

Web-Based System for Adaptable Rubrics: Case Study on CAD Assessment

Pedro Company^{1*}, Manuel Contero², Jeffrey Otey³, Jorge D. Camba⁴, María-Jesús Agost⁵ and David Pérez-López²

¹Institute of New Imaging Technologies, Universitat Jaume I, Castellón, Spain // ²I3B, Universitat Politècnica de València, Valencia, Spain // ³Zachry Department of Civil Engineering, Texas A&M University, TX, USA //

⁴Gerald D. Hines College of Architecture and Design, University of Houston, TX, USA // ⁵Department of Mechanical Engineering and Construction, Universitat Jaume I, Castellón, Spain // pcompany@uji.es // mcontero@upv.es // j-otey@tamu.edu // jdorribo@uh.edu // magost@uji.es // dapelo@i3b.upv.es

*Corresponding author

(Submitted December 23, 2015; Revised August 16, 2016; Accepted August 23, 2016)

ABSTRACT

This paper describes the implementation and testing of our concept of adaptable rubrics, defined as analytical rubrics that arrange assessment criteria at multiple levels that can be expanded on demand. Because of its adaptable nature, these rubrics cannot be implemented in paper formats, neither are they supported by current Learning Management Systems (LMS). The main contribution of this work involves the adaptable capability of different levels of detail, which can be expanded for each rubric criterion as needed. Our rubrics platform provides specialized and intuitive tools to create and modify rubrics as well as managing metadata to support learning analytics. As an example of a practical assessment situation, a case study on Mechanical Computer Aided Design (MCAD) systems training is presented. The validation process in this scenario proved the effectiveness of our adaptable rubric platform for supporting formative assessment in a multifaceted and complex field such as MCAD. The system also showed the potential of collecting user interaction metadata, which can be used to analyze the evaluation process and guide further improvements in the teaching strategy.

Keywords

Adaptable rubrics, Rubrics platform, Formative assessment, Learning Management Systems

Introduction

Rubrics are a recognized instrument to support authentic assessments to describe student achievement (Andrade, 1996; Andrade, 2000). A rubric can be defined as a scoring tool that provides a set of criteria to assess a piece of work and includes gradations of quality or performance for each criterion. Rubrics can increase student-learning outcomes by making teachers' expectations explicit and by showing students how to meet these expectations (by presenting what level of quality is expected from their work). Rubrics are also useful to help students develop a critical sense of their own work by providing them with criteria to become more thoughtful judges of the quality of their own and others' work.

True assessment emphasizes the application and use of knowledge to solve complex tasks that involve contextualized problems. Rubrics help students to understand the criteria for judgment from the beginning of their instruction (Montgomery, 2002). As tasks become more complex, there is often a gradual degradation of the structure and comprehension of the rubric. This problem is manifested especially when analytical rubrics are used (rubrics that break the evaluation down to simple components that are scored separately and then combined to produce the global evaluation). Quality criteria are difficult to use, both by teachers and students, if they become too abstract. A typical approach to rectify this issue is to disaggregate the complex criteria into a series of more understandable criteria of lower conceptual difficulty. A problem arises when a compact list of abstract or dense criteria is replaced by a long list of simpler ones, which in many cases can make them impractical and time-consuming. This situation is accentuated when each criterion is weighted to reflect its relative importance. Current Learning Management Systems (LMS) do not provide a solution to this problem, and as a result, analytical rubrics are often avoided in complex situations.

In this context, the concept of an adaptable rubric emerges as a powerful mechanism to support different learning styles and rhythms. We define adaptable rubrics as those that provide multiple levels of detail, which can be expanded on demand. The level of detail can be adjusted and adapted to a specific teaching scenario and/or the students' level of understanding of quality concepts. If a student finds a particular criterion or its performance levels too difficult to understand, he/she can deploy an additional level of detail (if provided) for that specific criterion, where it is divided into several sub-criteria with a lower abstraction level.

In this paper, we present a new computer-assisted rubric platform specifically designed to support adaptable rubrics. The main features of this platform are:

- provides feedback (showing detailed scores and levels of performance, if requested).
- supports different learning rhythms and styles (different levels of detail are deployed on demand by students, based on their choice).
- collects metadata that could be used to support adaptive behavior in the future.
- automates the management of different weights among scoring criteria during rubric creation.

The platform is generic, as it can be used to manage any type of rubric. The implementation strategy, validation, and lessons learned while developing and testing our platform are presented in the paper. As an example of the application to a highly difficult and complex assessment problem, the developed system was used in a Mechanical Computer Aided Design (MCAD) training scenario at the undergraduate and graduate college level.

The paper is structured as follows: second section describes the state of the art in platforms for scoring rubrics and confirms the lack of support for adaptable rubrics. Third section describes the architecture of the proposed system. The description includes design specifications and the most relevant implementation details. Two experiments aimed at validating the system are described in fourth section. The first experiment was based on spreadsheet forms while the second was based on our new platform. Results confirm that the new tool does not negatively affect the evaluation process. Furthermore, the first experiment demonstrates the weaknesses of spreadsheet forms while the second experiment demonstrates that the new platform provides richer and more meaningful information than other systems. We conclude by highlighting the lessons learned which will be used to guide future developments and improvements of the platform.

Related work

Many authors claim that rubrics can both teach and evaluate (Andrade & Du, 2005; Jonsson & Svingby, 2007; Reddy & Andrade, 2010). In fact, the formative use of scoring rubrics has proven useful under many circumstances (Panadero & Jonsson, 2013). Unfortunately, the use of rubrics in complex contexts is a non-trivial challenge. For instance, MCAD systems are complex software tools that require not only a thorough understanding of the various functionalities provided by the system, but also the application of efficient strategies to create high quality CAD models. Based on related research, this issue can be considered a particular type of the “content-based constructed responses” problem (Liu et al., 2014).

Computer-assisted grading rubrics are essential to improve the efficiency and effectiveness of grading (Anglin et al., 2008; Auvinen et al., 2009). Although some commercially available Computer Assisted Assessment (CAA) tools provide automatically contextualized feedback (Santos et al., 2009; Cebrián-Robles et al., 2014), and some Learning Management Systems (LMS) provide rubric functionality (Atkinson & Lim, 2013; Isbell & Goomas, 2014), currently available implementations only support static rubrics, which are not flexible in adapting to different learning scenarios.

Building on the general concepts of using rubrics to capture judgment (Mertler, 2001; Karkehabadi, 2013) and the evaluative use of rubrics in higher education (Reddy & Andrade, 2010), this paper focuses on the role of rubrics as instruments for acquiring or reinforcing complex skills (Manson & Olsen 2010; Smit & Birri 2014), assessing engineering questions (McCormick et al., 2015), and supporting the process of formative assessment (Popham, 1997) by advising students about their progress and assisting them in their development (Panadero & Jonsson, 2013).

Contrary to holistic rubrics—which only score the overall process or product as a whole without judging the individual components separately (Nitko, 2001)—analytic rubrics are typically used for formative feedback (Mertler, 2001), as they allow the individual scoring of factors—or dimensions—of the product or performance (Moskal, 2000). However, a common problem when introducing performance assessment is evaluating complex competences in a credible way (i.e., whether or not observations of complex behavior can be performed in a reliable and trustworthy manner (Jonsson & Svingby, 2007)). In complex assessment scenarios, when comparing instructor and student judgments, assessment differences can be attributed to the fact that the concepts under evaluation may still be so foreign that students are unable to recognize them (Orsmond, Merry, & Reiling, 1996). This lack of consensus between the instructor assessment and the student self-assessment—i.e., the lack of inter-rater reliability—is useful in detecting problems in the understanding of quality criteria. This type of data can be processed by applying proper statistical analyses to the assessed rubrics (Zaiontz, 2015).

Additionally, we are also interested in metadata that describes how users interact with the platform while completing the rubrics. It has been argued that rubrics should be complemented with “anchors,” i.e., written descriptions, examples that illustrate the various levels of attainment, or work samples (Jonsson & Svingby, 2007). To the best of our knowledge, no results have been reported regarding the use of metadata to discover improved methods to increase the effectiveness of anchors.

Although advanced trainees may not require detailed explanations of each evaluation criterion, more detail may be desirable by novice users. Self-directed learning skills (involving self-assessment and task selection skills) are needed to choose an appropriate on-demand learning pathway (Taminiau et al., 2015). Adaptable rubrics should allow students to display more detail and score low-level criteria on demand. If a particular criterion in the rubric is too abstract or difficult to understand, the additional level of detail can provide a clearer description of the expected performance levels (Company et al., 2015). A major challenge with adaptable rubrics involves guaranteeing consistent scoring for each different combination of detailed-level responses. In this arena, spreadsheets have proven impractical, as its implementation requires significant programming proficiency and extracting relevant information from them is time consuming (Company et al., 2016). Furthermore, they can hardly convey performance criteria descriptors, which are critical components of rubric design (Tierney & Simon, 2004).

To distinguish user-driven rubrics from system-driven rubrics, we adopt the terms adaptable and adaptive. While an adaptable rubric can be modified by the user to adapt to different needs, an adaptive rubric is able to adapt or change itself, depending on the usage pattern. The origins of adaptive rubrics are briefly summarized by Georgiadou et al. (2006).

According to Economides and Roupas (2007), the majority of the Computer Adaptive Testing (CAT) systems “do not offer any advanced support and functionalities to the examinee.” In analyzing current available LMS systems such as Moodle, Sakai, Blackboard or Desire to Learn (D2L), it was found that none of them implement functionality to support adaptable rubrics. They only support static rubrics that cannot deploy additional levels of detail to improve students’ understanding.

Electronic platform for adaptable rubrics

This section describes the design and implementation of a dedicated rubrics platform, with support for adaptable rubrics, and the ability to output metadata to analyze the evaluation process to guide further improvements.

Design specifications

Apart from supporting adaptable rubrics, the main goal of our design was to centralize both rubrics and assessment results in order to improve data exchange and reduce the time required to program and manipulate a rubric. An additional goal was to make the rubric creation process easier by providing specialized and intuitive tools to enter new rubrics into the system and edit existing ones. Finally, metadata management to support learning analytics was also included. Metadata describing the context of the evaluation process is required to better understand current problems and guide future improvements. We define metadata as the dataset collected by tracking how users interact with the platform while completing the rubrics. Spreadsheet-forms (which cannot collect such data) and proprietary “black-box” rubric platforms (which do not grant direct access to the data) are unsuitable.

Two main specifications were defined for our rubrics platform: (1) the system must be adaptable and allow for rubrics with varying levels of detail, and (2) the rubrics should be easy to reconfigure by allowing instructors to adjust them to changing scenarios.

The first specification implies that every participant that uses the rubric to score a particular task should be able to select the desired level of detail. The rationale behind our adaptable rubrics is the “expand–contract strategy,” which adapts the rubrics to trainee progress and assists them in comprehending the different dimensions of the rubric (Company et al., 2015). Although technically there is not a limit to the number of supported levels, our first prototype has been intentionally limited to three levels. The second specification is intended to allow instructors to customize rubrics based on the evolution of student learning. In addition, our long-term goal involves linking the rubrics to appropriate anchors, as defined by (Jonsson & Svingby, 2007).

Three additional design specifications were also considered: (3) the system must provide immediate feedback of the evaluation scores; (4) instructors should be able to easily extract and process the information of the completed forms, and (5) forms should prevent incomplete or inconsistent scoring.

From a user standpoint, the platform manages three types of accounts: students, instructors, and administrators. After a successful login to the system, students are allowed to perform evaluations using rubrics (both self and peer evaluation). Instructors are allowed to manage rubrics, manage student users and groups, assign rubrics to groups, schedule rubric assignments, and manage results. A global administrator can create new instructors.

Prototype implementation

Before implementing a dedicated rubrics platform, our team tested the concept of adaptable rubrics using a spreadsheet-based mock-up, which was used in a complex assessment scenario, showing promising results regarding students' improved management and understanding of the rubric (Company et al., 2016). A spreadsheet was configured to support rubrics with multilevel criteria, which could be unfolded using buttons. A five-level Likert scale implemented by radio buttons represented performance levels. Active X controls and Visual Basic for Applications (VBA) were used for the implementation. An example of deployment of sub-criteria is presented in Figure 1. In this case, when level 2 sub-criteria for criterion 1 are unfolded, criterion 1 is blocked, and the student can only mark sub-criteria 1.1 and 1.2. As a result, the score for criterion 1 is automatically calculated and displayed as the total score for this criterion.

MODEL

OVERALL SCORE
INCOMPLETE
0.00%

Criterion	Assertion	Weight	No/ Never	Almost never	Some-times	Almost always	Yes/ Always	Score
1	The model is valid	0%	Score the sub-criteria					0.00%
1.1	The file of the model can be found and opens in neutral state	0.0%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
1.2	The file contains the expected model, and the model can be used	0.0%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
2	The model is complete	25%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
3	The model is consistent	20%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
4	The model is concise	25%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
5	The model is clear	15%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%
6	The model conveys design intent	15%	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	0.00%

Figure 1. Spreadsheet-based adaptable rubric

Implementation

Spreadsheet-based mock-up showed many limitations to create a general tool to support the creation and use of adaptable rubrics, so our team opted to develop Annota eRubrics, a web-based framework to manage adaptable rubrics and users. Three user profiles were defined: student, instructor, and global administrator. Student users can fill out rubrics and visualize completed rubrics during the evaluation period defined by the instructor. Instructors can manage rubrics (create new, import from CSV or XML files, export to files, and assign rubrics to students during a certain period of time), manage students (register new, import from CSV files, and manage groups of students), and access assessment results (view and export to files). Finally, global administrators (who can also be instructors) can register new instructors.

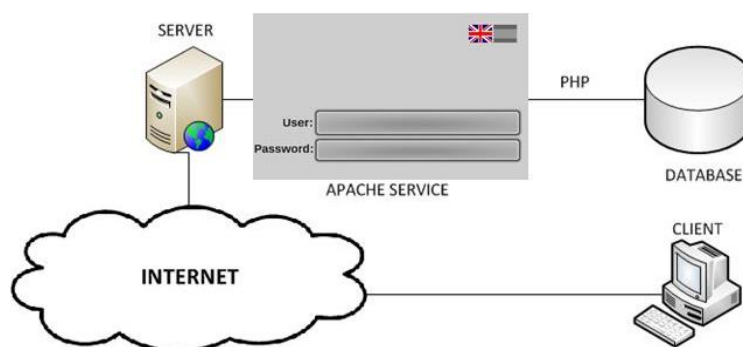


Figure 2. Annota framework

The general structure of the framework is shown in Figure 2. Data is stored in a MySQL database and accessed via an Apache web server by a number of PHP scripts triggered by the framework front-end running over Unity3D Webplayer (Creighton, 2010). Only name and e-mail address are required to create a new user. During registration, a custom link is automatically emailed to the users so they can set up their password and enter basic demographics information to complete the process.

A screen capture of a sample MCAD rubric with 6 main criteria (for the 6 CAD quality dimensions to be assessed) is shown in Figure 3.

		No/Never	Almost never	Sometimes	Almost always	Yes/Always
+ 1	The assembly is valid	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
+ 2	The assembly is complete	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
+ 3	The assembly is consistent	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
+ 4	The assembly is concise	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>
+ 5	The assembly is clear	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
+ 6	The assembly conveys design intent	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Figure 3. MCAD high level assembly (6 main criteria) rubric as implemented in the Annota platform

An example of the mechanism to unfold lower level sub-criteria is illustrated in Figure 4. By clicking the “+” (unfold) or “-” (fold) signs located on the upper left corner of the criteria (criteria 1 and 1.2 have been unfolded in Figure 4), users can dynamically adapt the rubric to their own rhythms and learning styles. Instructors can pre-configure criteria as “folded” or “unfolded” to determine which criteria will be viewed folded/unfolded by default when students load the rubric for the first time.

- 1	The assembly is valid	Evaluated by subcriteria				
+ 1.1	The file of the assembly can be located and opens in a neutral state	No/Never	Almost never	Sometimes	Almost always	Yes/Always
- 1.2	The file contains the expected assembly, and the assembly can be used	Evaluated by subcriteria				
1.2.1	The file contains the expected assembly	No/Never	Almost never	Sometimes	Almost always	Yes/Always
1.2.2	Assembly tree is free from error messages	No/Never	Almost never	Sometimes	Almost always	Yes/Always

Figure 4. Annota rubrics are adaptable by allowing users to fold and unfold levels of detail

Each criterion has an associated scoring weight configured by the instructor (Figure 5). Assessment results are recalculated every time the student fills out a criterion. By default, weights and assessment results are not visible to the student but can be activated at any time. This activation/deactivation is part of the metadata collected by the system.



Figure 5. Example MCAD criterion with an associated scoring weight (in this case a 10.00 % weight)

As a result of our first experiment (described in the next section), contextual information bubbles were added to various options of the rubric. These bubbles display textual explanations for each level of deployment, also called performance criteria descriptors, and are automatically shown when the user hovers over a check box (Figure 6). These bubbles provide a complementary explanation of the Likert level associated to a specific criterion. The number of bubble activations and the bubble activation times are also part of the collected metadata.



Figure 6. Example of bubble with a textual explanation for a particular level of deployment

From an instructor standpoint, a “student view” is available, which allows instructors to view the rubric as members of the student group. This mode is for testing purposes only, so results are not saved. In addition, instructors can grade students by importing a CSV file. This functionality allows the use of spreadsheets to grade an entire user group with a single mouse click.

Experiments

To validate our platform, we selected a complex assessment problem in the area of MCAD training. We conducted two experiments with various groups of MCAD students. The first experiment validates adaptable rubrics for feedback and evaluation, but states the limitations of spreadsheet-based adaptable rubrics. It also provides a comparison base to determine that the new form, used in the second experiment, does not negatively affect the behavior of students. The second experiment also shows how our new platform for adaptable rubrics can offer additional guidelines to help evaluation, when required. Finally, the second experiment illustrates a future development in which the metadata generated by the new platform can be used as a resource to identify weaknesses, measure effectiveness, and lead to further improvements.

Experiment 1

The goal of our first experiment is to test the feasibility and analyze the benefits of the adaptable rubrics concept. We use a learning/teaching scenario based on the combination of suitable tutorials and materials during instruction with a set of adaptable rubrics for feedback and evaluation. The context is MCAD training.

Our sample was a multi-disciplinary group of senior engineering students at a US university. Students were divided into two groups: experimental (EG) and control (CG). Both groups were given a complete set of lecture notes, which included detailed explanations of the different dimensions of MCAD quality (we refer to it as theoretical background). The lecture notes given to the EG also included detailed explanations and examples on how to apply rubrics to self-evaluate in-class exercises (practical guidelines). This information was intentionally removed from the lecture notes that were given to the CG.

Participants were asked to create 3D solid models of two parts (see Figure 7). As an additional requirement for the second part, participants were explicitly instructed to make their models flexible and reusable by allowing a series of specific design changes to be performed successfully and efficiently. Changes for the part 2 EG included: Increasing the distance between the cylinders with holes (for example, from 2 to 3 inches), modifying the height of the Ø3/4 cylinder (for example, from 7/8 to 1 inch), and rotating the arm (for example, 90 degrees). Changes for the part 2 CG included: Increasing the distance between the center points of the countersunk holes (for example, from 90 mm to 150 mm), increasing the height of the main cylinder (for example, from 65 mm to 100 mm), and rotating the arm (for example, 90 degrees).

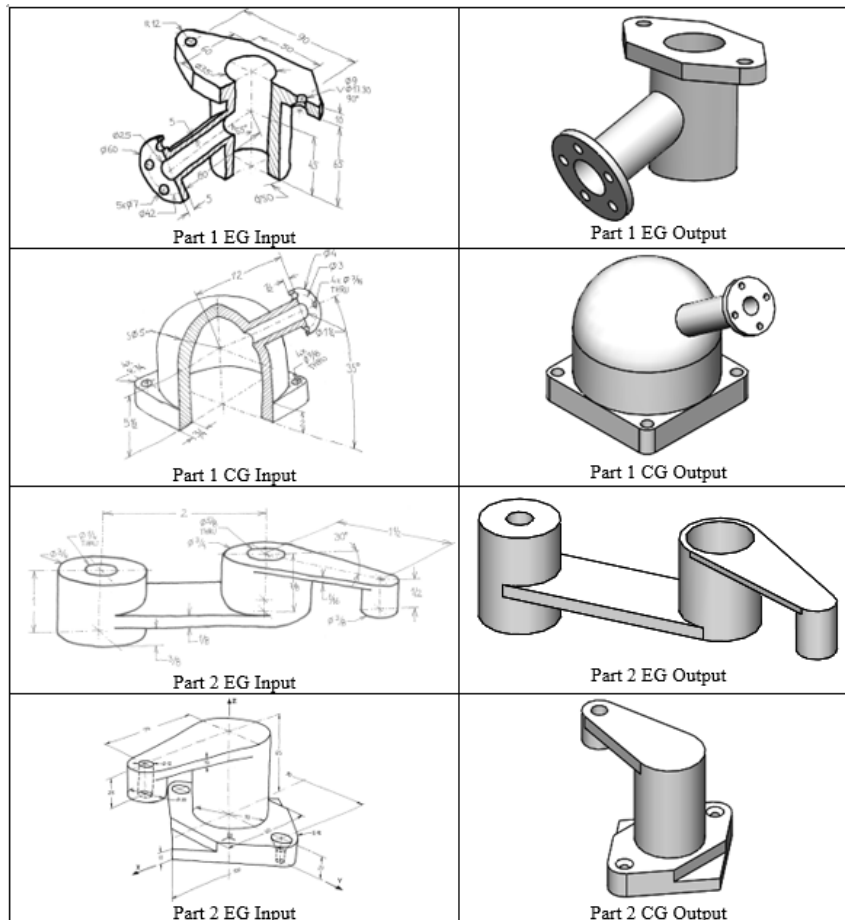


Figure 7. Parts used by the EG and the CG in experiment 1

All subjects were asked to self-evaluate their work using spreadsheet-based adaptable rubrics. After submitting the first self-evaluation, they were given a sample “solution:” step-by-step instructions of an efficient strategy (defined by our research team) to model the parts, which also included detailed explanations on how to evaluate the quality of the models. Provided with this information, they were asked to re-evaluate their original models a second time.

The rubric used to evaluate the experiment was necessarily multidimensional, as is the nature of the quality of CAD models. Thus, a multilevel criteria tree was selected. The detailed sub-dimensions schema used for the experiment is shown in the tree diagrams of Figures 8 and 9.

The initial sample size of the study was 50 subjects in the EG and 49 subjects in the CG, but only 29 in the EG and 26 in the CG completed all tasks.

As expected, all the inconveniences of spreadsheet-based forms soon emerged. For example, 8.8% of the rubric forms (29/330) were returned unlocked and evaluated simultaneously at various levels. This issue resulted in inconsistent scores between the main levels and their corresponding sub-levels. Also, the process of extracting the information from the forms was time consuming and prone to errors. The only metadata we could obtain from those forms was that only 7 out of 330 rubrics were assessed without showing the score (thus, the usefulness of such feedback was demonstrated). Finally, by analyzing student submissions, we determined that 53% of the rubrics were always used at the third level (175/330), while 8% of the assessments only deployed the first level (26/330). The remaining assessments used Level 1 for some dimensions and Levels 2 or 3 for others, thus demonstrating that the rhythm of assimilation of quality concepts varies among students and validates the need for adaptable rubrics.

We searched for dissimilarities in the understanding of quality criteria by comparing significant differences between inter-rater evaluations. Results for Part 1 EG and Part 1 CG illustrated similar behavior in both user groups. Figures 10 and 11 visually summarize the detailed quantitative results by showing the differences between the instructor, self, and peer evaluations. Each quality criterion is marked with a grayed background if

the student and the instructor agree within a threshold (the threshold Delta was set to 0.25, which is equivalent to a jump between two consecutive levels in a five-level Likert item mapped into the range [0, 1]), and is void for larger differences.

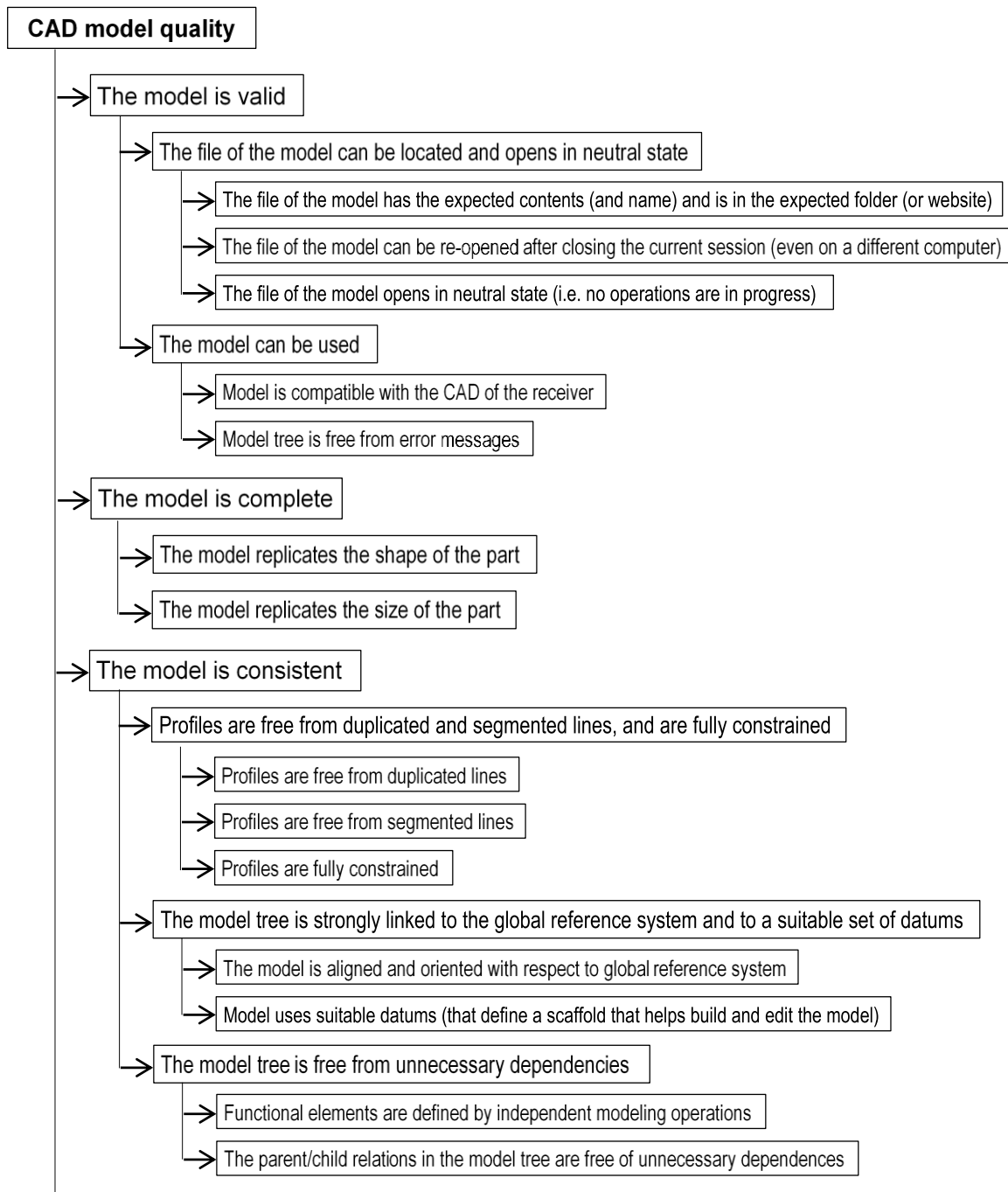


Figure 8. Tree structure of the assessment criteria used to evaluate the three first main criteria of CAD model quality

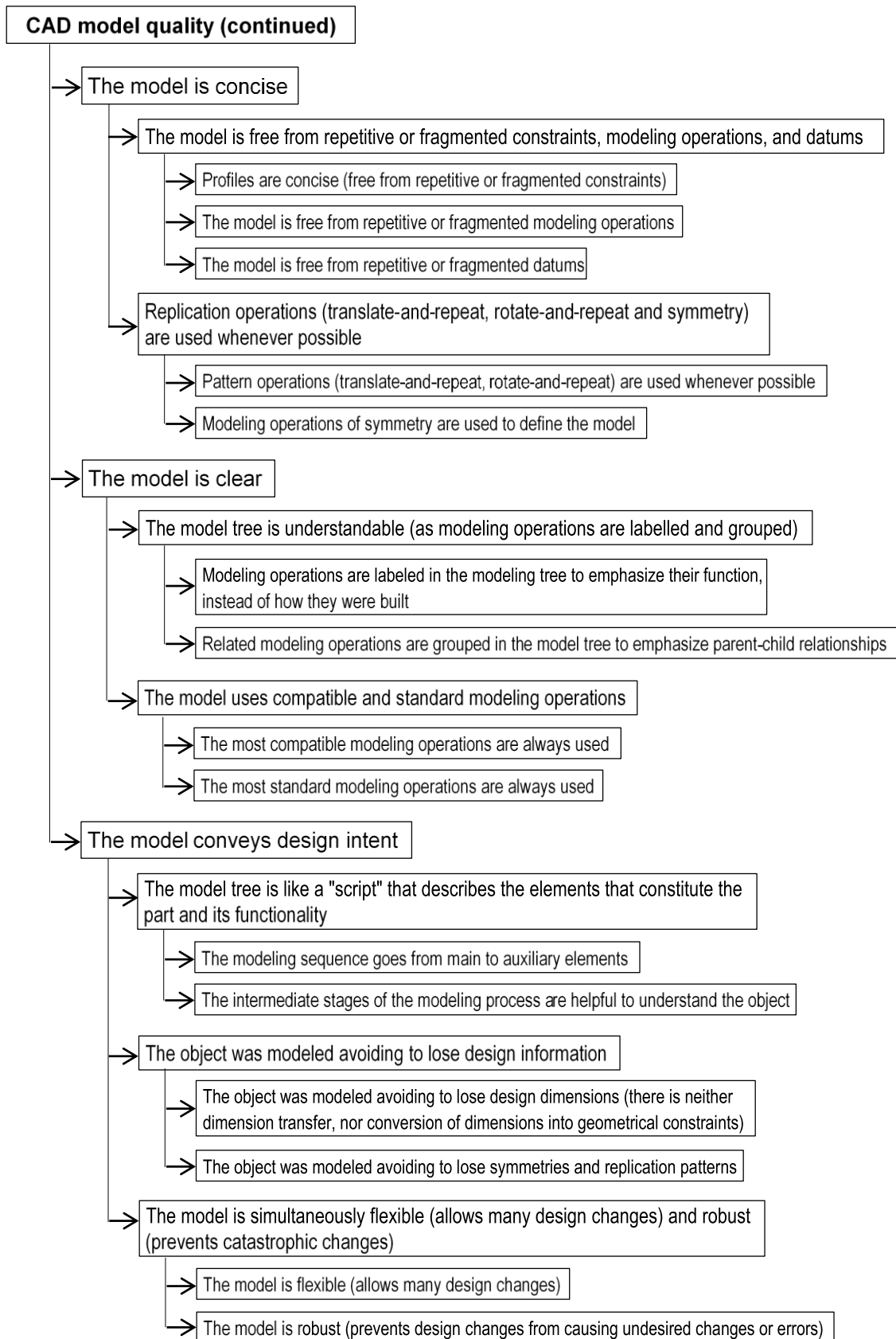


Figure 9. Tree structure of the assessment criteria used to evaluate the three last main criteria of CAD model quality

Part 1	Self1 to Instructor $\Delta= 0.25$							Self 2 to Instructor $\Delta= 0.25$							Peer to Instructor $\Delta= 0.25$								
	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements		
CG34							5							6								5	
CG21							1							4									3
CG25							1							4									3
CG46							3							4									2
CG37							0							4									1
CG13							2							2									2
CG33							3							3									4
CG35							5							5									4
CG9							6							5									6
CG44							6							5									5
CG38							4							4									4
CG42							3							2									4
CG6							1							1									1
CG27							4							4									1
CG17							4							3									3
CG43							4							4									1
CG49							3							5									4
CG4							4							4									5
CG23							0							2									2
CG8							2							2									2
CG30							5							6									5
CG29							5							5									3
CG18							4							6									4
CG5							5							3									5
CG15							0							1									4
CG48							3							3									3
% Agreement	73	58	69	46	46	27	53	96	73	69	65	38	31	62	96	65	46	50	50	23	55		

Figure 10. Visual map of the agreements between the instructor and self and peer evaluations for the CG (part 1)

Part 1	Self1 to Instructor $\Delta= 0.25$							Self 2 to Instructor $\Delta= 0.25$							Peer to Instructor $\Delta= 0.25$								
	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements		
EG29							3							2									2
EG28							2							2									1
EG30							4							5									4
EG31							3							3									3
EG48							4							5									4
EG35							0							1									2
EG15							1							2									2
EG44							5							5									5
EG6							1							4									2
EG13							6							5									5
EG27							2							2									2
EG46							4							4									5
EG9							0							1									1
EG5							2							2									3
EG39							1							3									3
EG19							4							5									5
EG23							6							6									6
EG17							1							3									1
EG12							4							4									4
EG26							1							1									2
EG38							2							4									5
EG47							6							6									6
EG3							5							4									6
EG34							2							3									2
EG25							3							5									5
EG8							2							3									4
EG43							2							2									6
EG37							0							0									0
EG33							3							3									3
% Agreement	62	59	52	31	45	24	45	86	59	59	34	62	28	55	90	59	66	31	62	34	57		

Figure 11. Visual map of agreements between the instructor and self and peer evaluations for the EG (part 1)

The Wilcoxon (non-parametric) test for related samples was applied, since normality could not be assumed (Kolmogorov-Smirnov test, at the 5% level of significance). Tables 1 and 2 summarize the significant

differences. In light of those results, the practical guidelines provided to the EG do not seem to have any effect on the results. The detailed instructions to model the parts provided for the second auto-evaluation do not seem useful either, since all student evaluations (first and second auto-evaluation, and peer evaluation) showed significant differences with instructor evaluation. Dimension 5 (Clear) is the only one that did not show any differences between evaluations. There was no improvement in the second self-evaluation with respect to the first one, which suggests that short repetitive exposure to rubrics does not increase reliability.

Table 1. Significant differences between evaluations for the CG (part 1)

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Average
Instructor and Self1	*			**		**	**
Instructor and Self2	*	*	*	**		**	**
Instructor and Peer		*	**	**		**	**

Note. ** $p = .01$; * $p = .05$.

Table 2. Significant differences between evaluations for the EG (part 1)

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Average
Instructor and Self1	**		*	**		**	**
Instructor and Self2	*	**		**		**	**
Instructor and Peer		**	**	**		**	**

Note. ** $p = .01$; * $p = .05$.

Different results were obtained for Part 2 EG and Part 2 CG. Students were explicitly asked to model the part so it would lend itself to specific design changes. This requirement of flexibility matches various sub-levels of Dimension 6 (Design Intent) in the rubric. As a result of these specific conditions, differences arise in the evaluation of Dimension 6 between the CG and the EG (Table 3). Figures 12 and 13 visually summarize the detailed quantitative results.

Part 2	Self1 to Instructor $\Delta = 0.25$							Self2 to Instructor $\Delta = 0.25$							Peer to Instructor $\Delta = 0.25$							
	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	
CG34							5							5								4
CG21							3							4								4
CG25							2							3								3
CG46							3							3								2
CG37							5							5								4
CG13							3							4								4
CG33							2							4								4
CG35							5							5								2
CG9							6							5								6
CG44							4							3								5
CG38							2							1								3
CG42							3							2								3
CG6							2							2								2
CG27							6							6								6
CG17							4							3								4
CG43							4							4								3
CG49							2							3								4
CG4							4							4								5
CG23							4							4								4
CG8							6							5								5
CG30							5							4								5
CG29							4							3								3
CG18							4							4								3
CG5							6							4								6
CG15							2							3								3
CG48							1							2								3
% Agreement	92	73	62	54	62	31	62	92	73	69	58	46	27	61	100	85	62	54	65	19	64	

Figure 12. Visual map of agreements between the instructor, self, and peer evaluations for the CG (part 2)

Part 2	First evaluation $\Delta = 0.25$							Second evaluation $\Delta = 0.25$							Peer evaluation $\Delta = 0.25$							
	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	
EG29							1							0								2
EG28							5							5								6
EG30							6							5								5
EG31							5							5								4
EG48							5							5								5
EG35							3							2								5
EG15							4							3								4
EG44							5							4								6
EG6							5							4								4
EG13							6							5								6
EG27							2							2								4
EG46							5							5								5
EG9							5							4								2
EG5							4							4								5
EG39							6							6								5
EG19							5							4								2
EG23							6							6								5
EG17							3							3								6
EG12							6							5								5
EG26							4							4								5
EG38							5							5								3
EG47							5							6								5
EG3							4							4								6
EG34							6							5								5
EG25							6							6								6
EG8							5							5								6
EG43							4							3								3
EG37							1							1								4
EG33							4							5								5
% Agreement	97	69	86	55	90	55	75	97	69	72	48	76	55	70	97	76	90	48	83	69	77	

Figure 13. Visual map of agreements between the instructor and self and peer evaluations for the EG (part 2)

While significant differences between instructor and student remained consistent throughout ongoing evaluations for CG, they decreased and eventually disappeared from the EG. We hypothesize that the experience gained in successive evaluations appears to improve student assessment only if the practical guidelines are combined with explicit metrics: the background of the EG (the explanations on how to self-evaluate exercises with rubrics) is boosted by specific checking requirements added to the exercise so that students can clearly identify with the criteria.

Table 3. Significant differences between evaluations of the CG and the EG for dimension 6 (part 2)

	Dim 6 (Design Intent) CG	Dim 6 (Design Intent) EG
Instructor and Self1	**	**
Instructor and Self2	**	*
Instructor and Peer	**	

Note. ** $p = .01$; * $p = .05$.

The spreadsheet-based prototype revealed the benefits of adaptable rubrics that provide varying levels of detail for each criterion and immediate feedback of the evaluation scores. Every subject was able to select the level of detail he/she required, while being blocked from incoherently marking rubrics when working at different levels. Unfortunately, an extensive use of the tool confirmed that spreadsheet forms are difficult to reconfigure and adapt; information from completed forms is difficult to extract and process, and spreadsheet forms do not fully prevent incomplete or inconsistent scoring. A more generic and flexible platform was required to maximize the benefits associated to the adaptable nature of the rubric.

Experiment 2

The second experiment confirms that the new platform does not affect the reliability of the evaluations, while providing richer and more complete information, as it outputs user interaction metadata that can be used to analyze the evaluation process and guide further improvements.

Our sample included one group of junior industrial engineering students from a Spanish university. All 47 students had basic knowledge of rubrics (MCAD quality concepts were explained using rubrics, but students were never required to practice).

First, students were asked to create a solid model of a specific part (a fixed arm of a pistol clamp) depicted in a detailed drawing (see Figure 14). Next, they were asked to create their own detailed drawing of the fixed arm they previously modeled. Finally, participants were required to assemble a virtual model of the pistol clamp (see Figure 15). All parts were provided, with the exception of the standard parts (available in the library) and the fixed arm (previously modeled by each subject).

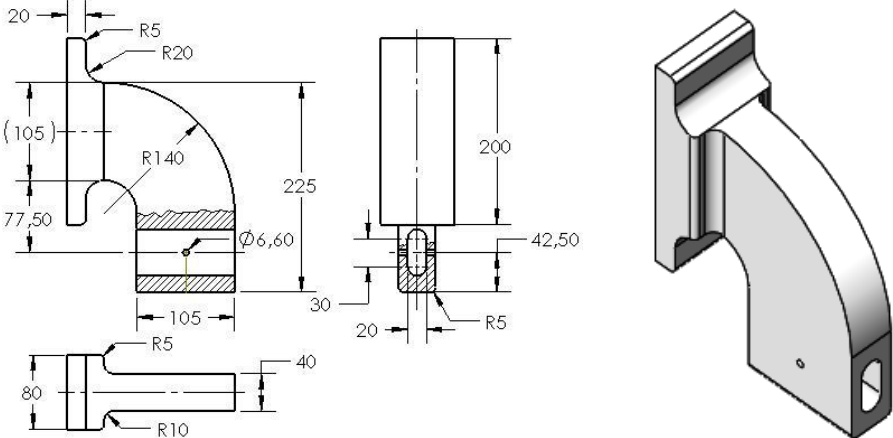


Figure 14. Input (left) and output (right) of the part used in experiment 2

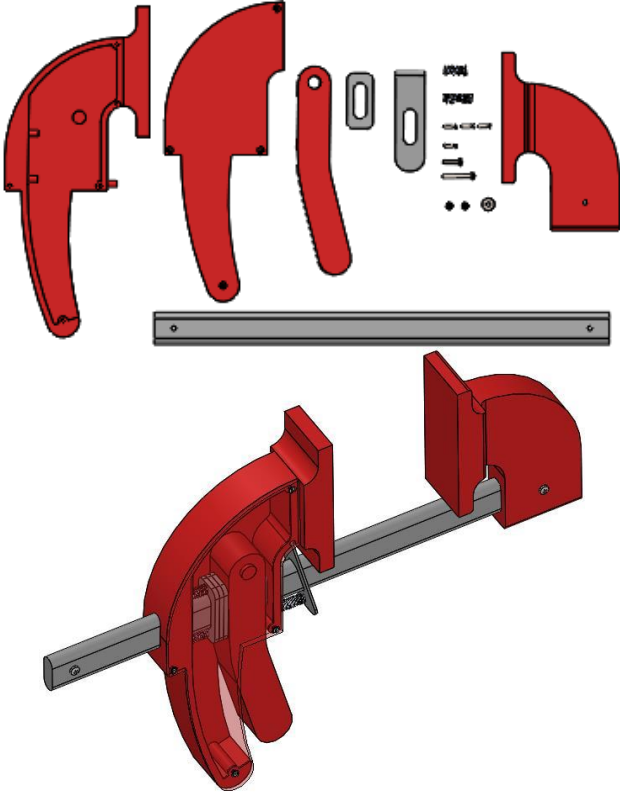


Figure 15. Input (left) and output (right) of the pistol clamp used in experiment 2

Figure 16 visually summarizes the detailed quantitative results of the experiments, as they show the differences between teacher and self-evaluations.

Subject	MODEL							DRAWING							ASSEMBLY							
	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	Valid	Complete	Consistent	Concise	Clear	Des. Intent	Agreements	
1							2							1								3
2							2							2								4
3							3							6								4
4							4							5								3
5							3							0								3
6							2							2								5
7							3							1								6
8							2							3								4
9							3							1								2
10							2							2								4
11							3							1								3
12							4							2								3
13							5							4								6
14							5							6								0
15							3							2								2
16							2							2								2
17							4							2								3
18							2							4								4
19							4							4								3
20							3							3								5
21							2							2								4
22							3							4								3
23							4							4								4
24							4							2								2
25							3							2								1
26							1							3								3
27							1							2								2
28							4							5								4
29							5							6								4
30							1							1								5
31							1							1								3
32							2							1								4
33							1							5								3
34							3							3								4
35							2							2								3
36							3							3								2
37							2							2								5
38							3							3								4
39							2							2								5
40							2							2								2
41							4							3								4
42							3							3								3
43							4							5								2
44							3							4								2
45							3							4								2
46							2							2								5
47							4							1								0
% Agreement	91	68	51	32	32	9	47	77	68	43	51	21	17	46	77	83	40	40	57	30	55	

Figure 16. Visual map of agreements between instructor and self-evaluations for the three tasks (modeling, drawing and assembling)

Comparison between visual maps of the Experiments 1 and 2 clearly show that students behave quite similarly with independence on whether they evaluate with spreadsheets or with the new platform; the percentages of agreements are similar for both experiments. Thus, we validate the hypothesis that the new platform does not negatively affect the results.

In order to find significant differences between instructor and students evaluations, a T-test for related samples was applied for the average ratings (since the Kolmogorov-Smirnov test revealed normality could be assumed at the 5% level of significance) and the Wilcoxon (non-parametric) test for related samples was applied for Dimensions 1 to 6 (normal distribution could not be assumed). Results are shown in Table 4.

Some collateral lessons learned resulted from Experiment 2. Information bubbles were incorporated to Annota to contextually display textual explanations for each level of deployment (see Figure 6). In addition, the new platform collects information about how users interact with the platform while completing the rubrics. In

particular, the platform tracks how long each detailed explanation was visible, so we can identify students who looked at the bubbles longer than average, and those who looked at them a shorter time.

Table 4. Significant differences levels between self-evaluations and instructor evaluation for experiment 2

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Average
Modeling task		*			**	**	**
Drawing task			**		**	**	**
Assembling task			**	**		**	**

Note. ** $p = .01$; * $p = .05$.

For all tasks (modeling, drawing, and assembly), students who looked at the bubbles longer than average showed less significant differences between their evaluations and the instructor's, as shown in Tables 5 and 6. Therefore, using bubbles to provide explanations seems to be a valid anchor to illustrate the various levels of attainment.

Table 5. Significant differences between self-evaluations and instructor evaluation for experiment 2, only for students who looked at the bubbles longer than average

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Average
Modeling task					**	**	**
Drawing task			**		**	**	**
Assembling task			**	**		**	**

Note. ** $p = .01$; * $p = .05$.

Table 6. Significant differences between self-evaluations and instructor evaluation for experiment 2, only for students who looked at the bubbles shorter than average

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Average
Modeling task		*		*	*	**	*
Drawing task				*	**	**	
Assembling task		*	*			*	*

Note. ** $p = .01$; * $p = .05$.

The ability of the platform to collect metrics based on user interaction helps to improve the teaching strategy. For example, by examining metadata about the use of bubbles for the modeling rubric (Table 7), we can see that criterion 6 displayed the highest number of bubbles, which can be related to the fact that criterion 6 is more difficult to understand.

Significant differences between student and instructor evaluations were found in all cases (Tables 4 to 6). Maximum and average times significantly increased for both drawing and assembly tasks (Table 7), which can be attributed to the fact that students were less exposed to rubrics of drawings and assemblies than they were to rubrics of part models. Under these circumstances, and based on Orsmond et al. (1996), we presume that the concepts of drawing and assembly quality are still so alien that many students need more time to read the explanations from the bubbles. More illustrative anchors may still be required.

Table 7. Time (seconds) each detailed explanation was displayed in experiment 2

	Valid	Complete	Consistent	Concise	Clear	Design Intent	Total
Modeling task							
Max time (seconds)	38.1	6.958	42.79	9.891	15.57	49.86	139.504
Average time (seconds)	4.573	1.609	3.497	2.114	2.292	2.906	14.032
Number of users of bubbles	35	39	39	39	38	45	47
Drawing task							
Max time (seconds)	14.320	27.672	148.351	44.921	13.195	14.910	161.092
Average time (seconds)	3.173	2.1	5.689	2.96	2.44	1.836	15.036
Number of users of bubbles	30	43	42	39	36	43	47
Assembling task							
Max time (seconds)	33.565	139.501	46.807	61.097	19.988	30.920	285.859
Average time (seconds)	3.967	6.355	3.006	4.865	2.937	3.003	22.549
Number of users of bubbles	40	46	44	43	44	46	47

Our Annota platform overcomes a number of disadvantages commonly found in spreadsheet forms. For example, it prevents duplicate evaluations of the same criterion at multiple levels, and provides for easily obtainable metadata. Additional improvements have been gradually incorporated, such as detailed explanations for each

level of deployment that can be shown during the evaluation process, if required. Our e-rubric platform can also be helpful to test the effectiveness of modeling guidelines and other materials, particularly if we are able to use the metadata provided by the platform to detect the materials that better reduce the inter-rater agreement.

Based on our results, we can confirm that the metadata captured by our rubrics platform provides valuable metrics (such as the interaction between the user and the anchors). The hypotheses that was postulated as a result of the metadata analysis will be explored in future experiments. If the hypotheses are true, further work will be required to more accurately correlate such complex information and obtain guidelines to improve the anchors used to recognize quality criteria.

Conclusions and future developments

In this paper, we have tested our concept of adaptable rubrics, and have described the design, implementation, and testing of a web-based adaptable rubrics system aimed at formative e-assessment. It has been proven valid, as demonstrated in a case study on MCAD training that encompasses multiple and heterogeneous criteria to assess, and exhibited a moderately good correlation between the self-assessment of the students, the peer-assessment, and the assessment provided by the instructors. The main novelty of the approach is its adaptability, as it allows users to dynamically access multiple levels of detail for each quality criterion. In our implementation, buttons that allow users to show or hide details, as needed, control the level of detail. In addition, the system outputs metadata, which is helpful to analyze the evaluation process.

Comparing Experiment 1 (where students used spreadsheet forms) and Experiment 2 (where the new adaptable rubrics system was used), we conclude that the new form does not negatively affect the behavior of students, as inter-rater evaluations are quite similar for both experiments, and the small differences seem to have other plausible causes than the difference in the form. Besides, we can speculate that our new platform supports adaptable rubrics that work on some multidimensional rubrics with non-trivial evaluation criteria, where the available literature states that other rubrics do not work accurately.

Some collateral lessons gleaned from the experiments include: (1) specific instruction on the quality concepts evaluated by the rubrics appears to be useful only if the evaluated concepts are directly linked to quantitative metrics; (2) complementary information automatically displayed during the interaction with the rubrics platform seems to have a beneficial effect, and (3) the capability to collect metadata looks promising, but still requires further development to improve the educational process.

Our platform lays the foundation for a future intelligent tutoring system based on adaptable rubrics, which is hoped to easily become also adaptive by establishing a set of interconnected tasks—linked to different rubrics—to automatically activate subsequent tasks that can better assist the student in future learning stages. This process is performed after the previous stage has been self-evaluated and the platform has analyzed the level of performance attained by the student.

In the short-term, we are designing a more comprehensive analysis of the metadata captured by the rubrics platform. This analysis will provide guidelines to improve the anchors and help students better understand rubric's criteria. In the long-term, we plan to use our system to merge our current MCAD teaching notes and materials (which have already been successfully tested separately) with self-evaluation rubrics in order to develop an online MCAD course.

Acknowledgements

This work was supported by the Spanish Ministry of Economy and Competitiveness and the European Regional Development Fund, through the ANNOTA project (Ref. TIN2013-46036-C3-1-R). The authors also wish to thank the editor and reviewers for their valuable comments and suggestions that helped us improve the quality of the paper.

References

Andrade, H. (1996). Understanding rubrics. *Educational Leadership*, 54(4), 14–17.

- Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership*, 57(5), 13–19.
- Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation*, 10(3). Retrieved from <http://pareonline.net/getvn.asp?v=10&n=3>
- Anglin, L., Anglin, K., Schuman, P. L., & Kaliski, J. A. (2008). Improving the efficiency and effectiveness of grading through the use of computer-assisted grading rubrics. *Decision Sciences Journal of Innovative Education*, 6(1), 51-73.
- Atkinson, D., & Lim, S. L. (2013). Improving assessment processes in higher education: Student and teacher perceptions of the effectiveness of a rubric embedded in a LMS. *Australasian Journal of Educational Technology*, 29(5), 651-666.
- Auvinen, T., Karavritra, V., & Ahoniemi, T. (2009). Rubryc - An Online assessment tool for effortless authoring of personalized feedback. In *Proceedings of the Annual Conference on Innovation and Technology in Computer Science Education (ITiCSE)* (pp. 377). doi:10.1145/1562877.1563019
- Cebrián-Robles, D., Serrano-Angulo, J., & Cebrián-de-la-Serna, M. (2014). Federated eRubric service to facilitate self-regulated learning in the European university model. *European Educational Research Journal*, 13(5), 575-584.
- Company, P., Contero, M., Otey, J., & Plumed, R. (2015). Approach for developing coordinated rubrics to convey quality criteria in CAD training. *Computer-Aided Design*, 63, 101-117.
- Company, P., Otey, J., Contero, M., Agost M. J., & Armiñana A. (2016). Implementation of adaptable rubrics for CAD model quality formative assessment. *International Journal of Engineering Education*, 32(2A), 749-761.
- Creighton, R. H. (2010). *Unity 3D game development by example: A Seat-of-your-pants manual for building fun, groovy little games quickly*. Birmingham, UK: Packt Publishing Ltd.
- Economides, A. A., & Roupas, C. (2007). Evaluation of computer adaptive testing systems. *International Journal of Web-Based Learning and Teaching Technologies*, 2(1), 70-87.
- Georgiadou, E., Triantafyllou, E., & Economides, A. A. (2006). Evaluation parameters for computer-adaptive testing. *British Journal of Educational Technology*, 37(2), 261–278.
- Isbell, T., & Goomas, D. T. (2014). Computer-Assisted rubric evaluation: Enhancing outcomes and assessment quality. *Community College Journal of Research and Practice*, 38, 1193–1197.
- Jonsson, A., & Svingby, G. (2007). The Use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2, 130-144.
- Karkehabadi, S. (2013). *Using rubrics to measure and enhance student performance*. Springfield, VA: Office of Institutional Research, Planning, and Assessment Northern Virginia Community College.
- Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items. *Prospects and Obstacles, Educational Measurement: Issues and Practice*, 33(2), 19-28.
- Manson, J. R., & Olsen, R. J. (2010). Diagnostics and rubrics for assessing learning across the computational science curriculum. *Journal of computational Science 1*, 55-61.
- McCormick, M., Lawyer, K., Wiggins, J., Swan, C., Paterson, K., & Bielefeldt, A. (2015). Sustainable engineering assessment using rubric-based analysis of challenge question responses. *Journal of professional issues in engineering education and practice*, 141(2). doi:10.1061/(ASCE)EI.1943-5541.0000211
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation*, 7(25), 1-9. Retrieved from <http://pareonline.net/getvn.asp?v=7&n=25>
- Montgomery, K. (2002). Authentic tasks and rubrics: Going beyond traditional assessments in college teaching. *College Teaching* 50(1), 34–40.
- Moskal, B. (2000). Scoring Rubrics: What, When and How? *Practical Assessment, Research & Evaluation*, 7(3). Retrieved from <http://pareonline.net/getvn.asp?v=7&n=3>
- Nitko, A. J. (2001). *Educational assessment of students* (3rd ed.). Upper Saddle River, NJ: Merrill.
- Orsmond, P., Merry, S., & Reiling, K. (1996). The Importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, 21(3), 239–250.
- Panadero, E., & Jonsson, A. (2013). The Use of scoring rubrics for formative assessment purposes revisited: A Review. *Educational Research Review*, 9, 129–144.
- Popham, W. J. (1997). What's wrong and what's right with rubrics? *Educational Leadership*, 55(2), 72-75.
- Reddy, Y.M., & Andrade, H. (2010). A Review of rubric use in higher education. *Assessment & Evaluation in Higher Education*, 35(4), 435-448.

Santos, P., Colina X., Hernández-Leo, D., Melero, J., & Blat, J. (2009). Enhancing computer assisted assessment using rubrics in a QTI editor. In *Proceedings of the 9th IEEE International Conference on Advanced Learning Technologies (ICALT)* (pp. 303-305). doi:10.1109/ICALT.2009.92

Smit, R., & Birri, T. (2014). Assuring the quality of Standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation*, 43, 5-13.

Taminiau, E. M., Kester, L., Corbalan, G., Spector, J. M., Kirschner, P. A., & Van Merriënboer, J. J. G. (2015). Designing on-demand education for simultaneous development of domain-specific and self-directed learning skills. *Journal of Computer Assisted Learning*, 31, 405-421.

Tierney, R., & Simon, M. (2004). What's still wrong with rubrics: Focusing on the consistency of performance criteria across scale levels. *Practical Assessment, Research & Evaluation*, 9(2). Retrieved from <http://pareonline.net/getvn.asp?v=9&n=2>

Zaiontz, C. (2015). *Real statistics using Excel*. Retrieved from <http://www.real-statistics.com/>