RaMon, a Rating Monitoring System for Educational Environments

Mikel Villamañe[1[0000-0002-4450-1056]], Mikel Larrañaga[2[0000-0001-9727-1197]] and Ainhoa Álvarez[2[0000-0003-0735-5958]]

[1] Escuela de Ingeniería de Bilbao (UPV/EHU), Bilbao, Spain
mikel.v@ehu.eus
[2] Escuela Universitaria de Ingeniería de Vitoria-Gasteiz (UPV/EHU), Vitoria-Gasteiz, Spain
{mikel.larranaga, ainhoa.alvarez}@ehu.eus

Abstract. When more than one rater is involved in the assessment and scoring of a work, the scores are affected by each rater's thinking processes, knowledge level and personal preferences among other issues. These idiosyncrasies are known as rater effects and can dramatically affect the evaluation process. Even when instruments such as evaluation rubrics are used to increase the fairness and impartiality of the evaluation, rater effects may be present and affect the scoring. Rater effects can remarkably influence the final score in those assessable elements in which various raters are involved. Therefore, identifying and trying to avoid those effects is crucial for a fair evaluation. However, identifying these effects is not always an easy task and scoring leaders need tools that help them in this process. In this paper RaMon, a system for monitoring raters and controversial evaluations using visualization techniques, is presented. The authors have tested the system using data from a course with more than 100 evaluations made by 15 raters which has helped to detect some rater-effects.

Keywords: Scoring leaders, rater effects, monitoring, visualizations.

1 Introduction

All formal educational environments imply some kind of assessment or scoring of the work done. In some cases, there is only one teacher involved in the evaluation, but in other cases, e.g., Final Year Projects or Doctoral Thesis, the evaluation is performed by several raters. When the evaluation is carried out by more than one rater, monitoring both the scores and the raters is required, as there can be an important rater effect in the final mark of a work [1]. Rater effects are systematic patterns in evaluation behaviours that can be produced in an unconscious way, due to the different personal perceptions and tendencies of the raters or on purpose to affect some student's score in a positive or negative sense. To guarantee the quality of the evaluation and its fairness, the rater effects have to be detected and avoided.

With the purpose of avoiding rater effects and guarantee a fair marking that truly reflects the student performance, the standardization of the assessment criteria is the first step [2]. However, accomplishing a uniform marking standard for all the students

in those works assessed by several raters is difficult, even with settled criteria. A staff member may indicate a very good performance level for a student on a particular criterion while another staff member may grade it just as adequate [3]. The analysis of these differences may reveal the different behaviors and cognitive process of the raters during the assessment and could further facilitate taking remediation actions such as the improvement of rater selection, training, or monitoring procedures into the evaluation processes. Those actions could help reducing or minimizing the impact of rater inaccuracy or bias in scores and improving the assessment procedure [4, 5].

In many situations the data gathered during an evaluation process may include different students, with different works and each work being scored by different raters, so its analysis to detect rater effects it is no trivial. So, it is important to provide software that automates some of the rater monitoring aspects [6]; for example, by analyzing statistics related to particular raters and automatically detecting some scoring patterns.

This paper presents RaMon (**Ra**ting **Mon**itoring), a system that helps monitoring evaluations and also detecting and measuring rater effects. The system provides automatic analysis of statistics and graphical visualizations to help detecting rater effects and *controversial* evaluations.

RaMon has been tested in the assessment of Final Year Projects (FYP). In the context of FYPs, assuring impartial and unbiased evaluations is very difficult due to the existence of different evaluation boards and the high amount of raters involved [7]. This field has been chosen for two main reasons: (1) the assessment of the FYPs has been identified as one of the major concerns and problems in FYP development, and (2) the authors of this paper have been intensively working in the improvement of the development and evaluation processes of Final Year Projects. In order to overcome the problems in FYPs, the authors proposed a methodology implying a formative rubric-based assessment [8, 9]. The implementation of the new methodology and the use of rubrics have helped making the assessment less obscure and more objective, as the evaluation criteria is known both by students and lecturers and a higher coherence and agreement level in the assessment has been achieved [10]. However, some controversies in several evaluations were observed and, thus, the need for a means to supervise the evaluation process in order to assure its fairness has arisen.

This paper first presents the necessity of monitoring ratings. Next, a visual monitoring system of ratings called RaMon is presented. After, the two main monitoring aspects of RaMon are presented: Monitoring of raters or controversial evaluations. Finally some conclusions are future work are presented.

2 Rating monitoring necessity

Monitoring ratings in those contexts where multiple raters are involved is crucial to assure a fair evaluation. In the literature, different rater effects have been identified [11]:

- *Leniency/Severity effect* is the rater's tendency to give significantly lower (severity) or higher (leniency) scores than those given by other raters.

- *Central tendency effect* is the tendency to give scores only from the middle of the scale, avoiding the highest and lowest values.
- *Randomness effect* consists on giving scores inconsistently with the other raters. This effect can appear if the rater does not know the evaluation criteria or has not the sufficient knowledge to assess the work.
- *Halo/Horn effect* is the bias in which the rater gives a student always similar grades based in some preconceived impression, rather than consider the assessment criteria for the work being evaluated.
- *Differential Leniency/Severity effect* is the tendency to bias in a positive or negative way the scores of a particular group of students for some purpose.

To identify these kinds of effects, statistical analysis, including summaries of score distributions that depict the performance of each rater, are usually carried out [6]. When analyzing the rating patterns of a rater, the mean scoring and the discriminability should, at least, be examined [12]. Mean scoring refers to the mean level of scoring of each rater whereas discriminability is related to the dispersion of all scores of different ratees from a rater. These data allow evaluating the *leniency/severity effect*. If the mean scoring of a rater is very high, maybe the rater is too lenient, or if the mean is too low, maybe the rater is being very severe.

This information can be visualized in different ways. For example, Fig. 1 shows this information for those raters who have carried out at least three evaluations in the assessment context used throughout the paper (FYPs). The visualization options include traditional boxplots (Fig. 1a) or violin plots (Fig. 1b) where in addition to the grade distribution, the density of the grades for each value is also present.

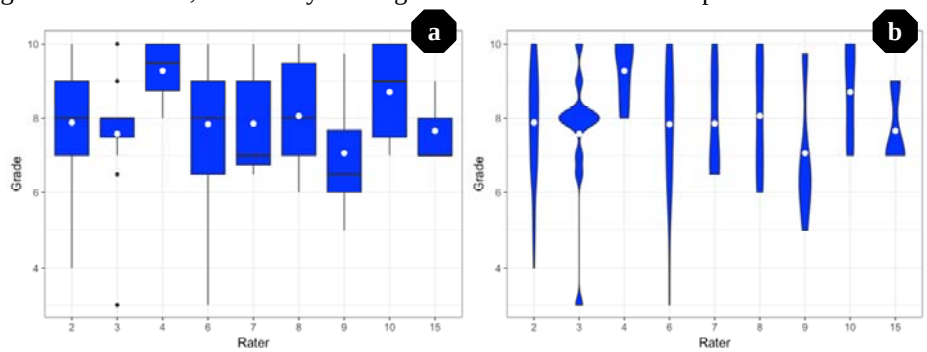

Fig. 1. Dispersion of all ratings and mean rating

In this figure, rater 4 presents a suspicious performance: a restricted score range and a high average. But is rater 4 really a lenient rater or this is mainly due to the high quality of the works assessed? To answer this question, the validity of the ratings should be also considered. This can be achieved using one of these two approaches, an accuracy framework or an agreement framework [6]. The former estimates the quality of the scores by comparing the scores of the raters with *true* scores whereas the latter compares the score of each rater with those given by the others.

The first approach is suitable, for example, in contexts in which students carry out a peer-review process and the lecturers provide a real evaluation. However, it cannot be implemented, for instance, for the Final Year Project evaluation where a *true* score is not available.

In addition, identifying the cases in which a controversial evaluation has occurred is necessary, because even if a rater effect has not been previously detected, an evaluation with significant differences among raters may indicate some kind of problem that needs to be analyzed.

Monitoring both raters and controversial evaluation allows detecting problems and taking remediation actions to improve the assessment fairness. Next section presents our proposal for RaMon, a system that relies on visualization techniques to provide monitoring of raters and controversial evaluations.

3 RaMon, a visual rating monitoring system

As stated in [14], visualization is an important part of the learning analytics area [15] which tries to improve the understanding of learning and its processes. Visualizations can help having a deeper insight into the evaluation process and help improving pedagogical interventions [16, 17]. RaMon relies on visualizations for monitoring both raters and rated assignments in order to detect rater effects and to find controversial evaluations.

In addition, RaMon allows the scoring leaders to define alarms that will raise whenever a rater or a rated assignment with an agreement or an accuracy below a settled threshold is detected.

Fig. 2 shows an example of an alarm when a controversial evaluation has been identified. Clicking on the alarm icon allows the user to visualize the information of the evaluations that raised the alarm.

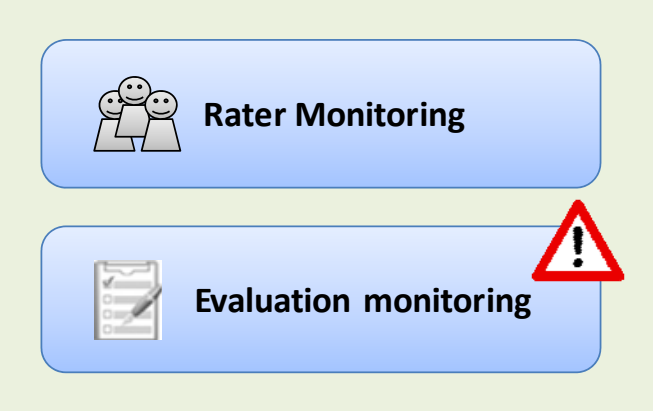

Fig. 2. Example of an alarm in RaMon when a controversial evaluation has been detected

When an alarm regarding raters' performance is highlighted, the system behaves in a similar way, allowing the user to directly access to the suspicious raters' information.

Next sections describe in detail some of the visualization capabilities that RaMon provides for monitoring both raters and controversial evaluations.

4 Monitoring of raters

As shown in **Fig. 1**, rater 4 presents a very small score range and a high mean rating. But, is really rater 4 a lenient rater? This behavior could also be due to the high quality of the works graded or the reduced amount of works assessed. In order to answer this question, more information is required, e.g, the information provided by either an accuracy framework or an agreement framework.

RaMon allows the user to analyze this information through the visualizations depicted in the next sections.

4.1 Distribution of ratings

When analyzing the dispersion of ratings, a box or violin plot such as those shown in **Fig. 1**are not enough to monitor the raters and extract accurate conclusions about the presence of rater effects.

One of the factors that can affect the scoring dispersion is the number of projects each rater has evaluated. When this number is very small, it is not rare to have a small scoring range. Therefore, RaMon can enrich the information provided with the number of projects evaluated by each rater as shown in **Fig. 3**. The users can choose to visualize the data using violin plots, as shown in the figure, or box plots according to their preferences.

In this case (see **Fig. 3**), rater number 4 has evaluated a relative small amount of projects. So, maybe, the small score range might be influenced by this factor. However, if we compare the score range for rater 4 and rater 3, the size of the range is very different whilst the amount of evaluated projects is similar.

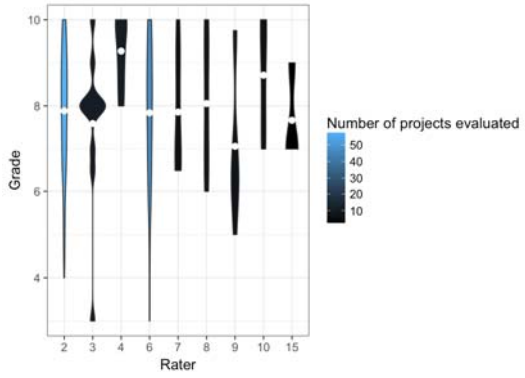

Fig. 3. Dispersion of all ratings together with the number of rated projects

Using an accuracy framework or an agreement framework can provide a higher insight of the raters' performance, including the fairness of the scores. As in FYP evaluation the *true* score is not available, an agreement framework is used and, thus, the visualization of the score distribution is enriched with the average agreement score of each rater (see **Fig. 4**).According to the information shown in **Fig. 1** or **Fig. 3**, rater 4 could be identified as suspicious of being lenient, i.e., giving always very good marks. However, analyzing data in **Fig. 4**, it can be seen that rater 4 has a high agreement score. Therefore, the small dispersion of the marks of rater 4 is probably due to the quality of the projects this rater has evaluated.

On the other hand, even if raters 7 and 8 have a higher dispersion they have a smaller agreement with the other members of the evaluation board. Although this agreement score does not allow detecting rater effects by itself, raters 7 and 8 should be analyzed in more detail using other statistics.

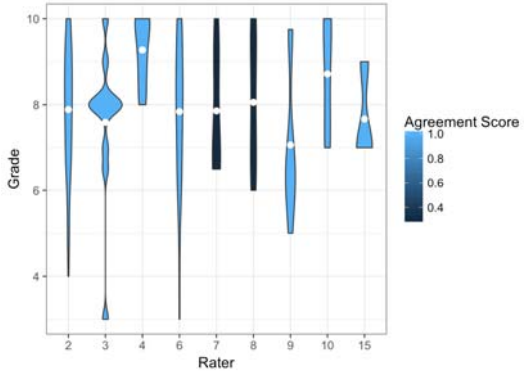

Fig. 4. Dispersion of all ratings together with the agreement score

As previously depicted, sometimes raters give scores based on personal criteria or interests (*Differential Leniency/Severity effect*). In many institutions, the supervisor of a FYP is a member of the evaluation board, and people might think they could perform differently depending on whether they are rating their pupil's work or other's. Therefore, detecting differences in the distribution of the ratings according to the role of the raters (supervisor or member of the evaluation board) might also be helpful.

RaMon provides different visualization, such as the violin plots shown in **Fig. 5**, to analyze the dispersion of their rating according to their role in the project rated: only member of the evaluation board or supervisor.

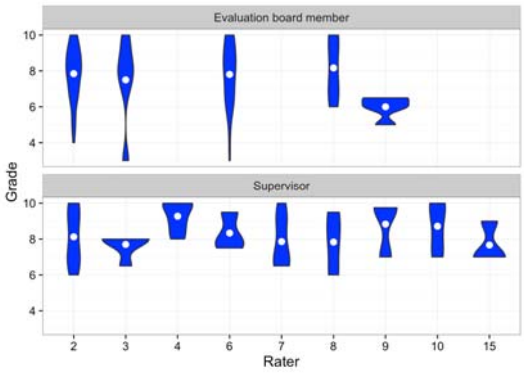

Fig. 5. Dispersion of the ratings divided depending on whether the rater was the supervisor or not

All the raters shown in **Fig. 5**, with the exception of rater 8, present very different plots according to their role. It can be inferred that those raters tend to give higher marks and with smaller dispersions when they supervise the project being evaluated (see for example rater 6). This behavior might be considered as an evidence of the *Differential Leniency effect*.

4.2 Deviations in the ratings

Analyzing the dispersion of the ratings is interesting but provides limited information for detecting whether the rater is lenient or harsh. In order to detect this aspect, the deviation of each rater from the *true* score is required. However, as mentioned above, this information is not available in FYP evaluation. In contexts where the *true* score is not available, RaMon uses the average score of the work. Moreover, when the rating is provided through the aggregation of the ratings for several elements, the system allows to compare the deviation from the projects average for the different components in order to detect in which aspects the raters are more critique.

For example, **Fig. 6** shows the deviations for the *Final report* (a) and the *Oral defense* (b). In this case, it can be derived that raters 7 and 8 are lenient whereas raters 2 and 6 seem to be more severe in the assessment of the *Final report* (**Fig. 6**a).

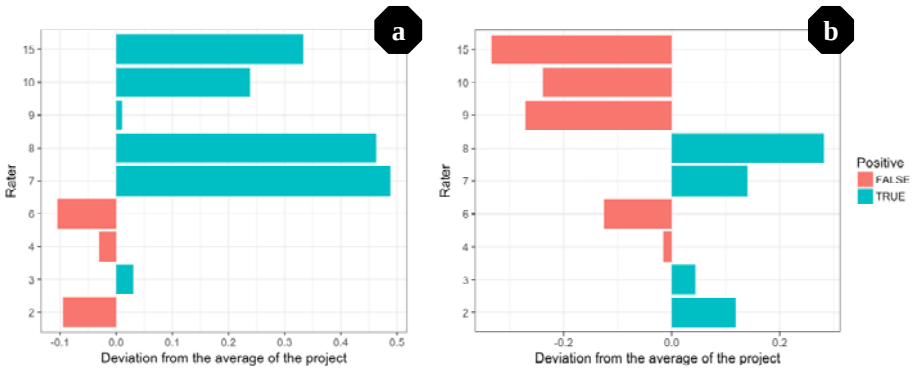

Fig. 6. Deviation from the project average. a) *Final report* and b) *Oral defense*

However, it is also interesting the analysis of both plots together to see differences according to the evaluable element being rated. For example, according to **Fig. 6**, rater 15 seems to be very lenient for the *Final report* (a), whilst being very severe for the *Oral defense* (b). This can be due to the fact that the rater gives greater relevance to the presentation and evaluates it more thoughtfully, or that the rater has not read the *Final report* very carefully and prefers not to be very severe in its evaluation.

RaMon also provides the means to analyze the behavior difference between those projects under the supervision of the rater and those in which he or she has only been member of the evaluation board to identify *Differential Leniency/Severity effects.*

Fig. 7 shows the deviation from the average of the *Final report* for different raters. In this figure, it can be detected that raters 3 and 9 are more severe when evaluating projects that have not supervised.

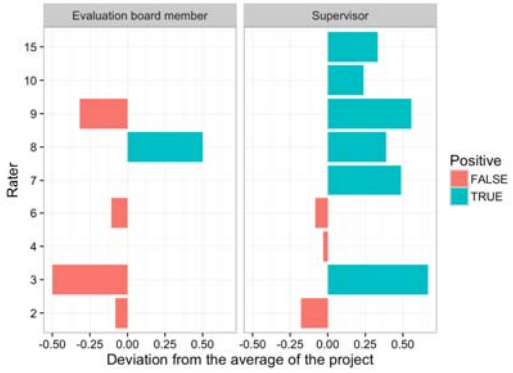

Fig. 7. Deviation from the average of the *Final report* depending on whether the rater was the supervisor or not

This kind of figure might enrich the analysis done from previous visualizations. For example, rater 4, who was suspicious of giving high marks according to the initial analysis, shows to be harsher than his or her counterparts in the evaluation board be-

cause his or her grades are a bit lower than the average even in the projects under his or her supervision.

4.3 Analysis of the rubrics

So far, all the visualizations shown have been limited to the overall score distribution, regardless the way the score has been computed. However, when the score is computed using evaluation rubrics, RaMon supplies further analysis capabilities.

These capabilities help analyzing the tendencies when performing a rubric-based evaluation. The frequency distribution of ratings, especially when graphically shown, helps detecting the raters tendency [6]. It makes evident whether raters tend to select the upper or lower categories (*Leniency/Severity effect*) or the middle ones (*Central tendency effect*).

For example, **Fig. 8** shows the frequency distribution of performance levels selected for each dimension of the *Oral defense* rubric. Analyzing this plot, it can be observed that raters 6 and 8 have a greater tendency to select higher performance levels for the projects under their supervision in certain dimensions (*Content* dimension for rater 6; *Content* and *Time* dimensions for rater 8) whilst using the whole range of performance levels for projects supevised by others.

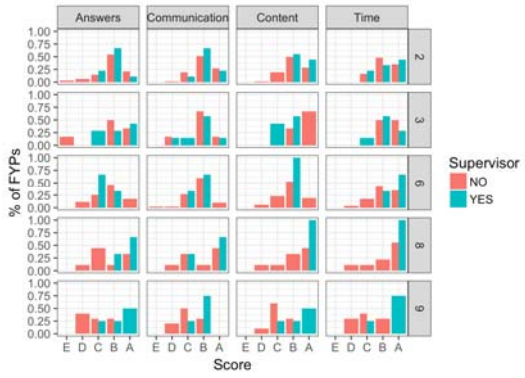

Fig. 8. Percentage of levels selected by each rater to the projects' oral presentation according to their role

In addition, RaMon can enrich this visualization by using different colors and transparency levels according to the number of evaluations made or the agreement level of the raters. This way, if the grades from a rater are very biased, but the rater has few evaluations, the rater effect can be considered less conclusive.

5 Monitoring of controversial evaluations

In order to identify controversial evaluations, RaMon can use an accuracy framework when a *true* assessment is available or an agreement framework otherwise. Plotting

the accuracy or agreement value can help identifying those evaluations in which suspicious behaviors are happening (**Fig. 9**).

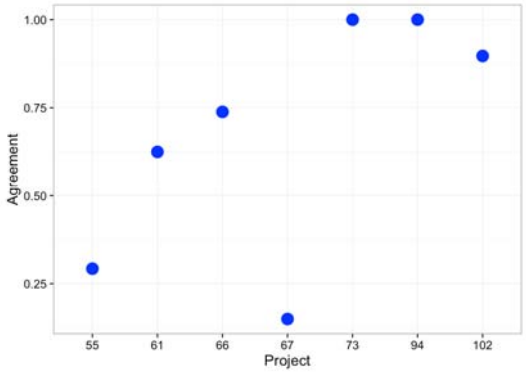

Fig. 9. Agreement of the members of the evaluation board for each project

Alternatively, RaMon can show the scores given to each assignment by each rater. **Fig. 10** shows all the assignments where rater number 7 has been part of the evaluation board. In the example used through this paper the evaluation board for each assignment was formed by 2 or 3 raters.

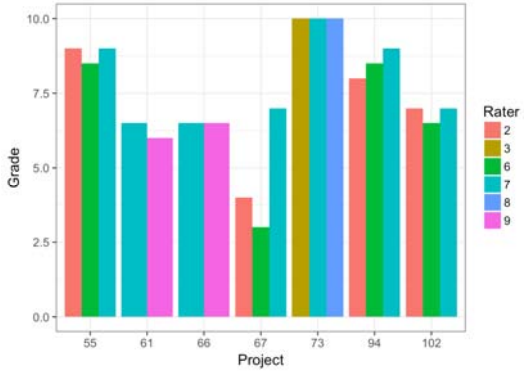

Fig. 10. Comparing the marks from the evaluation boards for the projects evaluated by rater 7

Analyzing either **Fig. 9** or **Fig. 10**, it can be observed that there is a problem in the evaluation of project 67. Its agreement score is very low and there is a rater (number 7) who has given a remarkable higher grade than the other components of the evaluation board. Therefore, this project should be analyzed in more detail.

Once this situation is detected for any project, RaMon offers different ways to analyze the details of the evaluation considering each dimension of the rubrics used. For example, in **Fig. 11** a heatmap for the *Final report* rubric for project 67 is shown.

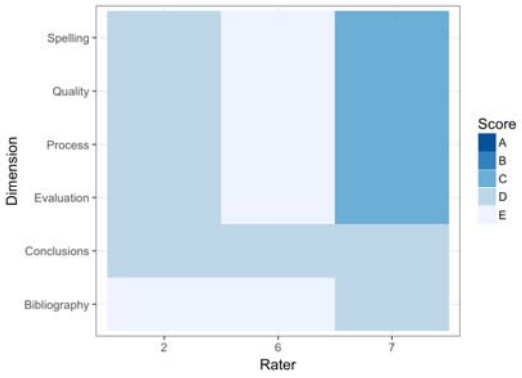

Fig. 11. Heatmap for the evaluation of the *Final report* for project 67

Although the raters agree on the performance level for the *Conclusions* dimension, and there are not great differences in the *Bibliography* dimension, there are great differences in all the other dimensions. In this case, rater 7 always gives significantly higher scores whilst rater 6 tends to give lower scores. This suggests that the evaluation of this work should be reviewed to assure a fair score.

6 Conclusions and future work

In educational environments where assessment is carried out by several raters, monitoring the evaluation results can be useful to assure the fairness of the process. In this paper RaMon, a system for Rating Monitoring in educational environments, has been presented.

RaMon supplies different visual ways of analyzing the information regarding the assessment that might help identifying different rater effects and controversial evaluations. To this end, RaMon uses diverse metrics (e.g. agreement or accuracy) to determine the quality of the ratings in addition to the mean scoring and dispersion of the grades.

RaMon has been applied in the context of the evaluation of Final Year Projects where more than 100 projects were evaluated by 15 raters. The visualizations provided have helped detecting different issues regarding both the raters and project evaluations.

For example, RaMon has helped finding differential lenient raters and identifying some controversial project evaluations (i.e. projects with low agreement among the evaluation board members).

Considering this information, remediation actions could be taken to improve the assessment process. For example, differential lenient raters can be warned to be more unbiased and controversial evaluations can be reviewed by other raters trying to achieve a fairer assessment.

In the near future, RaMon is going to be applied in more courses where multi-rater evaluations are carried out. Moreover, in some of these courses, the accuracy frame-

work is going to be used to analyze the rater effects in a student peer-review assignment where a *true* score, given by the teaching staff, is available.

In addition, the availability of more data about the academic record of each student, will allow analyzing the performance of a student along the time trying to detect the presence of *Halo/Horn effects* in the evaluations.

Acknowledgements. This work is supported by the Basque Government (IT980-16), the University of the Basque Country UPV/EHU (EHUA16/22) and SNOLA, officially recognized Thematic Network of Excellence (TIN2015-71669-REDT) by the Spanish Ministry of Economy and Competitiveness.

References

1. Engelhard Jr George, J., Wang, Jue: Unfolding Rater Accuracy in Performance Assessments. Rasch Meas. Trans. 28, 1489–1491 (2015).
2. Chan, K.L.: Statistical analysis of final year project marks in the computer engineering undergraduate program. IEEE Trans. Educ. 44, 258–261 (2001).
3. Teo, C.Y., Ho, D.J.: A systematic approach to the implementation of final year project in an electrical engineering undergraduate course. IEEE Trans. Educ. 41, 25–30 (1998).
4. Long, H., Pang, W.: Rater effects in creativity assessment: A mixed methods investigation. Think. Ski. Creat. 15, 13–25 (2015).
5. Wolfe, E.W.: Identifying rater effects using latent trait models. Psychol. Sci. 46, 35–51 (2004).
6. Wolfe, E.W.: Methods for monitoring rating quality: Current practices and suggested changes. Iowa City IA Pearson. (2014).
7. Valderrama, E., Rullan, M., Sánchez, F., Pons, J., Mans, C., Giné, F., Jiménez, L., Peig, E.: Guidelines for the final year project assessment in engineering. In: Actas de IEEE Frontiers in Education Conference. pp. 1–5. IEEE Computer Society, San Antonio, Texas, EE.UU. (2009).
8. Villamañe, M.: Análisis y mejora de los marcos actuales de desarrollo y evaluación de los Trabajos Fin de Grado mediante el uso de las TIC, (2017).
9. Villamañe, M., Ferrero, B., Álvarez, A., Larrañaga, M., Arruarte, A., Elorriaga, J.A.: Dealing with common problems in engineering degrees' Final Year Projects. In: Actas de IEEE Frontiers in Education Conference. pp. 2663–2670. IEEE Computer Society, Madrid (2014).
10. Villamañe, M., Álvarez, A., Larrañaga, M., Ferrero, B.: Desarrollo y validación de un conjunto de rúbricas para la evaluación de Trabajos Fin de Grado. ReVisión. 10, 17–27 (2017).
11. Myford, C.M., Wolfe, E.W.: Detecting and measuring rater effects using many-facet Rasch measurement: part I. J. Appl. Meas. 4, 386–422 (2003).
12. Wong, K.F.E., Kwong, J.Y.Y.: Effects of rater goals on rating patterns: Evidence from an experimental field study. J. Appl. Psychol. 92, 577–585 (2007).
13. Howard E. A. Tinsley, Weiss, D.J.: Interrater Reliability and Agreement. In: Howard E. A. TinsleySteven and D. Brown (eds.) Handbook of Applied Multivariate Statistics and Mathematical Modeling. pp. 95–124. Academic Press, San Diego (2000).

14. Kay, J., Bull, S.: New Opportunities with Open Learner Models and Visual Learning Analytics. In: Conati, C., Heffernan, N., Mitrovic, A., and Verdejo, M.F. (eds.) Actas de Artificial Intelligence in Education. pp. 666–669. Springer International Publishing, Cham (2015).
15. Siemens, G.: Learning analytics: envisioning a research discipline and a domain of practice. In: Actas de International Conference on Learning Analytics and Knowledge. pp. 4–8. ACM (2012).
16. Pardo, A., Dawson, S.: Learning Analytics: How can Data be used to Improve Learning Practice. In: P. Reimann, S. Bull, M. Kickmeier-Rust, R. K. Vatrapu & B. Wasson (Eds.), Measuring and visualizing learning in the information-rich classroom,. pp. 41–55. Routledge (2016).
17. Tervakari, A.M., Silius, K., Koro, J., Paukkeri, J., Pirttilä, O.: Usefulness of information visualizations based on educational data. In: Actas de IEEE Global Engineering Education Conference. pp. 142–151. IEEE Computer Society (2014).