Teachers as Designers of Formative e-Rubrics: A Case Study on the Introduction and Validation of Go/No Go Criteria

Pedro Company[1], Jeffrey Otey[2], María-Jesús Agost[3],
Manuel Contero[4], and Jorge D. Camba[5]

[1]Institute of New Imaging Technologies, Universitat Jaume I
Av. de Vicent Sos Baynat, s/n 12071 Castellón, Spain
pcompany@uji.es

[2]Zachry Department of Civil Engineering, Texas A&M University
213 DLEB, College Station, TX 77843-3136, USA
j-otey@tamu.edu

[3]Department of Mechanical Engineering and Construction, Universitat Jaume I
Av. de Vicent Sos Baynat, s/n 12071 Castellón, Spain
magost@uji.es

[4]Instituto de Investigación e Innovación en Bioingeniería (I3B)
Universitat Politècnica de Valéncia, Camino de Vera s/n, 46022 Valencia, Spain
mcontero@upv.es

[5]Computer Graphics Technology Department, Purdue University
Knoy Hall, 401 N. Grant St., West Lafayette, IN 47907, USA
jdorribo@purdue.edu

Abstract. Information and Communications Technologies (ICTs) offer new roles to teachers to improve learning processes. In this regard, learning rubrics are commonplace. However, the design of these rubrics has focused mainly on scoring (summative rubrics), whereas formative rubrics have received significantly less attention. ICTs make possible electronic rubrics (e-rubrics) that enable dynamic and interactive functionalities that facilitate the adaptable and adaptive delivery of content. In this paper, we present a case study that examines three characteristics to make formative rubrics more adaptable and adaptive: criteria dichotomization, weighted evaluation criteria, and go/no-go criteria. A new approach to the design of formative rubrics is introduced, taking advantage of ICTs, where dichotomization and weighted criteria are combined with the use of go/no-go criteria. The approach is discussed as a method to better guide the learner while adjusting to the student's assimilation pace. Two types of go/no-go criteria (hard and soft) are studied and experimentally validated in a Computer-Aided Design assessment context. Bland-Altman plots are constructed as discussed to further illuminate this topic.

Keywords. Formative rubrics, adaptable and adaptive rubrics, e-rubrics, go/no-go criteria.

1 Introduction

Information and Communications Technologies (ICTs) offer new roles to teachers to improve learning processes. ICTs present the opportunity to empower well known tools, such as rubrics, that are supported in a very simple way by current learning management systems (LMSs). Learning rubrics are scoring guides constructed of descriptors (or evaluative criteria) that establish assessment specifications. These criteria should align with the formative objectives [1] and are frequently referred to as "evaluation tables," since they are typically arranged in tabular format. The progressive development of electronic rubrics as a regular tool in learning management systems gives teachers a new intervention capability to facilitate the learning process. Adaptable and adaptive rubrics are made possible with these systems. The design of these types of rubrics will be an important task for the future development of personalized learning.

One of the main objectives of a rubric is to standardize and accelerate the evaluation process by highlighting the most relevant aspects of the subject matter. However, many authors claim that rubrics should go beyond assessment, as their continuous development has made them useful for informing and motivating the assessed subjects.

Rubrics can be classified according to different criteria [2]. Holistic rubrics offer a global view of the evaluation, whereas analytic rubrics provide a detailed view of different items. Similarly, general rubrics can be used for an entire course, while task-specific rubrics may focus only on one particular assignment or project. Summative rubrics produce a final global score, sorting subjects by those who pass and those who fail the evaluation. Formative rubrics provide performance feedback by conveying information about the strengths and weaknesses of the subjects [3]. The research reported in this paper focuses on analytic task-specific formative rubrics.

Formative rubrics help students determine their own progress throughout the training period. Since different formative levels have different needs, these rubrics must be designed, formatted, and applied based on their specific purpose and context of use. Of particular interest is self-evolving rubrics that can adapt automatically to the learning pace of each student.

This paper examines the design and use of formative rubrics in higher education, where rubrics for specialized content are common [4]. The use of computer technologies and formats is discussed as a means to maximize the benefits of adaptable and adaptive rubrics. To this end, certain characteristics such as criteria dichotomization, weighted evaluation criteria to provide various levels of importance, and go/no-go criteria have been linked to new strategies that adjust to the learning pace of each student. More specifically, the use of the go/no-go criteria in combination with dichotomization and weighted criteria as a strategy to better guide the learner is described.

The paper is structured as follows: in the next section, a brief review of the state of the art in rubrics is provided as a set of commonly accepted lessons learned. Next, an examination follows discussing the concepts of dichotomization and levels of importance in evaluation criteria, with emphasis on strategies targeting how they can be used to control the student's assimilation pace. Two types of go/no-go criteria are studied and validated in a mechanical computer-aided design context. General recommendations are discussed based on the results of two experimental studies. Obviously, inter-rater agreement is essential for formative rubrics, particularly if they are aimed at distance and self-paced training courses. Thus, finally the concept of "agreement" is examined with further statistical analyses to support the research hypothesis.

2 State of the Art

The state of the art on rubrics development can be summarized as a set of commonly accepted lessons learned. Rubrics can manage complex evaluation scenarios by defining subsets of homogeneous criteria called dimensions. In this regard, clustering techniques can be used to link descriptors in relatively homogeneous natural groupings. It has been stated that "dimensions are useful to work with hierarchical rubrics, which organize criteria in different levels" [5]. Dimensions are also useful to work with complex assessments, which evaluate heterogeneous criteria [6].

Each criterion is evaluated by measuring the degree of compliance or achievement level of a particular situation. The achievement levels should be introduced using the same terminology as the corresponding criteria. It is important to use a consistent scale throughout all achievement levels while avoiding the mixing positive and negative scales [7]. The number of achievement levels may vary depending on the situation. In some cases, the compliance level may be dichotomously determined. Whether or not a criterion is met is generally measured through a finite set of levels that discretize a continuum. A commonly used system for establishing discrete levels is based on Likert elements, usually with five achievement levels or points [8]. Rohrmann [9] states that category scaling enhances the usability of assessment instruments and well-defined qualifiers facilitate unbiased judgments.

To be noted, single point rubrics, introduced in 2000 by Dietz, as stated by Fluckiger [10], reduce the achievement categories to a single measure (the standard achievement or measurement of proficient performance), and provide feedback or evidence provided by the student demonstrating under-performance, accomplishment, or over-performance (work exceeding expectations). The goal is to provide students with a mechanism to assess their own learning in any subject area that requires an original response. Success has been reported utilizing them in an introductory programming course [11].

While absolute objective scoring is difficult to achieve - especially in self-assessment - achievement level categories provide unambiguous scales to properly rate quality. When providing performance scores to students, a preferred strategy involves moderate leniency, so confidence can be gradually built. Instead of penalizing students for each individual mistake, a proper assessment perspective may involve viewing those instances in a larger context to better determine whether they are significant enough to prevent awarding the highest rating.

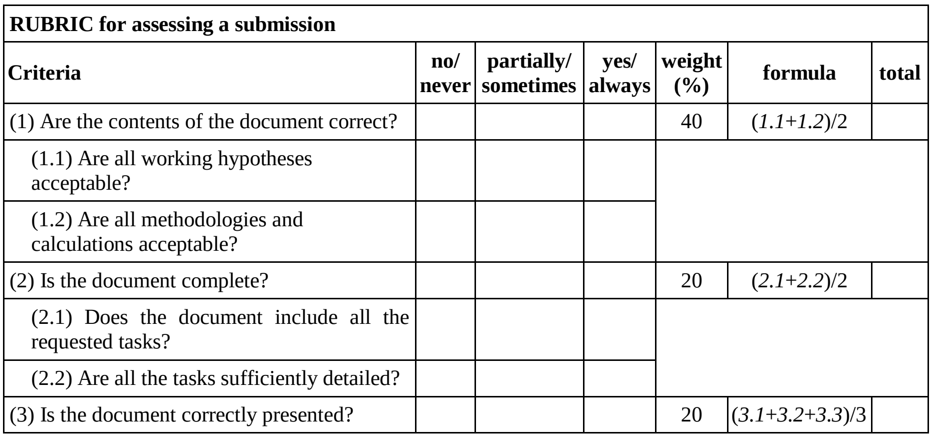
Rubrics are designed to homogenize evaluation processes. A common strategy toward this end is to complement rubrics with anchors; i.e. written descriptions or examples that illustrate the various achievement levels, or work samples [12]. Descriptions of good practices should be integrated into formative rubrics and used on demand, so the rubric user can be guided throughout the process of determining which criterion to check, how to do it, and the importance of the criterion in the overall result. Good practices are particularly important in distance and self-paced training courses, where the instructor is not always immediately available.

An important consideration when designing a rubric is the type of format. Static formats can be easily implemented but cannot provide feedback to the user, so they are only suitable for summative assessment. Formative assessment requires rubrics that can adapt. In this sense, formative rubrics must be dynamic [6].

Dynamic formats process information to provide feedback to the user, and can be adapted to specific cases and users with different levels of expertise. Two types of dynamic rubrics can be distinguished. Rubrics are adaptive if the instructor can design different rubrics for different stages of formation [13]. They are adaptable if users can vary the level of detail interactively to adapt the rubric to their learning rhythms [6]. Clearly, implementing these functionalities require the use of electronic rubrics (e-rubrics). To this end, dedicated applications to manage rubrics are needed, since standard tools such as spreadsheets are not fully adequate [14]. In the next sections, strategies to design adaptable and adaptive rubrics are discussed.

2.1 Adaptable Rubrics

Rubric criteria must be arranged according to gradually increasing levels of detail, so every user has the opportunity to select the level that better adapts to their understanding, thus optimizing the formative action. According to Company et al. [6], an effective strategy to accomplish this functionality with e-rubric tools allows the user to “fold” and “unfold” the detail level of the criteria interactively. A basic rubric with hierarchical criteria levels is illustrated in Figure 1. All rubric criteria are displayed unfolded, as interactivity cannot be shown in a static image.



RUBRIC for assessing a submission

Criteria	**no/ never**	**partially/ sometimes**	**yes/ always**	**weight (%)**	**formula**	**total**
(1) Are the contents of the document correct?				40	(*1.1*+*1.2*)/2	
(1.1) Are all working hypotheses acceptable?						
(1.2) Are all methodologies and calculations acceptable?						
(2) Is the document complete?				20	(*2.1*+*2.2*)/2	
(2.1) Does the document include all the requested tasks?						
(2.2) Are all the tasks sufficiently detailed?						
(3) Is the document correctly presented?				20	(*3.1*+*3.2*+*3.3*)/3	

(3.1) Is it well organized?						
(3.2) Is it well designed?						
(3.3) Is it free of grammatical errors?						
(4) Is the document clear?				20	(*4.1*+*4.2*)/2	
(4.1) Can the document be read easily?						
(4.2) Are all the concepts used in the document well defined?						
Final Score (*10.4+*2**0.2+*3**0.2+*4**0.2) *10**						

Fig. 1. Example of a rubric with criteria showing two levels of detail (unfold view). Italicized numbers in column "weight" represent criteria.

Dichotomous criteria can be defined when assessing simple aspects of a task, but also when itemizing the assessment with large amounts of criteria. Therefore, level decomposition indirectly favors the dichotomization of the assessment process.

Similarly, the role of weights as focal pointers is essential. Disagreement between student and instructor perceptions will naturally reveal discrepancies on the accuracy of the evaluated task. Therefore, making the perceived importance of each criterion explicit helps students focus on "what counts" (those criteria that the instructor prioritizes). Additionally, adjusting weights throughout the training period shifts focus from criteria that are already achieved to those that require a longer maturing time.

Finally, all levels must be described using the same terms as the main criterion, but each level should be characterized by appropriate qualifiers for each type of attribute. The attributes are the characteristics underlying the criteria, such as frequency, intensity or probability. Other authors such as Rohrmann [9] established achievement levels based on different attributes.

2.2 Adaptive Rubrics

To benefit from the formative nature of rubrics and accommodate the student's learning pace, general concepts are typically revealed gradually, as successive rubrics throughout a course. Alternatively, an unfold/fold strategy can also be adopted where only low level (deployed) criteria are shown at the beginning of a course, and other versions (displaying higher level criteria) are introduced only to students who have already mastered the low level criteria. Consequently, the unfold/fold strategy is useful both for the student, who can adapt the criteria to her optimal level of understanding, and the instructor, who can make criteria less specific throughout the learning process as needed.

For a more accurate adjustment to a user's level, an e-rubric tool must allow the configuration of gates, i.e. the different options that are available to the user based on their previous responses to particular criteria, so the rubric can automatically update with new and more complete content each time the user reaches a certain level of achievement in a previous (more basic) rubric. For example, a rubric can be configured so that only when a user obtains a minimum score of 8 points out of 10 in one lesson, will the contents of the next lesson be available. Another possibility is to have the system display messages that reinforce learning based on the results achieved. If the score does not reach a minimum threshold in certain questions, an automatic message can inform the student to review the particular lesson(s) related to those contents.

Gates provide active and customized feedback depending on user progress. The instructor has adaptive control over the gates from one stage to another, whereas users have adaptable control of the information needed to complete each stage. Therefore, adaptive rubrics must be coordinated with the lesson plans.

A sample course schedule, with the introduction plan of lower level criteria during a period of six weeks (following a bottom-up approach), is shown in Figure 2. These low-level criteria will later be replaced by more general criteria. Note that knowledge and/or procedures may become exclusionary in the last weeks (indicated with an "X"). In the next section, the use of go/no-go criteria is discussed as a means to implement this behavior.

	Weeks					
Criteria	**1**	**2**	**3**	**4**	**5**	**6**
M1			▨	▨	X	X
M1.1		▨	▨			
M1.2	▨	▨	▨			
M2				▨	▨	X
M2.1		▨	▨			
M2.2			▨			
M3			▨	▨	X	X
M3.1			▨			
M3.2		▨	▨			
M3.3	▨	▨	▨			
M4				▨	▨	X
M4.1		▨	▨			
M4.2			▨			

Fig. 2. Example of schedule as a basis for an adaptive rubric system. "X" indicates exclusion criteria.

3 Go/No-Go Criteria

Go/no-go criteria are defined as exclusion conditions inside descriptors. They establish basic conditions that must be met. Otherwise, the evaluation process is interrupted and the final score will be zero, regardless of other criteria. These go/no-go criteria must be explicitly identified. For example: "If the deliverable document contains many spelling mistakes, evaluation does not continue". Figure 3 demonstrates a hard go/no-go criterion, placed at the beginning of the rubric to avoid unnecessary assessments.

Rubric to assess the cover of a submission		
Instructions: (1) Mark the proper achievement level for each criterion (2) Assign a final score of zero (“0”) if a go/no-go criteria (*) is not met (3) Calculate the final score by using the formula shown in the last row		
Criteria	**No**	**Yes**
Does the deliverable have a cover? Do not continue assessing if there is no cover (*)		
Does the cover have a title?		
Is the author information available in the cover page?		
Is the date of the document available in the cover page?		
(E) Add the number of typos here		
Final score (10-*E*)* (Total Yes/4) *10		

Fig. 3. Example of rubric with a go/no-go criterion and a threshold parameter.

An alternative form of go/no-go is establishing a threshold parameter for pass/fail. An example of such a soft go/no-go criterion embedded in the final score (considering that after 10 errors, the grade becomes zero regardless of other criteria) is shown in Figure 3. As a result, catastrophic failures result in a no-go, while moderate failures reduce the final score, but do not prevent assessing other rubric dimensions. This soft go/no go is a recommended academic scoring alternative necessary to highlight critical failures, while avoiding unnecessary punitive student exam scores (so that maximum partial credit could be awarded). On the contrary, hard go/no-go criteria clearly send the message that some failures are unacceptable for already formed students. Soft go/no go criteria may be unsuited for non-academic environments.

3.1 Experimental Evaluation of Go/No-Go Criteria

The goal of this study was to validate the hypothesis that neither the use of hard nor soft go/no-go criteria affects the ability of students to self-assess their work. In other words, that the strong warning message sent by this type of criteria does not prevent the perception of the other criteria.

To this end, two pilot experiments were conducted (mid-term and final exam) to assess student understanding of CAD assemblies. These examples represent complex evaluation cases, where the following quality dimensions of CAD models/assemblies [5] were used:

1. Models are valid if they can be accessed successfully by suitable software with no errors or warnings;
2. Models are complete if all necessary product characteristics are provided for all design purposes;
3. Models are consistent if they do not crash during normal design exploration or common editing tasks;

4. Models are concise if they do not contain any extraneous (repetitive or fragmented) information or techniques;
5. Models are clear and coherent if they are understood at first glance;
6. Models are effective if they convey design intent.

Undergraduate students (beginning CAD users) at a Spanish university were introduced to a prototype system of assembly rubrics after being exposed to parts rubrics earlier in the semester. Detailed explanations of the assembly rubric dimensions were discussed and provided to the students prior to their exams. This material included thorough descriptions of the definition and significance of the quality dimensions, as well as clarifications of the detailed criteria used to measure the degree of accomplishment of such dimensions. The five achievement levels were defined as: No/Never, Almost Never/Rarely, Sometimes, Almost Always/Mostly, Yes/Always [9] and quantified as 0, 0.25, 0.5, 0.75 and 1, respectively.

Completion of rubrics was required and considered correct if it matched the primary instructor's (Instructor 1) evaluation (ideal). The primary instructor (Instructor 1) was the professor of record of the course; Instructor 2 was a faculty member at a different institution whose sole responsibility was to assess the student work.

The instructors had no prior collaborative grading experience. Correspondence was conducted electronically due to their geographic separation. Hence, any possible agreement between instructors cannot be assumed to correspond to any preceding mutual cooperation. Thus, it can be assumed that any agreement can be explained by the use of well-designed rubrics utilized by knowledgeable instructors with detailed understanding of the process of teaching CAD.

While the system was primarily developed to assess CAD model quality, the rubric itself can be assessed in terms of ease of understanding and use (which is an underlying research hypothesis). If a rubric is clearly understood, each rater (instructor and student) should produce similar assessments.

Experiment 1

For this experiment, students were required to assemble a fitness equipment pulley (Fig. 4 right) using four non-standard parts (previously modeled, as displayed in Fig. 4, left) and various standard parts.

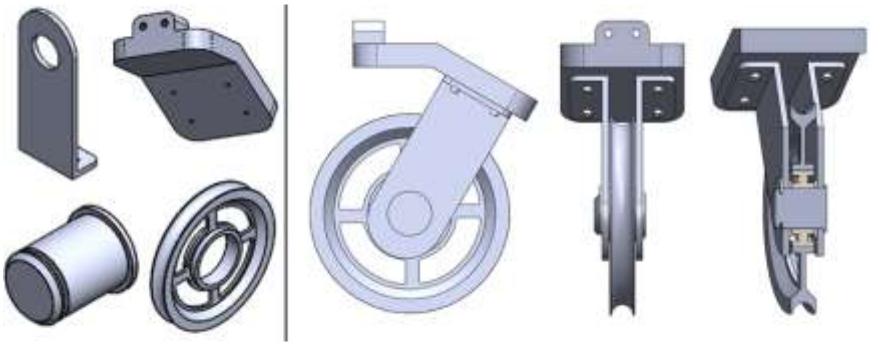


Fig. 4. Assembly and non-standard parts (left) used in Experiment 1.

The original intent, as explained to students, was to identify validity as a hard go/no-go switch, so the assembly would fail assessment if all linked files could not be located or used. Thus, weights were assigned as follows: valid (hard go/no-go criterion that multiplies the overall score using the remaining rubric dimensions), 0%; complete, 20%; consistent, 30%; concise, 20%; clear, 15%; and design intent, 15%.

Students were provided the solution after submitting their exams in order to self-assess their performance against an ideal solution. Although students were informed that Dimension 1 (validity) would be a hard go/no-go criterion (i.e., failure to submit a valid file would result in a non-passing grade for the exam), a soft go/no-go criterion was enforced by instructors (with up to half-credit being awarded to avoid unnecessarily punitive scoring).

A total of fifty students took the mid-term exam, though only forty-six submitted the self-assessment rubrics. Table 1 summarizes the results of the experiment which can be found at Otey [15] and illustrates the inter-rater reliability scores for the mid-term exam (for the student and both instructors). As described by Otey [15], the experiment demonstrated stronger agreement between instructors than either instructor with the students, for all dimensions. Agreement between instructors and students was obtained for the dimensions of validity, completeness, and consistency, however weak agreement exists for conciseness, clarity and design intent.

Table 1. Inter-rater reliability scores for midterm exam

% agreement	**Dim. 1 (Valid)**	**Dim. 2 (Complete)**	**Dim. 3 (Consistent)**	**Dim. 4 (Concise)**	**Dim. 5 (Clear)**	**Dim. 6 (Design Intent)**
Indiv. – Instr. 1	50.0	36.5	27.0	17.0	23.0	7.6
Indiv. – Instr. 2	51.9	38.0	28.8	21.0	21.0	11.5
Instr. 1. – Instr. 2	94.0	75.0	73.0	61.5	63.0	28.8

Experiment 2

Following a similar procedure as the first experiment, the final exam required assembling a mechanism. More specifically, students were asked to assemble a mechanical filter using four non-standard parts (previously modeled) and assorted standard parts. The assembly is shown in Figure 5.

Fifty-one students submitted self-assessment e-rubrics using our system. The students were assessed on assembly sequence and the use of sub-assemblies. This time, however, the students were informed that Dimension 1 (validity) would be assessed as a soft go/no-go criterion. As an example, a validity score of 0.5 would result in the remaining criteria receiving half value.

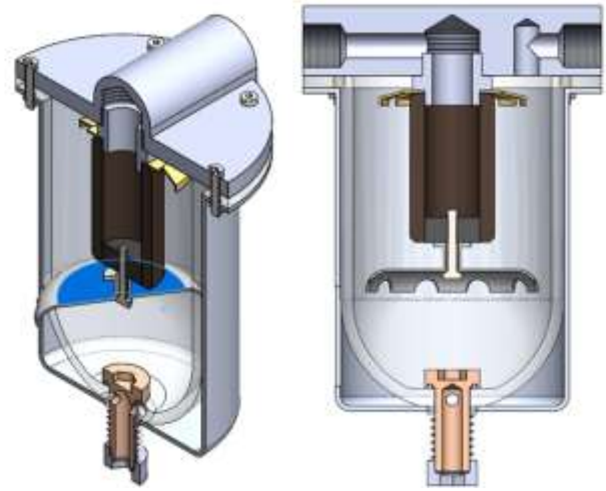


Fig. 5. Two section views of the filter assembly used in Experiment 2.

The inter-rater reliability scores for the final exam (for the student and both instructors) are shown in Table 2. Once again, there is greater agreement between the instructors than between instructor and students. There exists moderate to strong agreement for Dimension 1, between both instructors and between instructors and students. There is strong agreement between instructors for Dimensions 1, 2, 4, and 5 and little agreement between instructors and students for any dimension other than validity. It appears that there is no measurable increase in agreement for all dimensions other than validity, for instructors and students, between the mid-term and final exam. Reasons for the lack of increase could be attributed to time, as there was only three weeks between exams, which could have resulted in not enough time for many students to grasp missed concepts and improve their performance.

Table 2. Inter-rater reliability scores for final exam

% agreement	Dim. 1 (Valid)	Dim. 2 (Complete)	Dim. 3 (Consistent)	Dim. 4 (Concise)	Dim. 5 (Clear)	Dim. 6 (Design Intent)
Indiv. – Instr. 1	69.0	34.6	23.0	15.0	23.0	15.0
Indiv. – Instr. 2	69.0	30.0	15.0	15.0	23.0	15.0
Instr. 1. – Instr. 2	100.0	75.0	44.0	88.0	84.6	25.0

A slight increase can be observed when comparing Tables 1 and 2 in the agreement of some dimensions over time, which we speculate could be due to the previous exposure to the rubric. However, there are no significant changes in the inter-rater reliability. Thus, we can validate the hypothesis that switching from hard to soft go/no-go criteria does not affect the ability of students to self-assess their work, since the similarities in the assessments of raters (instructors and student) imply that the rubric is clearly understood.

Ideally, it would be useful to determine whether the correlation for each dimension significantly improved or decreased. However, since the r-value is bound between 0 and 1, it is exceedingly difficult to construct meaningful conclusions about this matter. A linear relationship between the correlation values cannot be assumed, but even if the change in correlation values were significant, would it be consequential? Even with perfectly defined rubric dimensions, it is impossible to remove all subjectivity, which may cloud any definitive judgment. In such cases, only the professional expertise of the investigator would guide those determinations. Regardless of this lack of statistical certainty, a pronounced general pattern emerges that reflects a positive directional improvement for a majority of rubric dimensions (between both instructor and student, and between instructors).

Finally, it is worth investigating in the future, whether the clearly increased agreement for Dimension 1 is only due to previous exposure, or to the fact that the explicit declaration of the go/no-go criterion as “soft” frees the students from a possible panic to recognize the failure, since it is no longer catastrophic.

4 Agreement

The authors understand agreement as the measurement of how much unanimity or consensus there is in the ratings given by different evaluators (both instructors and students). As explained above, previous joint experience between the instructors that rated the experiments did not exist. Besides, it is obvious that students possessed no previous background on rubrics with go/no go artifacts. Still, further investigation is warranted as to whether the agreement perceived in the previous experiments are consistent. Selecting the suitable statistical analysis requires careful attention since each test approaches measurement differently. In particular, inter-rater reliability and regression analyses, among others, are attempts to define how much agreement exists between raters. Next, a brief review the most related state-of-the-art is provided.

It has been stated that while there has been tremendous improvements in statistical analyses, approaches to improve the measurement of agreement have yet to be accomplished [16]. Kottner and Streiner [17] state that confusion between reliability and agreement estimation are “caused by conceptual ambiguities.” They believe that agreement is based on whether judgments are alike or the amount they fluctuate, while reliability is based on the ratio of variability between assessments. There also are misconceptions between repeatability and reproducibility. Watson and Petrie [16] define repeatability as agreement between different measurements on the same sample, while reproducibility uses the same techniques on an identical sample.

Much research literature on agreement and repeatability has its origins in the medical field, with new treatments being measured against traditional ones. McLaughlin reports that complications emerge when determining if a novel method is comparable to a conventional one [18]. In such cases, professional experience can bias true measurement of agreement. When developing new treatments or medications, repeatability is important to assist in outcome predictability, however, since no two patients are alike, there are realistic limits to definitively measuring agreement. Costa-Santos and colleagues [19] state that agreement is a challenging idea to capture statistically. They researched using Limits of Agreement (LA) and Intraclass Correlation Correlation Coefficient (ICC) to assess disagreement on medical diagnoses and found that inconsistent results were produced. In fact, McLaughlin [18] concludes that testing for significance is more rigid than for agreement. Chen and Barnhart [20] examined the use of Intraclass Correlation Coefficient and the Concordance Correlation Coefficient to measure agreement using 2D-echocardiograms. They propose using the Concordance Correlation Coefficient as it does not require Analysis of Variance (with its limiting assumptions).

The authors believe that the Bland and Altman approach is the most suitable one, as it states that agreement relies on repeatability, because if one method compared against another has poor repeatability, the increased variation in the measurement process will affect the agreement with the other method [21]. Van Stralen and colleagues [22] suggest that while the correlation coefficient is a simple calculation and is an accepted method to measure agreement, a major drawback is that it does not distinguish systematic and random errors between measurements. As such, they suggest using Bland-Altman plots, which are superior in instances of repeated measurements.

5 Bland-Altman

In order to shed further light on the applicability of the formative rubrics developed by the authors, Bland-Altman analysis was performed on the previously discussed experiments [15]. Bland-Altman has been historically used to access the effectiveness of different medical treatments. In this case, using it in an educational setting is a novel ambit and returns interesting results.

Bland-Altman describes difficulties in determining “true” values by estimating with indirect methods. According to Beckstead [23], Bland and Altman argue that Pearson’s coefficient is not suitable for determining if one value can be substituted for another. This agreement is dependent on:

- Elevation (difference between means of two measurements);
- Spread (standard deviation of differences);
- Scatter (as covariance increases, scatter decreases).

The Correlation coefficient *r* is not an appropriate index because it is a standardized covariance and is sensitive to scatter at the expense of its ability to respond to elevation and spread. Elevation and spread are systematic sources of bias [23].

Bland and Altman [24] state that in order to truly examine methods (or treatments), disagreement must be quantified, as agreement is not dichotomous. In other words, the amount of agreement must be numerically quantified, as statistical qualifiers such as correlation coefficients, regression, and means comparison are ill suited for such purposes.

Bland and Altman [21] determined that the use of correlation coefficients and regression analysis are ill-suited to distinguish whether new measurement techniques are sufficient to replace previously used ones. Their reservations about using correlation coefficients include the fact that *r* measures the relationship strength between two variables rather than the agreement that exists between them. They developed a process based on graphical techniques and straightforward computations which assist repeatability assessment, of which they were most interested.

Giavarina states that Bland-Altman plot analysis is an uncomplicated method to calculate the bias between mean differences and to gauge the agreement interval but does not determine acceptable limits [25]. He believes that correlation between methods is ambiguous and is ill-suited to determine method congruity.

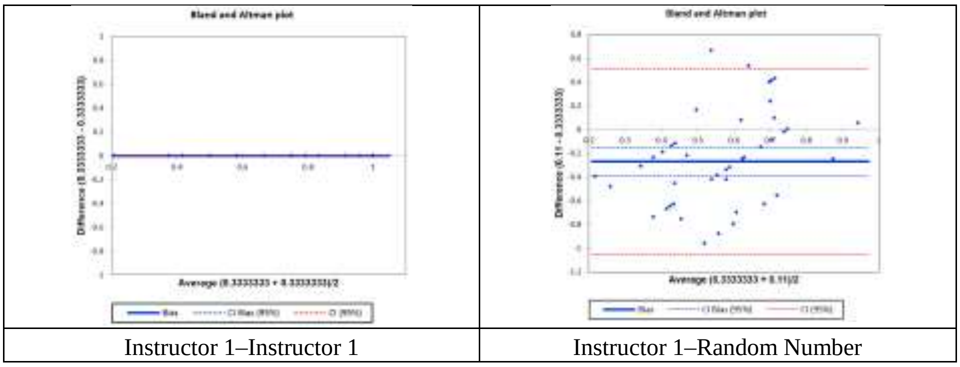
The Bland-Altman technique is used to display the results of this comparison. A plot was created illustrating the differences against the mean and applying a log transformation. If the mean difference is not close to zero, the approaches are systematically generating disparate results [26]. Bands of agreement for the differences, at the 95% level, are created and illustrate bias in both directions [26].

Bland-Altman plots are understood by resolving the following criteria:

- Does large bias (average discrepancy) exist between the methods? The answer is clinical in nature, not statistical;
- Large bands of agreement provide for inconclusive results;
- Does the difference between methods increase as the average increases?
- Is variability consistent throughout the plot [26]?

In order to provide a simple illustration of how to explain the interpretation of Bland-Altman results, example plots were created for the dimension of Design Intent for the midterm exam. Figure 6 shows the outcomes of Instructor 1 versus himself and for Instructor 1 versus random numbers (existing in the range of possible scores). As can be seen, in the plot of Instructor 1 versus himself, the data reflect a linear relationship with an extraordinarily small band of agreement, which is to be expected. In the plot of Instructor 1 versus random numbers, the bands are considerably wider and there is greater spread in the points, reflecting many outliers. Additionally, it should be noted that no pattern exists in the graphical representation of the data.

The differences between values should be near zero for close relationships to be assumed. The difference between Instructor 1 and himself is 0.000, while the difference between Instructor 1 and random numbers is -0.269. In examining these sample plots, it can be easily seen that there is more correlation in the first example than in the second, as not only evidenced by the differences in values, but graphically exhibited in the Bland-Altman plots.



Instructor 1–Instructor 1

Instructor 1–Random Number

Fig. 6. Example Bland-Altman plots for Instructor 1 vs. himself and Instructor 1 vs. Random Number.

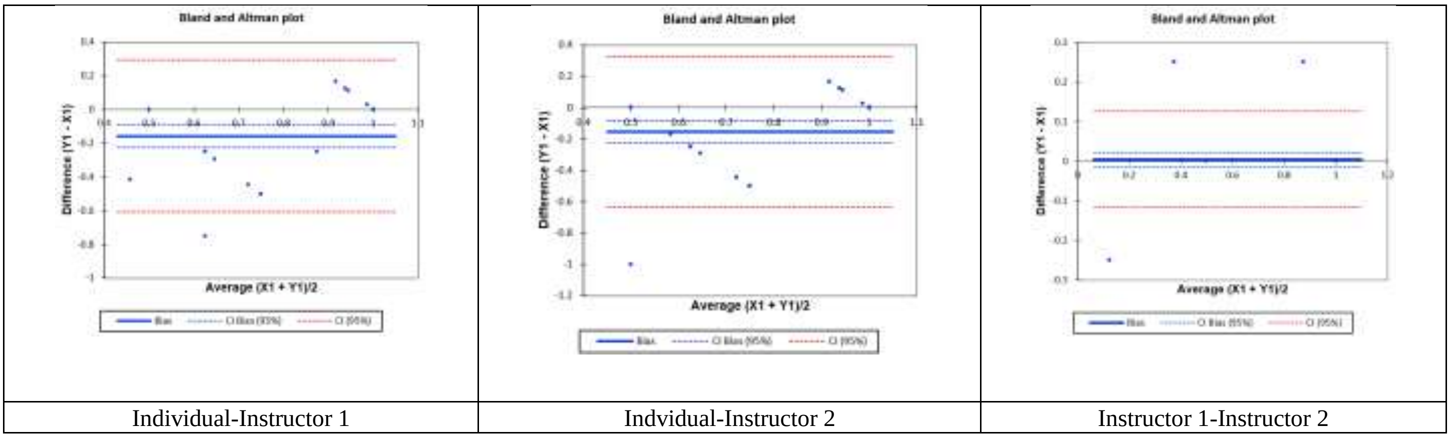
Midterm Exam

Table 3 reflects the differences between each instructor and student, and between instructors for the six rubric dimensions previously discussed. As can be seen, most values are relatively close to zero, other than “Design Intent.” These determinations, as a reminder, are made through professional experience and are only utilized to add one more validation step to full acceptance of the rubrics. These values, even including those for “Design Intent,” corroborate previous findings attained by other, more traditional statistical tests.

Table 3. Bland-Altman differences for midterm exam

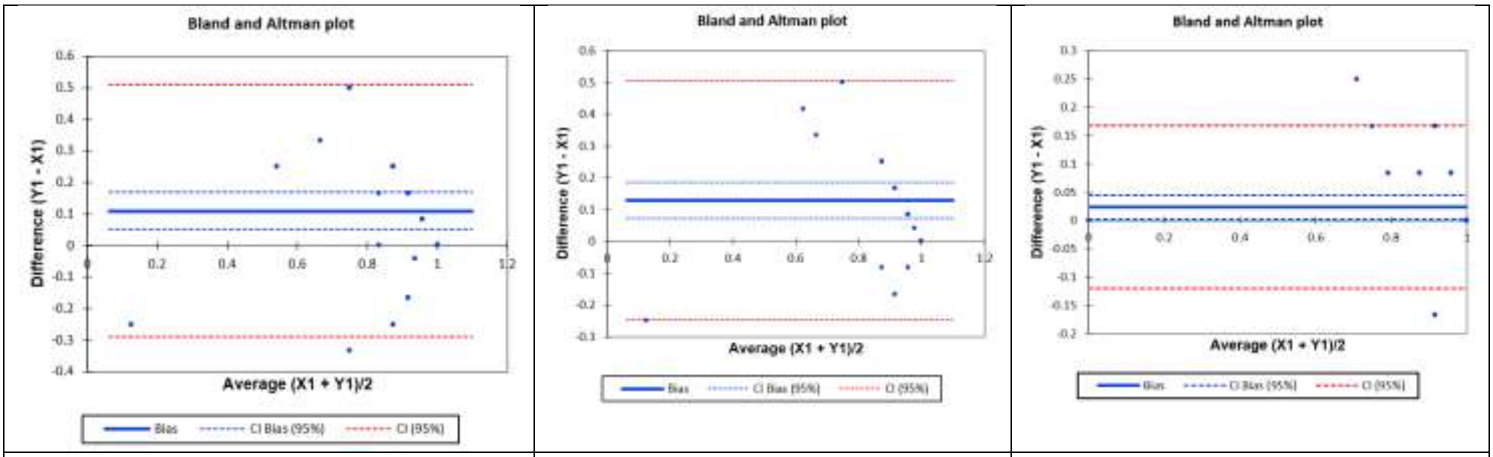
Difference	**Dim. 1** **(Valid)**	**Dim. 2** **(Complete)**	**Dim. 3** **(Consistent)**	**Dim. 4** **(Concise)**	**Dim. 5** **(Clear)**	**Dim. 6** **(Design Intent)**
Indiv. – Instr. 1	-0.157	0.110	0.012	-0.070	0.002	-0.339
Indiv. – Instr. 2	-0.152	0.130	0.045	-0.037	0.048	-0.230
Instr. 1. – Instr. 2	0.005	0.023	0.030	0.025	0.043	0.105

Figures 7 through 12 illustrate the Bland-Altman plots for the six rubric dimensions, between instructors and students. Most difference values lie between the acceptable band limits and mirror the difference values shown in Table 3. The limits of agreement appear to be reasonable. Variability appears to be consistent throughout the plots, irrespective of the rubric dimension.



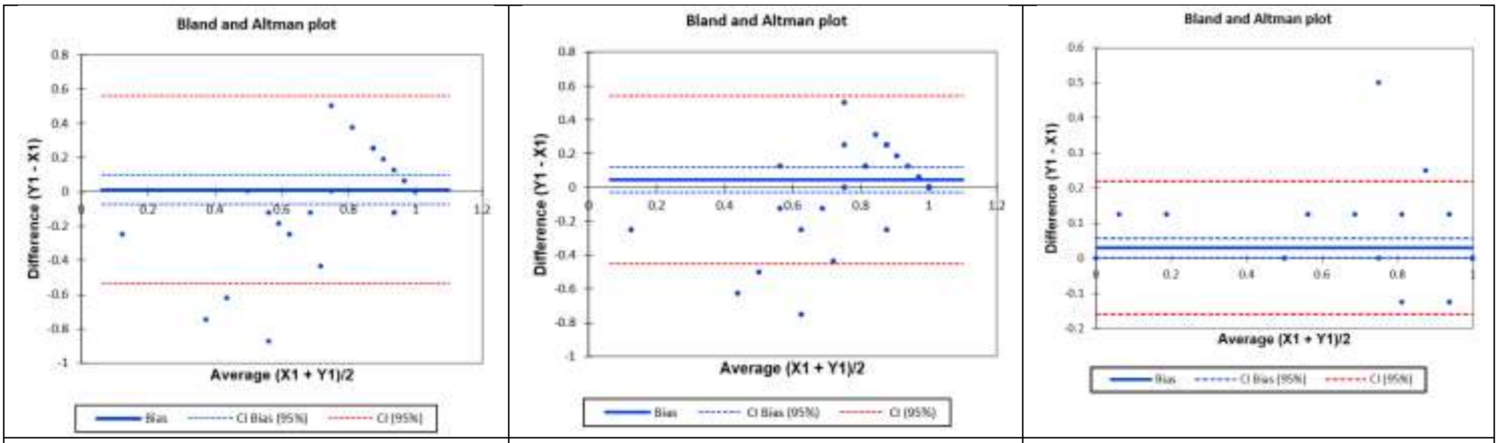
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 7. Bland-Altman plots for Validity on the midterm exam.



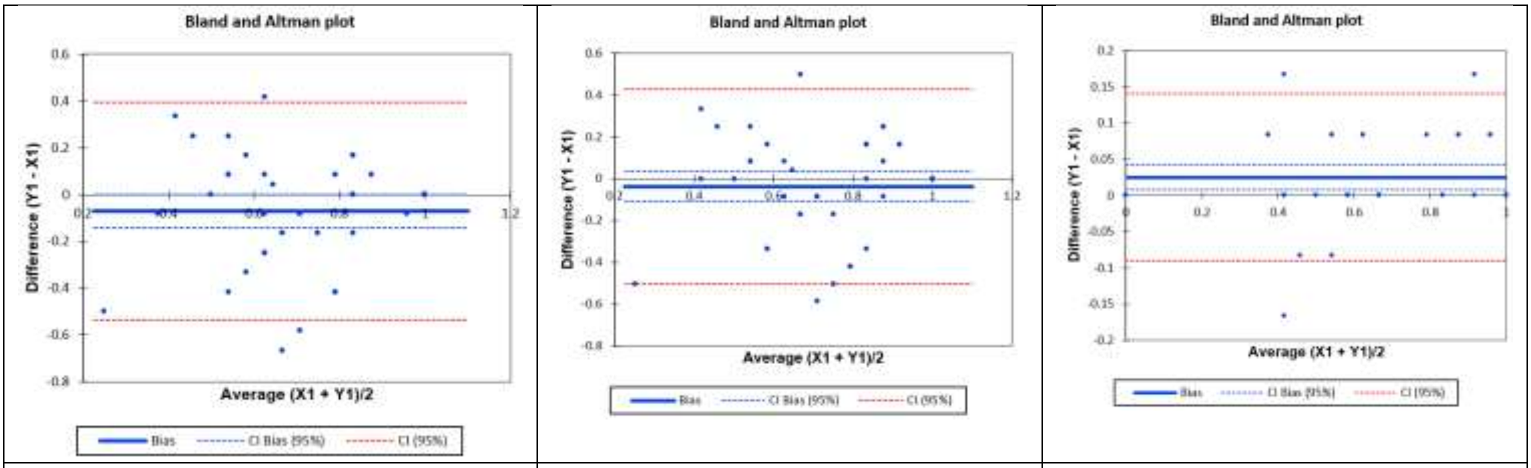
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 8. Bland-Altman plots for Completeness on the midterm exam.



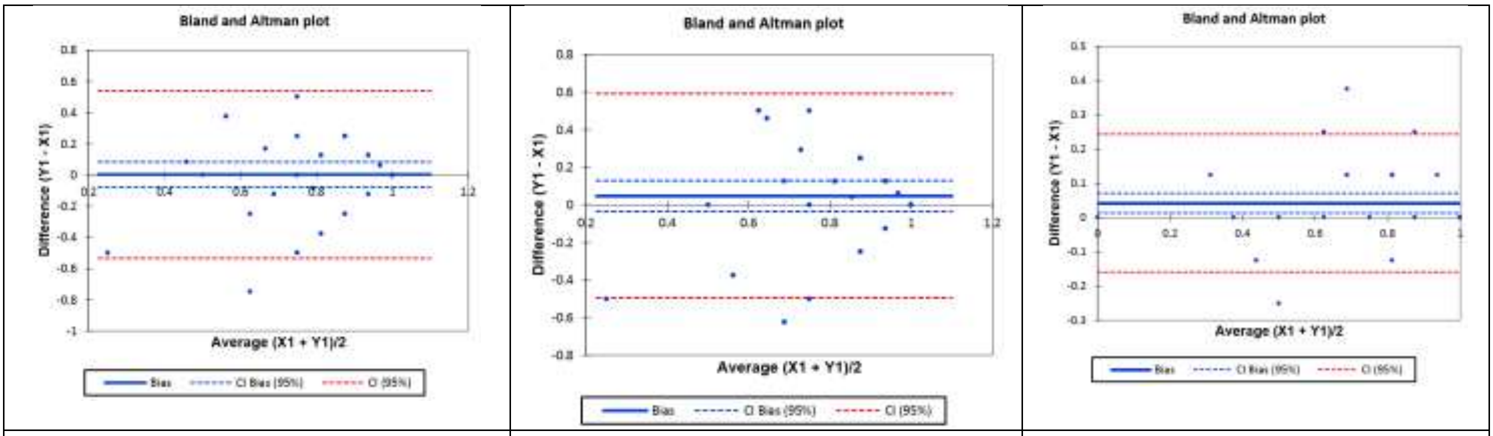
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 9. Bland-Altman plots for Consistency on the midterm exam.



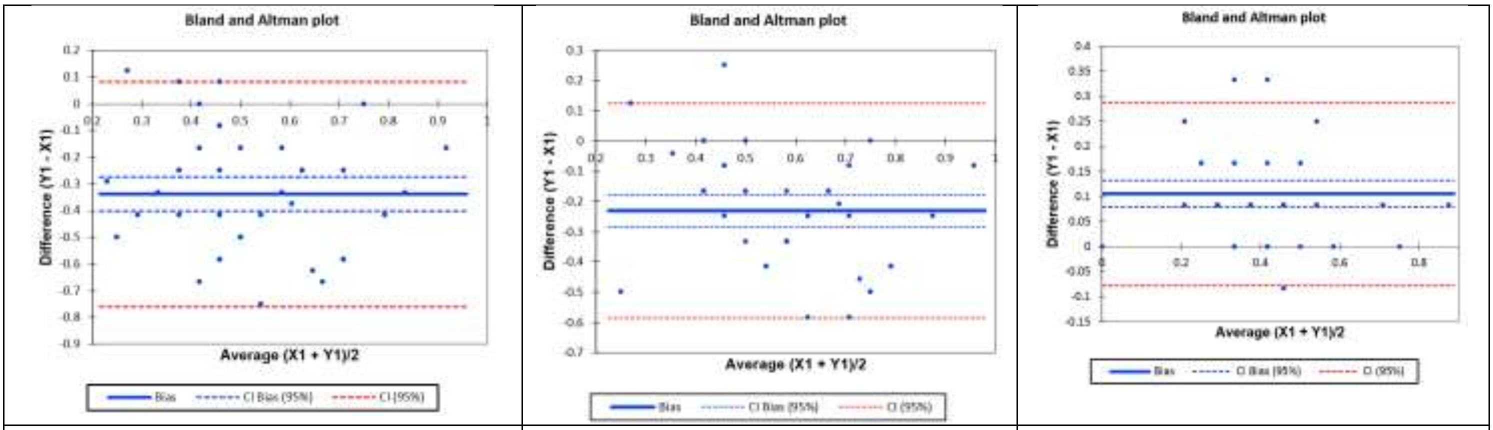
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 10. Bland-Altman plots for Conciseness on the midterm exam.



Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 11. Bland-Altman plots for Clarity on the midterm exam.



Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 12. Bland-Altman plots for Design Intent on the midterm exam.

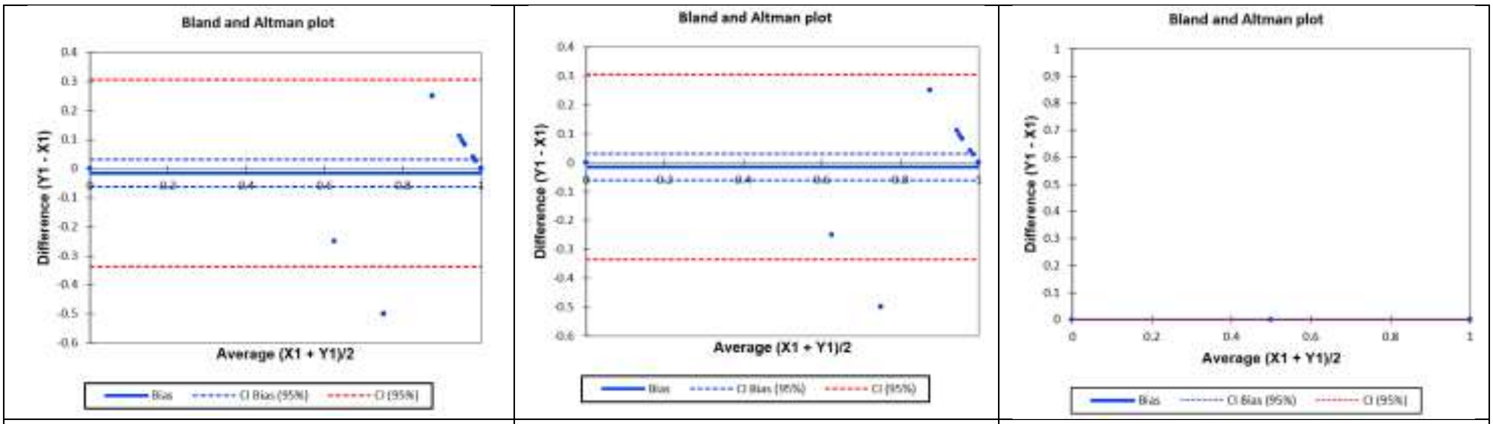
Final Exam

Table 4 reflects the differences for the six rubric dimensions. As noticed in the midterm exam, an overwhelming majority of the values are substantially close to zero, with “Design Intent” being furthest away, at least between Instructor 1 and the students. All values appear to be lower than the corresponding ones on the midterm exam, reflecting calibration thorugh experience. Once more, these findings validate previous discoveries using other statistical analyses.

Table 4. Bland-Altman differences for final exam

Difference	Dim. 1 (Valid)	Dim. 2 (Complete)	Dim. 3 (Consistent)	Dim. 4 (Concise)	Dim. 5 (Clear)	Dim. 6 (Design Intent)
Indiv. – Instr. 1	-0.015	0.025	-0.110	0.044	0.093	-0.144
Indiv. – Instr. 2	-0.015	0.046	-0.039	0.065	0.115	-0.052
Instr. 1. – Instr. 2	0.000	0.021	0.071	0.021	0.022	0.092

Figures 13 through 18 once again, illustrate the Bland-Altman plots for the six rubric dimensions, between instructors and students. As seen previously, most difference values lie between the acceptable band limits and mirror the difference values shown in Table 4. The limits of agreement appear to be reasonable and variability appears to be consistent throughout the plots.

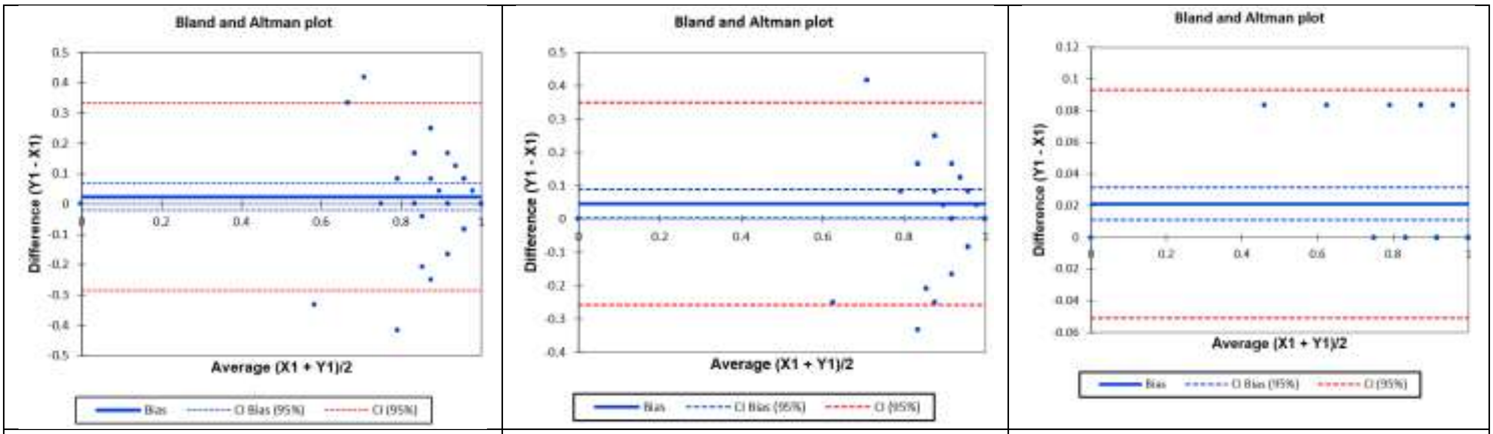


Individual-Instructor 1

Indvidual-Instructor 2

Instructor 1-Instructor 2

Fig. 13. Bland-Altman plots for Validity on the final exam.

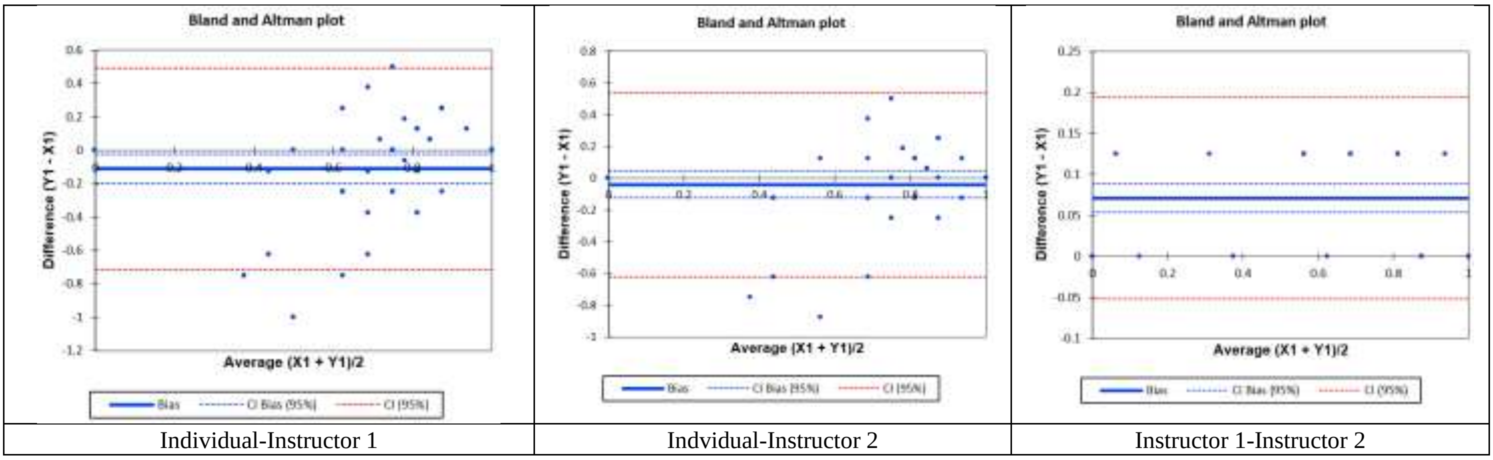


Individual-Instructor 1

Indvidual-Instructor 2

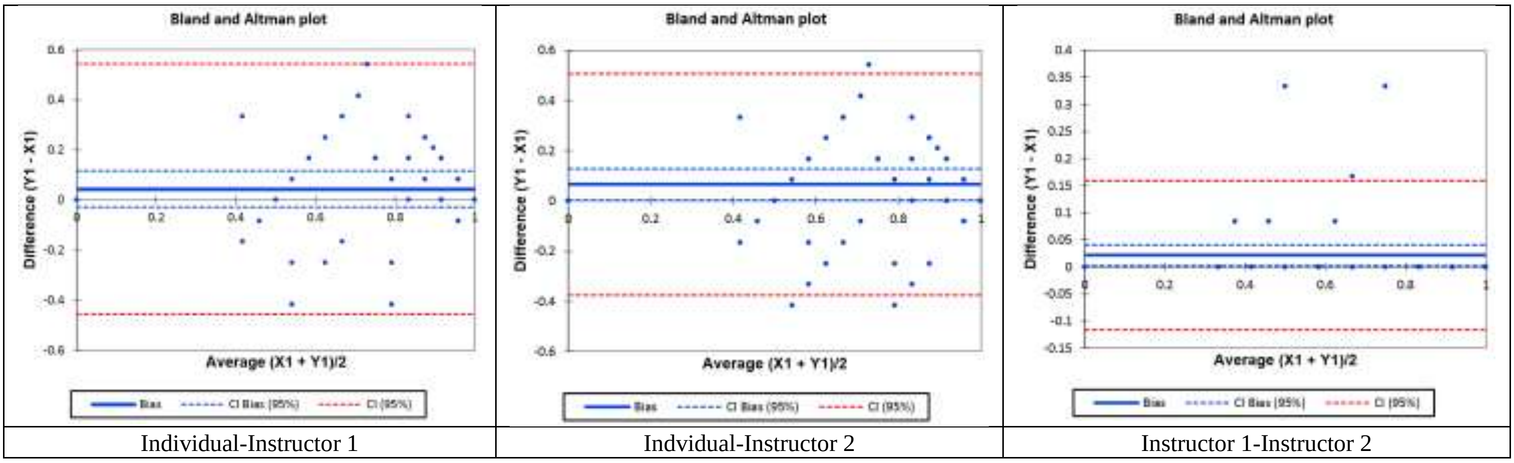
Instructor 1-Instructor 2

Fig. 14. Bland-Altman plots for Completeness on the final exam.



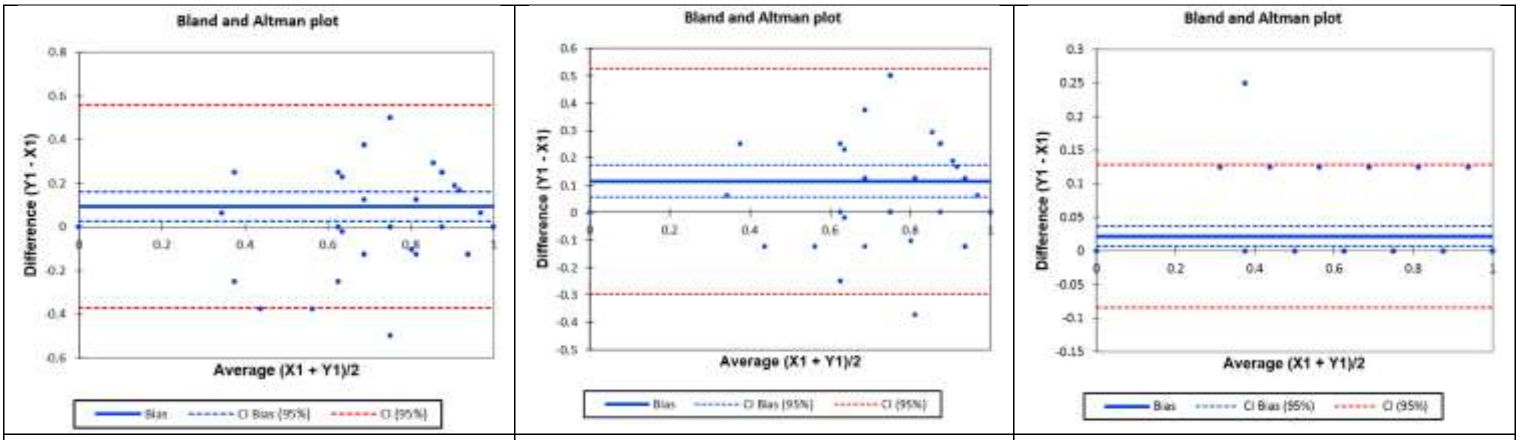
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 15. Bland-Altman plots for Consistency on the final exam.



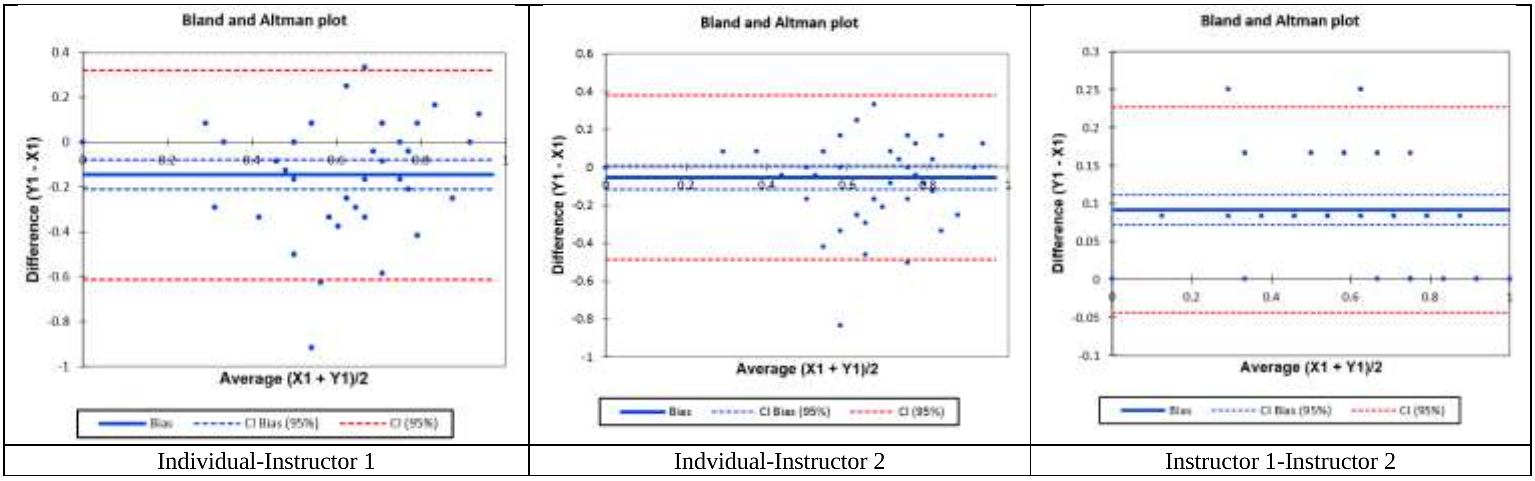
Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 16. Bland-Altman plots for Conciseness on the final exam.



Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 17. Bland-Altman plots for Clarity on the final exam.



Individual-Instructor 1 | Indvidual-Instructor 2 | Instructor 1-Instructor 2

Fig. 18. Bland-Altman plots for Design Intent on the final exam.

In a previous study, examining identical data from the midterm and final exams, it was shown that there was strong to moderate similitude for the dimensions of validity, completeness, conciseness, and clarity, while little correlation exists for consistency and design intent between instructors, with more agreement between instructors than between instructors and students [15]. Statistical tools included Wilcoxon Signed Rank Test, Kolmgorov-Smirnov test, Pearson Correlation, and Inter-rater Reliability. In running the Bland-Altman analysis in the current study, it can be easily seen that the previous determinations are confirmed, with very little outliers existing beyond the confidence bands. There exist more outliers with design intent, which is expected, although visual inspection of these plots confirms earlier research and conclusions. As a reminder, Bland-Altman is not a definitive statistical evaluation, but can be used to verify if hypothesis appears to be true and relies on professional expeience with the studied subject. As such, it results in addional support of prior determinations.

6 Conclusions

Rubrics, particularly summative rubrics, are mainly used to standardize and facilitate evaluation processes. However, rubrics can also become formative tools to convey performance information to students. This paper revisited the concept of rubrics to further extend some aspects related to making formative rubrics more adaptable and adaptive: criteria dichotomization, weighted evaluation criteria, and go/no-go criteria.

Formative rubrics should be discussed as learning content is presented, not after teaching. They should also adapt to the learning rhythm of students. According to the subject who controls adaptation, two types of dynamic rubrics are distinguished. Rubrics are adaptable if users can interactively vary the level of detail of criteria, and they are adaptive if the instructor can design different rubrics for different stages of formation. E-rubrics allow incorporating dynamic tools for adaptable and adaptive rubrics. The possibility to unfold or fold the level of detail of the criteria and query the anchors associated to the levels of attainment, allow the students to obtain an improved understanding of the matter.

Additionally, adaptive rubrics allow instructors progressive introduction of the evaluable concepts. Therefore, adaptive rubrics should be coordinated with the teaching guide. Timetables are suggested to plan the pace to introduce lower level criteria during the first weeks (following a bottom-up approach), and replace them later by more general criteria, whose knowledge may become exclusionary (by way of go/no-go criteria) at the end of the instructional period.

Go/no-go criteria (when a failure in one criterion is so critical that it prevents analyzing other aspects of the subject’s performance), are recommended, however they must be explicitly identified (and included as such) in the descriptor. Finally, go/no-go criteria can become soft by including a threshold parameter (e.g. after ten errors the assigned grade becomes zero, regardless of satisfying other rubric criteria).

Bland-Altman is not a definitive statistical test, as all interpretations are based on the actual experience with the subject. Bland-Altman sheds light on a question, and can be used to further reinforce other determinations, but will not provide an absolute answer in which to either reject or accept a hypothesis. Bland-Altman is used primarily in the medical field to assess different treatments, which is why there are so many questions about “agreement.” One patient’s medical condition is never identical to another therefore professional expertise is vital. Using Bland-Altman in an educational setting is a novel idea that warrants future exploration. If anything else, Bland-Altman serves as a mechanism to identify and visualize outliers.

Acknowledgements. This work was partially supported by grant DPI2017-84526-R (MINECO/AEI/FEDER, UE), project “CAL-MBE, Implementation and validation of a theoretical CAD quality model in a Model-Based Enterprise (MBE) context.”

References

[1] Popham W.J. (1997) What's wrong—and what's right—with rubrics. Educational Leadership, 55(2), pp. 72-75

[2] Educational Research Service (2004). Focus on: Developing and using instructional rubrics. Educational Research Service.

[3] Panadero E., Jonsson A. (2013) The use of scoring rubrics for formative assessment purposes revisited: A review. Educational Research Review, 9, pp. 129-144.

[4] Reddy Y.M., Andrade H. (2010) A review of rubric use in higher education. Assessment and Evaluation in Higher Education. 35(4), pp. 435-448.

[5] Company P., Contero M., Otey J., Plumed R. (2015) Approach for developing coordinated rubrics to convey quality criteria in MCAD training. Computer-Aided Design, 63, pp. 101–117.

[6] Company P., Contero M., Otey J., Camba J.D., Agost M.J., Perez-Lopez D. (2017) Web-Based system for adaptable rubrics: Case study on CAD assessment. Educational Technology & Society, 20, 3, 24-41.

[7] Tierney R., Simon M. (2004) What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. Practical Assessment, Research and Evaluation, 9(2), http://www.pareonline.net.

[8] Likert R. (1932) A technique for the measurement of attitudes. Archives of Psychology, 22 140, 55.

[9] Rohrmann B. (2007) Verbal qualifiers for rating scales: Sociolinguistic considerations and psychometric data. Project Report, University of Melbourne/Australia.

[10] Fluckiger J. (2010) Single point rubric: a tool for responsible student self-assessment. The Delta Kappa Gamma Bulletin, 76(4), pp. 18-25.

[11] Estell J.K., Sapp H.M., Reeping D. (2016). Work in progress: Developing single point rubrics for formative assessment. ASEE’s 123rd Annual Conference & Exposition, New Orleans, LA, USA, June 26-29. Paper ID #14595.

[12] Jonsson A., Svingby G. (2007) The Use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review, 2, pp. 130-144.

[13] Georgiadou E., Triantafillou E., Economides A.A. (2006) Evaluation parameters for computer-adaptive testing. British Journal of Educational Technology, 37, 2, 261–278.

[14] Company P., Otey J., Contero M., Agost M.J., Almiñana A. (2016) Implementation of adaptable rubrics for CAD model quality formative assessment purposes. International Journal of Engineering Education. Vol 32(2A), pp. 749–761.

[15] Otey J. (2017) A Contribution to conveying quality criteria in mechanical CAD models and assemblies through rubrics and comprehensive design intent qualification. PhD Thesis, Submitted to the Doctoral School of Universitat Politècnica de València.

[16] Watson P.F., Petrie A. (2010) Method agreement analysis: A review of correct methodology. Theriogenolgy, 73, 9, pp. 1167-1179.

[17] Kottner J., Streiner D.L. (2011) The difference between reliability and agreement. Journal of Clinical Epidemiology, 64, 6, pp. 701-702.

[18] McLaughlin P. (2013) Testing agreement between a new method and the gold standard-how do we test. Journal of Biomechanics, 46, pp. 2757-2760.

[19] Costa-Santos C., Bernardes J., Ayres-de-Campos D., Costa A., Costa C. (2011) The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. Journal of Clinical Epidemiology, 64, 3, pp. 264-269.

[20] Chen C.C., Barnhart H.X. (2013) Assessing agreement with intraclass correlation coefficient and concordance correlation coefficient for data with repeated measures. Computational Statistics and Data Analysis, 60, pp. 132-145.

[21] Bland J.M., Altman D. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. The Lancet, 327, 8476, pp. 307-310.

[22] Van Stralen K.J., Jager K.J., Zoccali C., Dekker F.W. (2008) Agreement between methods. Kidney International, 74, 9, pp. 1116-1120.

[23] Beckstead J.W. (2011) Agreement, reliability, and bias in measurement: Commentary on Bland and Altman (1986:2010). International Journal of Nursing Studies, 48, pp. 134-135.

[24] Bland J.M., Altman D. (1999) Measuring agreement in method comparison studies. Statistical Methods in Medical Research, 8, pp. 135-160.

[25] Giavarina D. (2015) Understanding Bland Altman analysis. Biochemia Medica, 25, 2, pp. 141-151.

[26] GraphPad. (1995) Interpreting results: Bland-Altman. Retrieved from https://www.graphpad.com/guides/prism/7/statistics/bland-altman_results.htm?toc=0&printWindow.