010.

[14] S. Saarinen, C. Tovey, and J. Goldsmith, "A Model for Intransitive Preferences," Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[15] M. Gray, and B. R. Bergmann, "Student teaching evaluations," Academe, vol. 89, no. 5, pp. 44-46, 2003.

[16] M. Platt, "What student evaluations teach," Perspectives on Political Science, vol. 22, no. 1, pp. 29-40, 1993.