3rd International Conference on Higher Education Advances, HEAd'17
Universitat Politècnica de València, València, 2017
DOI: http://dx.doi.org/10.4995/HEAd17.2017.5583

An agreement-based approach for reliability assessment of Students' Evaluations of Teaching

Vanacore, Amalia; Pellegrino, Maria Sole
Department of Industrial Engineering, University of Naples "Federico II", Italy

Abstract

Students' Evaluations of Teaching (SETs) are the most common way to measure teaching quality in Higher Education: they are assuming a strategic role in monitoring teaching quality, becoming helpful in taking the major formative and summative academic decisions. The majority of studies investigating SETs reliability focus on the instruments and the procedures adopted to collect students' evaluations rather than on the capability of the students as teaching quality assessors. In order to overcome this lack, a study has been carried out with the aim of measuring SETs reliability in terms of inter-student agreement and intra-student agreement. The results of our study show that the majority of students provided substantially repeatable evaluations whereas only a few students provided almost perfectly repeatable evaluations; the evaluations provided by different students generally slightly agreed, which means that the students did not share the same opinions and beliefs on teaching quality.

Keywords: *teaching quality assessment; reliability; inter-student agreement; intra-student agreement.*

1. Introduction

Measuring the student experience is assuming increasingly importance in Higher Education (hereafter, HE) representing a widespread method for evaluating teaching quality whose importance is relevant for taking the major formative and summative academic decisions (Berk, 2005; Gravestock & Gregor-Greenleaf, 2008; Onwuegbuzie *et al.*, 2009).

Student ratings, also known as Student Evaluations of Teaching (SETs), have dominated as the primary measure of teaching quality over the past 40 years (*e.g.*, Centra, 1979; Seldin, 1999; Emery *at al.*, 2003; Gaertner, 2014) forming the basis for the rankings of HE institutions. Although widely used, SETs are one of the most controversial and highly-debated measures of teaching quality: many researchers argue that there is no better option that provides the same sort of quantifiable and comparable data on teaching quality (McKeachie, 1997; Abrami, 2001) but, on the opposite, others point out significant biasing factors for SETs.

The fear that students cannot provide reliable teaching quality evaluations is, by far, one of the primary concerns about SETs. As a matter of fact, even highly motivated students can base their current evaluations on their past teaching experience, which can substantially vary depending on the college or university attended and/or on the student individual belief toward the degree (Ackerman *et al.*, 2009). Students who are generally satisfied/dissatisfied with the course and/or the instruction can bias the results upward/downward (Sliusarenko *et al.*, 2013). In addition, it is known that demographic (*e.g.*, gender and age; Thorpe, 2002; Fidelman, 2007; Kherfi, 2011) as well as logistic (*e.g.*, class size; Kuo, 2007) factors can influence SETs. The above considerations call into question the opportunity to consider the students as able to provide reliable evaluations on teaching quality. For this reason, differently from the majority of available studies, which rather focus on the instruments and the procedures adopted to collect SETs, our study aims at investigating the peculiar abilities of the students as teaching quality assessors by measuring SETs reliability in terms of inter-student and intra-student agreement. Particularly, the former allows evaluating the students' ability to provide the same score, on average, as the other students whereas the latter, also known as *repeatability*, allows evaluating the students' ability to score consistently a given quality item in different occasions.

2. Measuring inter-student and intra-student agreement: kappa-type indexes

The easiest approach for assessing the degree of agreement among repeated evaluations would be to simply calculate the observed agreement. This approach, however, provides a biased measure of agreement, especially when a rating scale with a few categories is adopted. In order to avoid this problem, inter-student and intra-student agreement will be assessed using the well-known kappa-type indexes, where the observed agreement is corrected for the agreement expected by chance. Specifically, the degree of inter-student agreement is assessed by calculating the s statistic proposed by Marasini *et al.* (2014), that is a rescaled measure of the probability of observed agreement p_a^s corrected with the probability of agreement expected by chance alone $p_{a|c}^s$:

$$s = (p_a^s - p_{a|c}^s) \big/ (1 - p_{a|c}^s) \tag{1}$$

Being r the number of students who rated twice (*i.e.* replications) the same n quality items on a $k \geq 3$ points ordinal scale, r_{hi} and r_{hj} the number of students who assigned the h^{th} quality item into i^{th} and j^{th} category during first and second replication, respectively; w_{ij} the corresponding weight, introduced in order to account that some disagreements (*i.e.* on categories that are at least two steps apart) are more serious than others (*i.e.* on neighboring categories), the observed proportion of agreement and the proportion of agreement expected by chance alone can be obtained as:

$$\hat{p}_a^s = \frac{1}{n}\sum_{h=1}^{n}\hat{p}_h; \quad p_{a|c}^s = \frac{1}{k} + \frac{1}{k^2}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} w_{ij} \tag{2}$$

where $\hat{p}_h$ is the proportion of agreement on h^{th} quality item given by:

$$\hat{p}_h = \left(\sum\nolimits_{i=1}^{k} r_{hi}\left(r_{hi} - 1\right) + 2\sum\nolimits_{i=1}^{k-1}\sum\nolimits_{j=i+1}^{k} w_{ij} r_{hi} r_{hj}\right) \Big/ \left(r\left(r-1\right)\right) \tag{3}$$

The degree of intra-student agreement, instead, is assessed using the weighted version of Brennan-Prediger coefficient (1981) proposed by Gwet (2014), that is a rescaled measure of the probability of observed agreement p_a corrected with the probability of agreement expected by chance alone $p_{a|c}$:

$$K_W^U = (p_a - p_{a|c}) \big/ (1 - p_{a|c}) \tag{4}$$

The chance measurement system adopted in Brennan-Prediger coefficient is the uniform one. Being n the number of quality items rated twice on a $k \geq 3$ points ordinal scale by the same student, n_{ij} the number of quality items classified into i^{th} category in the first replication and into j^{th} category in the second replication, the observed proportion of agreement $\hat{p}_a$ and the proportion of agreement expected by chance alone $p_{a|c}$ are:

$$\hat{p}_a = \sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{1} n_{ij} w_{ij}; \quad p_{a|c} = \left(\sum\nolimits_{i=1}^{k}\sum\nolimits_{j=1}^{1} w_{ij}\right) \Big/ k^2 \tag{5}$$

The values of kappa-type indexes range between -1 and 1, with negative values meaning disagreement. The index magnitude can be interpreted by adopting the Landis and Koch (1977) benchmark scale. According to this scale, there are 5 categories of agreement corresponding to as many ranges of coefficient values: slight, fair, moderate, substantial and almost perfect agreement for coefficient values ranging between 0 and 0.2, 0.21 and 0.4, and 0.41 and 0.6, 0.61 and 0.8 and 0.81 and 1.0, respectively.

3. Case Study

The case study was conducted at the Department of Industrial Engineering of University of Naples "Federico II" and consisted of 3 supervised experiments (hereafter, E.1, E.2, E.3) carried out on classes of students attending the course of Statistical Quality Control (SQC) in 3 successive academic years. All three involved classes included more than 20 students; all of them obtained the first level degree in Management Engineering from the University of Naples "Federico II" and thus they can be reasonably assumed homogeneous in curriculum and instruction.

Students were asked to fill two evaluation sheets (each with a specific rating scale) in order to collect their quality evaluation for a set of $n = 20$ items (regarding, for example, organization, workload and readings) of the SQC course they were attending. The first evaluation sheet used a Numeric Rating Scale (NRS) with scores ranging from 0 to 10 whereas the other used a Verbal Rating Scale (VRS) with agreement grades: "strongly disagreeing with the statement", "slightly agreeing with the statement", "quite agreeing with the statement" and "strongly agreeing with the statement". For comparability purposes, students' evaluations on the NRS were rescaled to the 4-points VRS using the following cut-off ranges: 0 to 2, 3 to 5, 6 to 8 and 9 to 10.

Each experiment consisted of two sessions: the first evaluation session (*i.e.*, S.I) took place at mid-term course and the second evaluation session (*i.e.*, S.II) took place the following lesson. Between S.I and S.II there was no new lesson and no interaction with the teacher, therefore no change in quality evaluation was expected. In order to guarantee evaluation traceability while preserving anonymity, each student signed her/his evaluation sheets with a nickname, which enabled to match student's ratings provided in the two evaluation sessions in order to estimate intra-student agreement. Only those students who rated all quality items in both experimental sessions were retained as participants in the study (*viz.* 17 students in E.1, 18 students in E.2 and 17 students in E.3).

The collected data were used to estimate the inter-student and intra-student agreement on NRS (hereafter, $\hat{s}_{\text{NRS}}$ and $\hat{K}^{U}_{W|\text{NRS}}$, respectively) and the inter-student and intra-student agreement on VRS (hereafter, $\hat{s}_{\text{VRS}}$ and $\hat{K}^{U}_{W|\text{VRS}}$); the intra-student agreement coefficients were both computed adopting the linear weighing scheme (Cicchetti & Allison, 1971).

3.1. Study results

The value of $\hat{s}_{\text{NRS}}$ and $\hat{s}_{\text{VRS}}$ for E.1, E.2 and E.3 are reported in Table 1.

Table 1. Inter-student agreement on NRS and VRS

Experiment	E.1	E.2	E.3
$\hat{s}_{\text{NRS}}$	0.395	0.300	0.600
$\hat{s}_{\text{VRS}}$	0.380	0.528	0.277

The results for intra-student agreement for each student participating in E.1, E.2 and E.3, are reported in Table 2 and plotted in Figures 1 against the 5 regions of intra-student agreement on NRS and intra-student agreement on VRS identified according to the Landis and Koch's benchmark scale.

Results in Table 1 highlight that the inter-student agreement is at most moderate, so that it is not possible to assume that the involved students shared the same opinions about teaching quality; the difference between the two rating scales is irrelevant only for students of E.1, however results do not allow preferring a rating scale over the other.

The intra-student agreement was generally higher than the inter-student agreement: 73% of students were at least substantially repeatable on both NRS and VRS whereas 19% of them were even almost perfectly repeatable on both NRS and VRS. In addition, the majority of students show over the years values of $\hat{K}^{U}_{W|\text{VRS}}$ higher than those of $\hat{K}^{U}_{W|\text{NRS}}$ although for about half of them the repeatability on the two rating scales belong to the same agreement categories and only for few (*i.e.*, 10) students $\hat{K}^{U}_{W|\text{NRS}}$ and $\hat{K}^{U}_{W|\text{VRS}}$belong to no-adjacent categories of agreement.

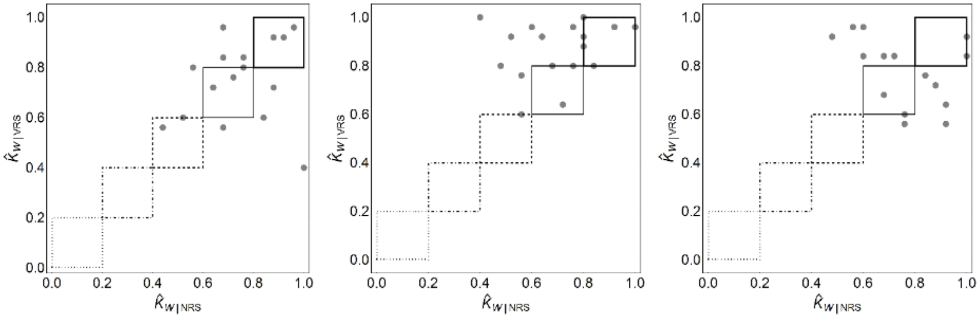

Figure 1. Intra-student agreement on NRS (as abscissa) and VRS (as ordinate) for each student participating in E.1. (on the left), E.2. (in the middle) and E.3. (on the right)

Table 2. Intra-student agreement on NRS ($\hat{K}^{U}_{W|\mathrm{NRS}}$) and VRS ($\hat{K}^{U}_{W|\mathrm{VRS}}$)

	E.1		E.2		E.3	
Student	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$	$\hat{K}^{U}_{W\|\mathrm{NRS}}$	$\hat{K}^{U}_{W\|\mathrm{VRS}}$
1	0.76	0.80	0.56	0.76	0.92	0.56
2	0.88	0.72	0.80	0.92	0.56	0.96
3	0.68	0.96	0.48	0.80	0.72	0.84
4	1.00	0.40	0.84	0.80	0.60	0.96
5	0.68	0.84	0.52	0.92	0.84	0.76
6	0.76	0.84	1.00	0.96	0.88	0.72
7	0.92	0.92	0.64	0.92	0.68	0.68
8	0.96	0.96	0.76	0.80	0.68	0.84
9	0.64	0.72	0.60	0.96	0.60	0.84
10	0.44	0.56	1.00	0.96	0.76	0.60
11	0.72	0.76	0.56	0.60	0.48	0.92
12	0.84	0.60	0.92	0.96	0.92	0.64
13	0.76	0.84	0.80	1.00	0.72	0.84
14	0.56	0.80	0.72	0.64	0.76	0.56
15	0.68	0.56	0.80	0.88	1.00	0.84
16	0.52	0.60	0.68	0.80	1.00	0.92
17	0.88	0.92	0.40	1.00	1.00	0.92
18			0.76	0.96		

4. Conclusions

This research aimed at investigating the reliability of Students' Evaluations of Teaching by evaluating intra- and inter-student agreement.

With respect to intra-rater agreement, the results of our study highlight that, on average, the 65% of involved students could be considered repeatable assessors of teaching quality, since they provided quality evaluations that were consistent over time. Specifically, for NRS, the percentage of at least substantially repeatable students ranges, across the three experiments, between 66% and 82%, whereas, for VRS, the percentage of at least substantially repeatable students ranges between 71% and 94%. These results seem to suggest that even if the NRS is the most common rating scale, the students were able to express their opinion more consistently using a verbal rather than a numeric rating scale.

On the other hand, focusing on inter-student agreement, results seem to suggest that the whole class of students could not be considered homogeneous in terms of beliefs and/or opinions and/or knowledge about teaching quality, being the inter-student agreement at most moderate, independently of the specific class of students and the adopted rating scale.

The obtained results cannot of course be generalized since, although the experiments were repeated over three academic years, they involved only students attending the same course. In order to overcome this weakness, an interesting development could be to conduct the same experiment on different university courses.

References

Abrami, P. C. (2001). Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms. *New Directions for Institutional Research*, 2001(109), 59-87.

Ackerman D., Gross B.L. & Vigneron F. (2009). Peer Observation Reports and Student Evaluations of Teaching: Who Are the Experts?. *The Alberta Journal of Educational Research*, 55(1), 18-39.

Berk R. A. (2005). Survey of 12 Strategies to Measure Teaching Effectiveness. *International Journal of Teaching and Learning in Higher Education*, 17(1), 48-62.

Brennan, R. L., & Prediger, D. J. (1981). Coefficient Kappa: Some Uses, Misuses, and Alternatives. *Educational and Psychological Measurement*, 41, 687–699.

Centra, J. A. (1979). *Determining Faculty Effectiveness. Assessing Teaching, Research, and Service for Personnel Decisions and Improvement*. Jossey-Bass Publications, ERIC Number: ED183127.

Cicchetti, D. V., & Allison, T. (1971). A new procedure for assessing reliability of scoring EEG sleep recordings. *American Journal of EEG Technology*, 11(3), 101-110.

Emery, C. R., Kramer, T. R., & Tian, R. G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education*, 11(1), 37–46.

Fidelman, C.G. (2007). *Course Evaluation Surveys: In-class Paper Surveys Versus Voluntary Online Surveys.* ProQuest.

Gaertner, H. (2014). Effects of student feedback as a method of self-evaluating the quality of teaching. *Studies in Educational Evaluation*, 42, 91-99.

Gravestock, P., & Gregor-Greenleaf, E. (2008). *Student course evaluations: Research, models and trends.* Toronto: Higher Education Quality Council of Ontario.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.

Kherfi, S. (2011). Whose Opinion Is It Anyway? Determinants of Participation in Student Evaluation of Teaching. *The Journal of Economic Education*, 42(1), 19-30.

Kuo, W. (2007). Editorial: How reliable is teaching evaluation? The relationship of class size to teaching evaluation scores. *IEEE Transactions on Reliability*, 56(2), 178-181.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1), 159-174.

Marasini, D., Quatto, P., & Ripamonti, E. (2014). A Measure of Ordinal Concordance for the Evaluation of University Courses. *Procedia Economics and Finance*, 17, 39-46.

McKeachie, W. J. (1997). Student ratings: Their validity of use. *American Psychologist*, 52(11), 1218–1225.

Onwuegbuzie, A. J., Daniel, L. G., & Collins, K. MT (2009). A meta-validation model for assessing the score-validity of student teaching evaluations. *Quality & Quantity*, 43(2), 197-209.

Seldin, P. (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions (Vol. 10).* Jossey-Bass.

Sliusarenko, T., Ersbøll, B. K. & Clemmensen, L. K. H. (2013). *Quantitative assessment of course evaluations.* Doctoral dissertation, Technical University of Denmark Danmarks Tekniske Universitet, Department of Informatics and Mathematical Modeling Institut for Informatik og Matematisk Modellering.

Thorpe, S. W. (2002). *Online student evaluation of instruction: An investigation of non-response bias.* 42nd Annual Forum of the Association for Institutional Research.