

CAD Training for Digital Product Quality: A Formative Approach with Computer-Based Adaptable Resources for Self-Assessment

ABSTRACT

As the engineering and manufacturing sectors transform their processes into those of a digital enterprise, future designers and engineers must be trained to guarantee the quality of the digital models that are created and consumed throughout the product's lifecycle. Formative training approaches, particularly those based on online rubrics, have been proven highly effective for improving CAD modeling practices and the quality of the corresponding outcomes. However, an effective use of formative rubrics to improve performance must consider two main factors: a proper understanding of the rubric and an accurate self-assessment. In this paper we develop these factors by proposing CAD training based on self-assessment through online formative rubrics enriched with adaptable resources. We analyzed self-assessment data, such as time spent, scoring differences between trainee and instructor or use of the adaptable resources, of fourteen different CAD exams. Results show that resources are more effective when used without any incentives. The comparison of assessments by quality criterion can facilitate the identification of issues that may remain unclear to trainees during the learning process. These results can guide the definition of new strategies for self-training processes and tools, which can contribute to the higher-quality outcomes and CAD practices that are required in model-based engineering environments.

Keywords: Model-Based Enterprise; CAD model; Enhanced quality; E-rubric; Adaptable resource; Formative self-assessment.

1. INTRODUCTION

As organizations continue their digital transformations toward Industry 4.0, workers will need relevant training to remain competitive; employers will need new mechanisms to assess employees' capabilities; and the incumbent workforce will need to demonstrate competency in a variety of new tools. At a fundamental level, these needs must be addressed by developing the technical literacy, skills and tools required to work with digital product data. The Model-Based Enterprise (MBE) paradigm involves the adoption of a model-based product definition as the unique source of truth to guide all the processes and information flows throughout the entire lifecycle of a product (Bertoline et al. 2019). In MBE environments, 2D drawings are replaced by 3D digital product models which are used as the basis for communication, documentation, and decision making to predict and validate product behavior and performance. The creation and manipulation of CAD models become critical tasks, as the quality of these models largely determines the effectiveness of the associated downstream processes and the ability to consume product data. The knowledge about the kind and quality of data of previous steps is to maintain the design intent (Xu and Galloway 2005), and about the following process steps is to guarantee their feasibility (Dankwort et al. 2004). In this context, quality of CAD models can be understood as a multidimensional construct (Contero et al. 2002) where morphology (geometrical and topological correctness), syntax (organizational aspects) and semantics (ability to alter and reuse) are the main axes.

To leverage the benefits of the digitization, automation, and connectivity of value chains, future designers and engineers must be trained to guarantee the quality of the digital models that are created and consumed throughout the product's lifecycle in an Industry 4.0 environment. Indeed, CAD education is largely considered one of the ten major challenges in computer-aided design (Piegl 2005). As organizations

transition to MBE practices, the development of individual and corporate CAD competence (not simply procedural, but strategic) will become even more indispensable than it already is (Hamade and Artail 2008; Hamade 2009; Ye et al. 2004; Bodein et al. 2013). From an academic perspective, the exploration of simplified and intuitive interpretation of technical contents, which has been called Education-driven-research (EDR), is a key issue (Rossignac 2004). An effective strategy to develop skills that are based on CAD quality consists in exposing future designers and engineers early to the concepts of CAD quality dimensions and educating them in the strategic use of the related techniques to ensure these criteria are met (Company et al. 2015). To this end, the use of rubrics has been proposed as a method to convey quality-oriented strategies to CAD trainees and prevent the Einstellung effect (Company et al. 2015; Luchins 1942).

An educational rubric is a scoring tool that includes criteria for rating dimensions of performance and standards of accomplishment for those criteria (Jonsson and Svingby 2007). Both content and processes are important in quality assessment (Parke 2001). Rubrics are assumed to enhance the consistency of scoring across students, assignments, and raters (Jonsson and Svingby 2007).

A well-designed rubric must mitigate inconsistencies in the scoring process by minimizing errors due to rater training, rater feedback and the clarity of descriptions of criteria (Reddy and Andrade 2010). These requirements are especially relevant in complex rubrics (i.e., those that assess multiple dimensions) such as those required for CAD training. Our concept of CAD quality criteria in digital product models aligns with the one described by (Company et al. 2015; 2017), which expands the morphology, syntax and semantics, to distinguish up to six dimensions: validity, completeness, consistency, conciseness, clarity and conveyance of design intent. Resources are provided to support the verification of the performance level that the trainee achieves in each quality criterion.

Due to their characteristics, rubrics are useful not only for evaluation purposes, but also for formative assessment (Jonsson and Svingby 2007; Popham 1997; Andrade and Du 2005; Reddy and Andrade 2010), which makes them ideal for educational settings where the instructor may not be present or readily available, such as asynchronous, massive, and/or online courses. Reliance on self-directed online training is increasing, as it has proved to have a positive influence in students' motivation and learning achievement (Nikou and Economides 2016). With the need to promote a quality-oriented behaviour on future engineers and designers, we anticipate that accurate self-assessment mechanisms and autonomous evaluation tools (Morris 2019) will play a key role in their development and technical training.

We assume that the efficient use of formative rubrics for CAD quality depends on two premises: a) the student must understand the rubric, and b) the student must be able to accurately score himself/herself. The first premise builds on the recent findings showing that students expect learning analytics features “to support their planning and organization of learning processes, provide self-assessments, deliver adaptive recommendations, and produce personalized analyses of their learning activities” (Schumacher and Ifenthaler 2018). In our view, the first premise is favored by using adaptable rubrics (Company et al. 2016), which can be adjusted on demand to support different CAD learning styles and rhythms, as described by (Hamade 2009). Furthermore, many electronic platforms for adaptable rubrics can manage metadata from the assessment process, which provides a rich data source for learning analytics (Company et al. 2017). The second premise is to attain a precise score. Studies show that many trainees cannot identify their weaknesses due to inaccurate self-assessments. Surprisingly, there seems to be little information in the scientific literature on the effectiveness of rubrics for self-assessment (Jonsson and Svingby 2007).

Our approach involves enhancing rubrics with complementary resources to facilitate efficient formative use based on the premises established previously (understanding of the rubric and proper scoring). To this end, indicators of rubric use (especially for complex ones) and resources for self-assessment must be analyzed to determine their effectiveness. These indicators can provide new insights on the impact of the

rubrics in different contexts and for different types of users. For example, if the trainee-instructor inter-reliability can be improved, rubrics can be a reliable support tool in large courses or distance learning scenarios.

The main goal of this work is to evaluate the impact of these adaptable resources for their refinement in order to achieve more accurate self-assessments. Ultimately, the enhancement of formative rubrics can help CAD trainees to gain the advanced modeling skills demanded by model-based engineering processes.

In this paper, we use complex rubrics as self-assessment tools, and also as an assessment tool for the instructor, with the aim of making the trainees more aware of quality aspects that have not been correctly assimilated. Alternatively, the study analyses aspects to be improved so that the rubrics can be used in autodidactic, massive or online courses, by trainees interested in receiving feedback about their actual level of progress. The integration of adaptable resources into complex rubrics is examined to improve the self-assessment process and ensure that all users understand the criteria the same way. First, we review application tools for rubrics to achieve quality results based on the assumptions described previously and introduce a set of adaptable resources that are applicable to rubrics. Methods to detect ineffective self-assessment are also reviewed. In section 2, we summarize the relevant state of the art and define the main hypotheses of our study. Section 3 describes a series of experiments to validate our hypotheses. In these experiments, a group of engineering students self-assessed their performance using complex formative rubrics. The results of these experiments are reported in section 4 and discussed in section 5. Finally, conclusions are reviewed in section 6.

2. REVIEW OF TOOLS FOR QUALITY IMPROVEMENT THROUGH SELF-ASSESSMENT

In this section, we review some adaptable resources and tools for improving self-assessment processes in formative rubrics. We describe the adaptable resources that will be studied in our work and some methods to prevent ineffective self-assessment.

2.1. Adaptable Resources for Complex Rubrics in Formative Contexts

Effective formative rubrics should include resources to support trainees in achieving quality results. Rubrics are typically embedded in Learning Management Systems (LMS) to generate reports and provide feedback to instructors, such as percentages of the maximum intended performance and similar statistical measures. Specialized tools have been developed to improve the formative assessment of more complex skills, such as the use of video-modeling examples (Ackermans et al. 2019), and to support advanced types of formative rubrics for complex subjects such as CAD (Company et al. 2016). In the latter approach, the authors included different quality dimensions (Company et al. 2017) where sets of related criteria were combined based on their purpose. An important feature in these rubrics is the ‘unfold’ tool, shown in Fig. 1. Dimensions can be ‘folded’ (or collapsed) and ‘unfolded’ (or expanded) to reveal different levels of detail and adapt to the user’s learning rhythm. This feature involves the use of a resource to achieve the first premise established in the previous section: *the student must understand the rubric*.

The platform also provides annotations or ‘dynamic bubbles’ that inform users about the performance levels while interacting with the rubric (Company et al. 2017). These annotations can be useful for preventing heterogeneous evaluations and simplifying self-assessments, as users receive immediate feedback that explains the scoring. This resource can be particularly valuable to address our second premise: *the student must be able to accurately score himself/herself*.

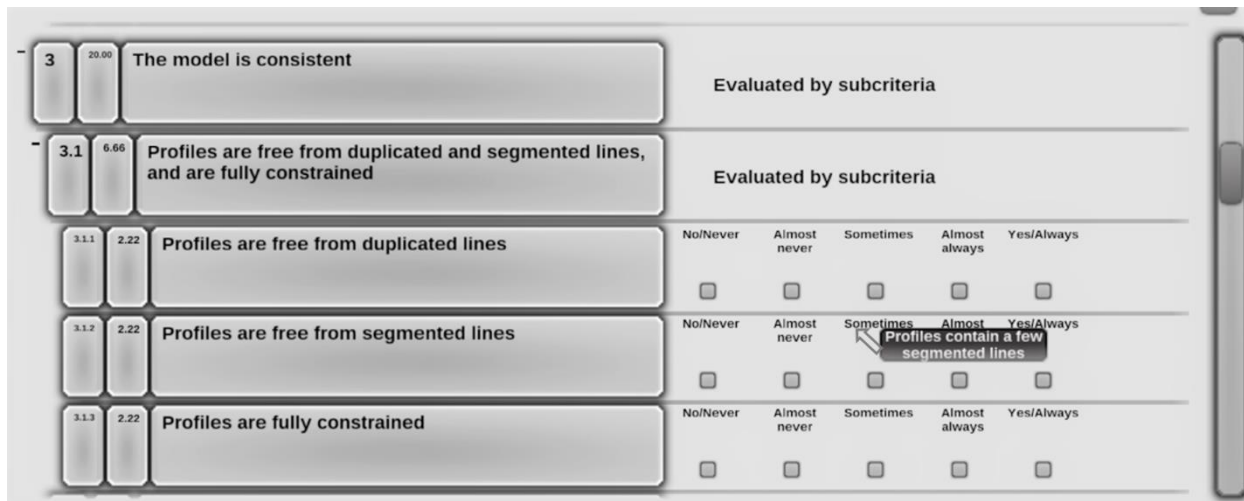


Fig.1 Example of ‘fold/unfold’ (left) and dynamic bubbles (right). Partial view from a rubric implemented in the ANNOTA platform (Company et al. 2017)

2.2. Methods for Detecting Ineffective Self-Assessment

The self-assessment of knowledge and abilities has been extensively studied (Sundström 2005), particularly in the higher education domain (Sundström 2005; Boud and Falchikov 1989; Falchikov and Boud 1989, Falchikov and Goldfinch 2000). In these studies, authors identified a tendency toward overly positive self-evaluations (Domínguez et al. 2016), which may be attributed to a reflection of private aspirations and personal desires (Willard and Gramzow 2009) or to a desire to create a good impression in others. It has been observed that people tend to think of themselves in a positive way. Positive biases have been identified in different types of self-reports, including physical attributes, socially valued behaviors and performance on tests of intelligence, athleticism and driving (Willard and Gramzow 2009). Students also tend to exaggerate their academic performance, so biased self-assessment with rubrics deserves research attention.

Many of the classic methods to detect deception are focused on conscious deception. Evidence suggests that poor self-assessment is common (Evans et al. 2005; Baxter and Norman 2011) and mostly focused toward lie-detection or detection of cheating (Reinhard et al. 2011; McManus et al. 2005). However, this work focuses on unintentional deception. The unconscious tendency of a person to see herself in a favorable light is known as *self-deception* (Paulhus and John 1998). Self-deception has been associated with self-adaptation and competence, unconsciously promoting the positive qualities (self-deceptive enhancement) and denying the negative ones (self-deceptive denial). Unconscious self-deception is associated with personality traits such as extraversion, emotional stability, or intellect (Dodaj 2012) and related to the concept of goal projection that leads towards better performance (Willard and Gramzow 2009). Occasionally, flawed self-assessments are unconsciously due to an information deficiency (Dunning et al. 2004).

Although traditional methods for detecting deception do not fit our purposes, some features are discussed, as they can be adapted to our study. First, we consider physiological or cognitive processing and neurosciences measurements, such as the observation of parameters in facial expressions of emotion (Ekman and O’Sullivan 2006). These techniques are based on the idea that emotional stimuli can cause certain physical reactions. If deceiving requires a higher cognitive effort than telling the truth, then subtle physiological changes might be detected (e.g., eye movements, the bioelectric activity of certain areas of the brain, etc.). However, reactions can also be caused by other factors, such as nervousness, joy or fear.

Therefore, these techniques are not applicable in the context of our study, where unconscious irregular behaviors can be caused by an excess of confidence or reluctance to complete the task.

The idea that lying involves more complex cognitive activities than being honest is also applied in lie-detection techniques based on the response speed. It is estimated that the act of conceiving a lie requires approximately 30% more time than telling the truth (Gregg 2007; Agosta and Sartori 2013; Walczyk et al. 2009). Some authors contend that longer response times are not only associated to false responses, but also to social desirability (Holtgraves 2004). In our study, the time patterns of users who fill out a rubric with made-up responses is unknown. It is possible that they invest little time because they do not understand the question and moved on, or because they are making an extra effort to understand it. In this regard, the times used for self-assessment could be analyzed and connected to other metadata in order to identify patterns of behavior.

Scales of socially desirable responding have also been used as an indicator of deliberate response distortion on personality questionnaires (Dodaj 2012). Even when people are aware of the facts, they may be motivated to alter them due to the desire to create or maintain a positive public impression. This tendency to deliberately convey a favorable impression is called *impression management* and may contribute to the inaccurate self-reporting of performance (Evans et al. 2005). This kind of response distortion also occurs in situations of anonymous responding (Paulhus 1984), which suggests that there might be a part of unintentional distortion in self-presentation that involves personality traits.

An assessment rubric taps into social desirability, as it exposes the level of knowledge of a particular subject. However, not all participants show the same level of self-deception and personality traits. It is expected that subjects showing higher levels on certain traits such as optimism or self-confidence will assess themselves higher. Researchers (Willard and Gramzow 2009) indicate that the tendency to exaggerate academic performance varies depending on the context. Costs and benefits of positive biases in self-evaluation depend on their underlying motives. Exaggeration in a private context may reflect an adaptive tendency to project positive objectives in self-reports. In a more public context, exaggeration is associated to social desirability and not to positive affection.

In our view, the conclusions discussed in the previous paragraph are interesting and relevant to our goals: since self-assessment is typically performed in a private context, if positive biases are observed they must be due to positive affection and not exaggeration. However, self-assessment could also occur in a public context, for example when rewards are offered for alignment of scores between the student and the instructor (i.e., the instructor reviews the trainee's self-assessment). Therefore, the following aspects must be analyzed:

- Students' personality traits (e.g. optimism) and the positive biases in their self-assessments based on the assessment of the instructor.
- The time spent performing the self-assessment and its possible connection to other process parameters.
- The potential influence of the context and the differences between private and public self-assessments.
- The use of adaptable resources.

2.3. Hypotheses

Based on the adaptable resources incorporated into the rubrics and the reviewed techniques for detecting ineffective self-assessment, we defined the following hypotheses:

- H1. Improved self-assessment mechanisms can close the gap between trainees' self-assessments (which tend to be higher) and instructors' assessments.

- H2. The time spent in self-assessment activities is correlated with the inter-reliability level between trainees and instructors.
- H3. Knowing that the instructor will review the self-assessments can influence the trainees' responses.
- H4. The use of adaptable resources in rubrics can improve the accuracy of the self-assessment.
- H5. Separating quality dimensions into categories in a rubric can help to identify specific aspects of CAD quality that are not well understood.

3. MATERIALS AND METHODS

3.1. Experimental Design

To validate our hypotheses, we conducted a study at a Spanish university with undergraduate students majoring in Mechanical Engineering and Industrial Engineering Technology and enrolled in a required junior-level Engineering Graphics course. All students took a Technical Graphics Communication course during their freshmen year. Students were introduced to CAD quality dimensions with the support of complex rubrics. Specifically, we used the MCAD rubrics implemented in the ANNOTA platform (Company et al 2017), which offered resources for criteria deployment and annotation tools ('unfold' and 'dynamic bubbles'). The assertions for dimensions and sub-dimensions of quality for these rubrics are described in (Company et al. 2015; 2016).

Students were asked to self-assess their work after the midterm and final exams. The rubrics were made available online for forty-eight hours starting from the moment the exams were submitted. Over every course of one semester, students were instructed on the creation of 3D CAD models and assemblies (Company and Gonzalez-Lluch, 2013). The assessed tasks corresponded to modeling and assembly tasks similar to those used during instruction, as shown in Fig. 2. Students were encouraged to inspect a solved model (also available online) before self-assessing their own work. However, all rubrics had to be completed in one sitting. A total of 14 different exams (14 experiments) with a mean number of valid students' self-assessments (N, valid sample size) were considered, as shown in Table 1.

3.2. Data Collection and Analysis

All data was reviewed after each experiment to verify validity. Only two cases of self-assessment were discarded due to an excessive total time (exceeding three days) in exams 5 and 6. Excessively short times were not discarded, as they represent an undesirable but actual behavior (students who filled out the rubric without putting a significant effort into the self-assessment). Cases in which the instructor assigned zero points to a model or when the file was invalid were also discarded.

The trainee – instructor inter-reliability was studied using the *TIS* factor. The use of the resources was measured as follows: for the unfold tool, the percentage of students who unfolded criteria at least once was recorded; for the bubbles, the average number of times that they were queried (in any level of performance) per student was considered. In terms of frequency of use, the variable *Unfold_num* indicates the number of times that the unfold tool was used during a self-assessment. *Bubbles_X_num* indicates the number of times that bubbles were shown for the level of performance X during a self-assessment. The parameters and metadata collected and analyzed from the rubrics are shown in Table 2.

Variable	Description
<i>Time</i>	Total time spent in the self-assessment.
<i>TIS (Trainee-Instructor Scoring)</i>	Difference between the total score from the self-assessment and the score from the instructor assessment.
<i>Unfold_use</i>	Percentage of students using the unfold tool for each exam.
<i>Bubbles_use</i>	Average number of times that bubbles were used (for all performance levels) for each exam.
<i>Unfold_num</i>	Number of times that the unfold tool was used during a self-assessment.
<i>Bubbles_X_num</i>	Number of times that bubbles for the level of performance X are shown during a self-assessment (for X=1 to 5)

Table 2. Variables used in the study.

Descriptive statistics for the variable *TIS* were analyzed. Bivariate correlations between *Time* and *TIS* were applied to determine whether the time devoted to self-assessment may be indicative of a more accurate evaluation.

The analysis on the use of adaptable resources was based on the variables *Unfold_use* and *Bubbles_use* as well as the bivariate correlations between the frequency of use of the resources (*Unfold_num* and *Bubbles_X_num*) with parameters such as *Time* and *TIS*. The Spearman's coefficient was applied, as variables could not be considered normally distributed. Significant correlations were considered when $p < 0.05$.

To study differences in the means and the distribution of *TIS*, depending on the quality criterion under consideration, ANOVAs (with Bonferroni correction in the *post hoc* when the Levene test showed critical levels > 0.05 ; otherwise Games-Howell) and Kruskal Wallis tests were applied, respectively. Kruskal Wallis tests were also applied for *Unfold_num* and *Bubbles_total_num*, depending on the criterion under consideration, as *Unfold_num* and *Bubbles_total_num* could not be considered normally distributed.

4. RESULTS

Descriptive statistics are shown in Table 3. The mean of the variable *TIS* (*Mean TIS*) was positive in all 14 experiments, which indicates that the final scores obtained in the self-assessments were, on average, higher than the scores from the instructor’s assessments.

In 11 of the 14 experiments, the percentage of cases in which the value of *TIS* was positive (% *TIS*>0) is greater than the percentage of cases in which it was negative (% *TIS*<0). The mean of the positive values of *TIS* is 2.13, and -0.83 for the negative values. Therefore, there were not only many more cases in which students self-assessed their work higher than the instructor did, but also the ratings were higher. The mean of the percentage of cases with null values of *TIS* (coincidence between self-assessment and instructor’s assessment) was 1.41. In addition, the mean of the difference of scores was greater when students were optimistic with respect to the cases in which self-assessments were lower than the instructor’s assessments.

Exam	N	Mean TIS	<i>TIS</i> >0		<i>TIS</i> <0		<i>TIS</i> =0	
			%	Mean	%	Mean	%	Mean
1	27	2.25	85.19	2.81	14.81	-0.97	0	0
2	25	2.11	88.00	2.43	12.00	-0.19	0	0
3	48	1.17	72.92	1.80	22.92	-0.61	4.16	0
4	45	1.86	80.00	2.55	20.00	-0.92	0	0
5	50	0.14	48.00	1.44	52.00	-1.05	0	0
6	50	0.53	46.00	2.16	52.00	-0.89	2	0
7	31	1.15	83.87	1.54	16.13	-0.90	0	0
8	25	3.08	96.00	3.21	4.00	-0.04	0	0
9	20	1.15	75.00	2.02	20.00	-1.88	5	0
10	29	2.15	79.31	2.95	20.69	-0.91	0	0
11	27	0.59	62.96	1.48	37.04	-0.93	0	0
12	36	0.54	61.11	1.51	38.89	-0.98	0	0
13	35	1.95	77.14	2.65	20.00	-0.45	2.86	0
14	35	0.12	45.72	1.26	48.57	-0.93	5.71	0
Mean		1.34	71.52	2.13	27.08	-0.83	1.41	0

Table 3. Descriptive statistics of the variable *TIS*. N: number of students (valid sample size).

The possible relationships of the variable *Time* with other metadata could also be examined, similarly to previous studies in lie-detection (Gregg 2007; Agosta and Sartori 2013; Walczyk et al. 2009). However, these studies do not exactly align with the focus of this work.

We cannot conclude that the variables *Time* and *TIS* were correlated, since statistical significance was only found in 3 of the 14 cases studied ($r(25) = -0.585$, $p = 0.001$; $r(23) = -0.606$, $p = 0.001$; $r(23) = -0.561$, $p = 0.004$). In these three cases, the relationship is negative, i.e., the more time invested in the self-assessment, the smaller the difference is between student and instructor scores.

A significant positive correlation was found between the time spent in self-assessment (expressed by the variable *Time*) and the frequency of use of the resources. Table 4 shows bivariate correlations (Spearman’s coefficient) between *Time* and *Unfold_num* and between *Time* and *Bubbles X_num*. In all cases, a significant correlation between the time spent with the frequency of use of the unfold tool was found ($p < 0.001$ in 12 of the 14 cases). For the relationship between *Time* and the use of bubbles, statistical significance was also found in all cases, at least for some levels of performance. It should be noted that the explanatory text in the bubbles only varies between the levels of performance in the level where the criterion is obtained. Knowing the gradation of achievement in five levels, it is indifferent to

the level for which the bubbles are consulted. The only cases with no statistical significance are shown in grey in Table 4.

Exam	Time - Unfold_num		Time - bubbles 1_num		Time - Bubbles2_num		Time - Bubbles3_num		Time - Bubbles4_num		Time - Bubbles5_num	
	Coef	p	Coef	p	Coef	p	Coef	p	Coef	p	Coef	p
1	0.740	0.000	0.732	0.000	0.762	0.000	0.718	0.000	0.711	0.000	0.658	0.000
2	0.689	0.000	0.740	0.000	0.789	0.000	0.688	0.000	0.687	0.000	0.528	0.006
3	0.552	0.000	0.574	0.000	0.617	0.000	0.594	0.000	0.569	0.000	0.552	0.000
4	0.659	0.000	0.492	0.001	0.442	0.003	0.524	0.000	0.681	0.000	0.470	0.001
5	0.696	0.000	0.581	0.000	0.463	0.001	0.469	0.001	0.489	0.000	0.682	0.000
6	0.693	0.000	0.500	0.000	0.527	0.000	0.509	0.000	0.592	0.000	0.650	0.000
7	0.627	0.000	0.700	0.000	0.626	0.000	0.664	0.000	0.662	0.000	0.688	0.000
8	0.537	0.006	0.671	0.000	0.728	0.000	0.680	0.000	0.761	0.000	0.656	0.000
9	0.558	0.011	0.436	0.055	0.441	0.052	0.717	0.000	0.542	0.014	0.416	0.680
10	0.714	0.000	0.663	0.000	0.377	0.044	0.554	0.002	0.489	0.007	0.528	0.003
11	0.460	0.016	0.485	0.010	0.452	0.018	0.351	0.073	0.578	0.002	0.656	0.000
12	.677	0.000	0.725	0.000	0.729	0.000	0.644	0.000	0.616	0.000	0.468	0.004
13	0.654	0.000	0.534	0.001	0.514	0.002	0.366	0.031	0.504	0.002	0.500	0.002
14	0.582	0.000	0.519	0.001	0.504	0.002	0.425	0.011	0.428	0.010	0.377	0.026

Table 4. Bivariate correlations between *Time* and *Unfold_num*, and between *Time* and *Bubbles X_num* (X=1-5). Coef.: Spearman’s coefficient. (Significance critical value: $p < 0.05$).

These results are not surprising, as using the resources is time-consuming. However, the use of such tools responds to an interest: users trying to improve their performance. Otherwise, the results would show a large number of “uses” (the number of times that the criteria were deployed, or the bubbles were visualized), that would not correspond to significantly higher times.

Yet, these adaptable resources are not widely used (see Table 5). The percentage of students that used ‘unfold’ in each experiment (*Unfold_use*) ranges from 34.83% to 52%. With respect to the average number of times that bubbles were used in the performance levels in each assessment (*Bubbles_use*), the minimum value is 13.1 and the maximum, 35.52.

We observed that resources were used more often when inter-reliability was rewarded. For the unfold tool, this was true in all cases except for exams 11 and 12. While in exams 3 to 8 there is variety in the subject, course and type of exam, exams 11 and 12 correspond to the first (partial) exam in the same class (subject B, Course 2), where less emphasis was put on those resources. This could justify the lower scores obtained for those cases.

The average percentage of use (*Unfold_use*) in cases where inter-reliability was rewarded was 40.85%, dropping to 31.33% when not rewarded. In the case of the bubbles, the increased use of the resources when rewarded was always true. The average number of times bubbles were used in cases when there was a reward was 30.67. When no reward was offered, the average only reached 18.03 times. In addition, a significant positive relationship was found between the use of the two resources combined (at least in some performance levels in the case of the use of bubbles) in 13 of the 14 experiments conducted. The only cases with no statistical significance are shown in grey in Table 5.

Exam	Inter reliability	N	Unfold _use (%)	Bubbles _use	Unfold_num	Unfold_num	Unfold_num	Unfold_num	Unfold_num
					- Bubs_1_num	- Bubs_2_num	- Bubs_3_num	- Bubs_4_num	- Bubs_5_num

	rewarded				Coef	p	Coef	p	Coef	p	Coef	p	Coef	p
1	NO	27	33.3	22.9	0.690	0.000	0.648	0.000	0.586	0.001	0.545	0.003	0.731	0.000
2	NO	26	30.8	19.2	0.689	0.000	0.676	0.000	0.672	0.000	0.544	0.000	0.494	0.010
3	YES	48	35.4	35.5	0.479	0.001	0.483	0.001	0.400	0.005	0.507	0.000	0.461	0.001
4	YES	45	42.0	30.7	0.644	0.000	0.625	0.000	0.605	0.000	0.373	0.012	0.557	0.000
5	YES	50	44.0	28.6	0.646	0.000	0.446	0.000	0.430	0.002	0.417	0.003	0.720	0.000
6	YES	50	48.0	30.7	0.599	0.000	0.449	0.001	0.462	0.001	0.501	0.000	0.710	0.000
7	YES	31	45.2	32.8	0.507	0.004	0.389	0.030	0.433	0.015	0.528	0.002	0.766	0.000
8	YES	25	52.0	28.6	0.643	0.001	0.634	0.001	0.719	0.000	0.531	0.006	0.398	0.049
9	NO	20	30.0	13.1	0.541	0.014	0.696	0.001	0.764	0.000	0.499	0.025	0.221	0.349
10	NO	29	31.0	14.6	0.561	0.002	0.601	0.001	0.796	0.000	0.397	0.033	0.388	0.038
11	YES	27	29.6	28.4	0.270	0.173	0.141	0.484	0.238	0.232	0.203	0.310	0.317	0.107
12	YES	36	30.6	30.1	0.690	0.000	0.632	0.000	0.599	0.000	0.581	0.000	0.619	0.000
13	NO	35	31.4	17.9	0.712	0.000	0.526	0.001	0.243	0.159	0.282	0.100	0.605	0.000
14	NO	35	31.4	20.5	0.649	0.000	0.647	0.000	0.238	0.168	0.315	0.065	0.362	0.033

Table 5. Use of the adaptable resources and correlations between frequency of use. Coef.: Spearman's coefficient. N: number of students (valid sample size). (Significance critical value: $p < 0.05$).

For the cases with significant correlation between the use of these tools and *TIS*, the relationship was negative, i.e. a greater use of the tools corresponds with higher levels of inter-reliability (smaller scoring differences). However, this correlation was only present in 2 of the 14 cases for the unfold tool, and in 5 cases for the use of the bubbles.

A greater number of cases of statistical significance were detected in the experiments in which the inter-reliability was not rewarded. For the unfold tool, the two cases with significant correlations had no reward. More relevant results were found for the use of bubbles. A significant relationship with *TIS* was found (the more use of bubbles, the lower the scoring difference) in 67% of cases (4 of 6) where alignment with instructor scoring was not rewarded. Otherwise, there were significant differences only in 12.5% of cases. Table 6 shows only the statistically significant cases, for clarity.

Exam	Inter reliability rewarded	N	TIS-Unfold_num		TIS-Bubbles_1_num		TIS-Bubbles_2_num		TIS-Bubbles_3_num		TIS-Bubbles_4_num		TIS-Bubbles_5_num	
			Coef	p	Coef	p	Coef	p	Coef	p	Coef	p	Coef	p
1	NO	27	-0.587	0.001	-	0.000	-	0.000	-	0.001	-	0.015		
2	NO	25	-0.438	0.029	-	0.000	-	0.000	-	0.001	-	0.000		
3	YES	48												
4	YES	45												
5	YES	50												
6	YES	50												
7	YES	31					-	0.012	-	0.005	-	0.011		
8	YES	25												
9	NO	20												
10	NO	29									0.438	0.017	0.380	0.042
11	YES	27												
12	YES	36												
13	NO	35				-	0.043	-	0.032					
14	NO	35												

Table 6. Bivariate correlations between *TIS* and *Unfold_num*, and between *TIS* and *Bubbles X_num* (X=1-5). Coef.: Spearman’s coefficient. N: number of students (valid sample size). (Significance critical value: $p < 0.05$).

Although used only by a fraction of the students, adaptable resources may be useful to improve inter-reliability (*H4. The use of adaptable resources in rubrics can improve the accuracy of the self-assessment*). Statistical significance with *TIS* was not demonstrated for all cases, but when such significance exists, it was always connected to improving inter-reliability. The resources were more useful when a reward from the instructor was not expected.

The previous results can help improve the understanding of the self-assessment process. However, to gain a better understanding of the quality dimensions defined in the rubric, it is important to consider each criterion separately. Our comparative study for *TIS* separately for each criterion helped identify the criteria that were more confusing to students. In modeling exams, criteria 4 and 6 showed the most differences between the scores assigned by the student and the instructor. For assemblies, criteria 6 and (to a lesser extent) 3 stood out over the others. The significant differences in mean and distribution between criteria are shown in Table 7.

In general, students seemed to have more difficulties with the criteria during intermediate exams. In the final exams, the number of statistically significant differences decreased (especially in the modeling tests). We assume that the students were able to address these difficulties during the intermediate exams before the final exam. However, the fact that the differences between criteria in the final exams decreased does not necessarily mean that the difference between assessment scores (*TIS*) also did. In fact, for the cases shown in Fig. 3, for example, the distribution for criterion 5 increased for the final exam, which confirms the existence of additional factors that may influence the *TIS*.

Finally, no significant differences were found in the use of the unfold and bubbles tools based on the evaluated criterion. Although greater differences between student and instructor scores were found for some of the criteria, this does not translate into an increased use of adaptable resources.

Exam	Type	Significant differences in TIS between criteria
1	Partial exam (Assembly)	6-1**; 6-2**, 6-3**, 6-4**, 6-5**
2	Final exam (Assembly)	6-1*; 6-2**, 5-3**, 6-4**, 6-5**
3	Partial exam (Modelling)	6-1**, 6-3**, 4-1**, 4-2**, 4-3**, 4-5**, 4-6*, 2-3**, 1-2**
4	Partial exam (Assembly)	1-2**, 1-3*, 1-5*, 1-6**, 2-4*, 2-6**, 3-6**, 4-6**, 5-6**
5	Final exam (Modelling)	6-1**, 6-2**, 6-3**, 6-4**, 6-5**
6	Final exam (Assembly)	6-1*; 6-2**, 6-4**, 6-5**, 3-5**
7	Partial exam (Modelling)	6-1**, 6-2**, 6-3**, 6-5**, 4-1**, 4-2**, 4-3**, 4-5**
8	Partial exam (Assembly)	6-1**, 6-2**, 6-4**, 6-5**, 3-1**, 3-2**, 3-4**, 3-5**
9	Final exam (Modelling)	6-1**, 6-2**, 6-4**, 2-4**, 3-4**, 5-4**
10	Final exam (Assembly)	5-1**, 5-2*, 6-1**
11	Partial exam (Modelling)	1-3*, 5-1*, 5-2*, 2-3*, 4-3**, 5-3**, 6-3**
12	Partial exam (Assembly)	1-5*, 6-1**, 6-2**, 6-3**, 6-4**, 6-5**
13	Final exam (Modelling)	2-1**, 6-1** 6-3**, 6-4**
14	Final exam (Assembly)	1-5*, 2-5*, 6-1**, 6-2**, 6-3**, 6-4**, 6-5**

Table 7. Criteria for which significant differences in mean and distribution of their *TIS* variable have been detected. (Significance critical value: ** $p < 0.01$; * $p < 0.05$).

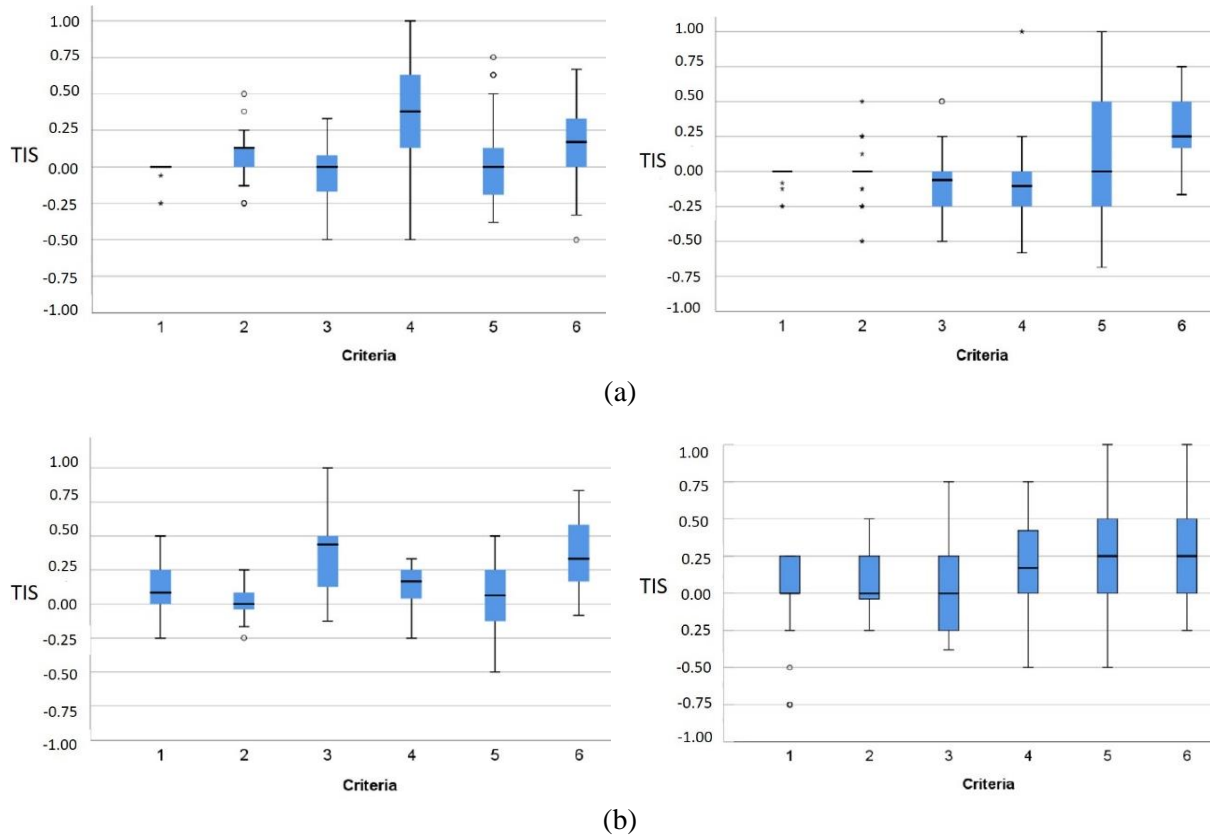


Fig.3 Box plots for *TIS* in each criterion. Comparison between partial (left) and final exams (right). (a) Modeling exercises (exams 3 and 5) and (b) Assembly exercises (exams 8 and 10)

5. DISCUSSION

As engineering organizations transition toward Industry 4.0 environments, the role of the digital product representation will become more relevant. Digital product models will be the authoritative source for product definition and the mechanism around which all business processes will revolve throughout the entire product's lifecycle. In these scenarios, ensuring the quality of all digital product data, which begins with the native CAD file, is crucial.

A key strategy to enforce high CAD quality practices in an organization is to incorporate quality assurance mechanisms, which should include automated methods to assess, repair, and optimize models that do not meet a predefined standard of quality, and the development of effective educational tools and training strategies, both at the academic and corporate levels, that leverage the technical literacy, skills and tools required to work with digital product data and emphasize the principles of CAD-quality and reusability. The introduction of CAD quality criteria can be enabled by formative rubrics.

Rubrics have traditionally been used to homogenize and accelerate evaluation processes, but their purpose can go much further. Formative rubrics combined with suitable resources and properly implemented in accessible online platforms can be used by students to self-assess their progress and determine possible weaknesses in their training. In this regard, a critical goal of formative rubrics is to inform and motivate.

In our view, two factors need to be considered in order to use formative rubrics effectively and improve

performance: understanding of the rubric, and accurate self-assessment. In this paper, we described a set of experiments designed to study the general tendency of CAD trainees when they self-assess using formative rubrics, and to determine the impact of some adaptable resources when used in combination with rubrics. From the review of the adaptable resources incorporated into the rubrics and the techniques for detecting ineffective self-assessment, we stated five hypotheses focused on closing the gap between trainees' and instructors' assessments and enhancing the accuracy of the self-assessment.

Our results confirmed the tendency toward overly positive self-assessment reported in previous studies (Myers and Ridl 1979). These results partially validated our first hypothesis (*H1. Improved self-assessment mechanisms can close the gap between trainees' self-assessments (which tend to be higher) and instructors' assessments*) and justified the need to improve self-assessment processes and provide adequate resources to increase accuracy.

We found no conclusive evidence of a statistically significant relationship between the time spent in self-assessment and the inter-reliability level (difference between trainee's and instructor's scoring). However, in the cases where this relationship is significant, it is a negative one (the higher the time spent, the lower the difference in scores), which means that investing more time in self-assessment tasks translates into more accurate results. However, although there seems to be a trend (based on our limited number of cases) toward obtaining better evaluations when investing more time, it is not possible to confirm H2 from these results (*H2. The time spent in self-assessment activities is correlated with the inter-reliability level between trainees and instructors*). Further studies are needed to provide a more conclusive answer.

A significant relationship was found between the time devoted to self-assessment tasks and the use of adaptable resources, as well as between the use of the two resources. Although using resources is time-consuming, the results reflect the users' intentions to produce accurate self-assessments. The increase in time may indicate that the information being displayed was carefully read. This seems to imply that further research on adaptable resources for rubrics-based self-evaluation tools is necessary but hardly sufficient condition for complex tasks like CAD training.

Although the use of adaptable resources was not prevalent (only 30-50% of users unfolded at least one criterion), it increased when inter-reliability was rewarded. This result supports the idea of social desirability and context influence (Willard and Gramzow 2009) and validates hypothesis H3 (*H3. Knowing that the instructor will review the self-assessments can influence the trainees' responses*). However, a high percentage in the use of adaptable resources did not always translate into high levels of inter-reliability; adaptable resources could help obtain more accurate self-assessments (*TIS* factor) only when resources were used without the expectation of additional rewards. Therefore, different results were obtained for H4 (*H4. The use of adaptable resources in rubrics can improve the accuracy of the self-assessment*) based on the use of rewards. It is recommended that rewards are applied only during the early stages of training so trainees can get familiar with the resources and appreciate their value but should be removed as soon as the use of resources is established.

The results from the comparisons of *TIS* values between quality criteria revealed that discrepancies between trainees' and instructors' assessments were more evident in certain criteria, which seems to fit with those that are more objective and dichotomic (a model is clearly perceived as valid or invalid), while differences seem to be easier in more subjective and even interrelated criteria (a model sometimes conveys better design intent at the cost of becoming less concise). Therefore, this subject remains open for future investigations. As expected, these differences are more frequent while instruction is in progress, which may mean that the student's understanding of the subject improved as the course progressed. These results validate hypothesis H5 (*H5. Separating quality dimensions into categories in a rubric can help identify specific aspects of CAD quality that are not well understood*). The inter-reliability analysis between criteria is recommended so the least understood aspects of the material can be identified and addressed early.

6. CONCLUSIONS

In this paper, we evaluated the impact of adaptable resources on self-assessment strategies for CAD model quality with complex rubrics. Our analysis of metrics from the rubrics provided valuable information for instructors to facilitate the development of CAD quality-oriented training. Various relationships were identified under specific conditions, although some were not fully validated, such as the relationship between the time spent in self-assessment and inter-reliability, or between the use of adaptable resources and inter-reliability. Further studies are required to provide a more conclusive answer and to determine the applicability of formative rubrics to other fields. Additional work is also needed to study other aspects that may influence self-assessment accuracy and reliability. For example, methods to detect inefficient or fraudulent assessments published in the literature have not been shown to be effective when applied to complex assessments, as the quality assessment of CAD models.

Our results can inform the development and use of resources in combination with formative rubrics to help CAD trainees acquire advanced modeling skills, particularly in self-paced and online learning scenarios. It has been shown that offering rewards for high levels of inter-reliability leads to a higher use of adaptive resources, which can be beneficial in the early stages of training. However, the strategy is not recommended in more advanced stages of the learning process, as the level of inter-reliability increases when resources are used without offering a reward. Our findings are useful for our long-term goal, as we are interested in developing a complete characterization of the factors involved in CAD quality-oriented training, so new educational strategies should be designed to provide trainees with the CAD competences that will be necessary to effectively operate in MBE environments.

References

- Ackermans, K., Rusman, E., Brand-Gruwell, S., Specht, M. (2019). Solving instructional design dilemmas to develop a Video Enhanced Rubric with modeling examples to support mental model development of complex skills: the Viewbrics-project use case. *Educational Technology Research and Development*, 67:983–1002.
- Agosta, S., Sartori, G. (2013). The autobiographical IAT: a review. *Frontiers in Psychology*, 41:519.
- Andrade, H., Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research and Evaluation*, 10(3):1-11.
- Baxter, P., Norman, G. (2011). Self-assessment or self deception? A lack of association between nursing students' self-assessment and performance. *Journal of Advanced Nursing*, 67(11):2406-2413.
- Bertoline, G.R., Hartman, N.W., Ross, W.A. (2019). *Fundamentals of Solid Modeling and Graphic Communication*. New York: McGraw Hill Education. ISBN: 978-1-260-08480-1.
- Bodein, Y., Rose, B., Caillaud, E. (2013). A roadmap for parametric CAD efficiency in the automotive industry. *Computer-Aided Design*, 45(10):1198-1214. doi:10.1016/j.cad.2013.05.006
- Boud, D., Falchikov, N. (1989). Quantitative Studies Of Student Self-Assessment In Higher-Education - A Critical Analysis Of Findings. *Higher Education*, 18(5):529-549
- Company, P., Contero, M., Otey, J., Plumed, R. (2015). Approach for developing coordinated rubrics to convey quality criteria in MCAD training. *Computer-Aided Design*, 63:101-117.
- Company, P., Contero, M., Otey, J., Camba, J.D., Agost, M.J., Pérez-López, D.C. (2017). Web-based system for adaptable rubrics: case study on CAD assessment. *Journal of Educational Technology & Society*, 20(3):24-41
- Company, P., González Lluch, C. (2013). *CAD 3D con SolidWorks® Tomo I: Diseño básico*, Publicacions Universitat Jaume I; <https://doi.org/10.6035/Sapientia86>.

- Company, P., Otey, J., Contero, M., Agost, M.J., Almiñana, A. (2016). Implementation of adaptable rubrics for CAD model quality formative assessment. *International Journal of Engineering Education*, 32(2A):749-761.
- Contero, M., Company, P., Vila, C., Aleixos, N. (2002) Product data quality and collaborative engineering. *IEEE Computer Graphics and Applications*, 22(3):32-42.
- Dankwort, C.W., Weidlich, R., Guenther, B., Blaurock, J.E. (2004) Engineers' CAx education - it's not only CAD. *Computer-Aided Design*, 36(14):1439-1450. doi:10.1016/j.cad.2004.02.011
- Dodaj, A. (2012). Social desirability and self-reports: testing a content and response-style model of socially desirable responding. *Europe's Journal of Psychology*, 8(4):651-686.
- Domínguez, C., Jaime, A., Sanchez, A., Blanco, J.M., Heras, J. (2016). A comparative analysis of the consistency and difference among online self-, peer-, external- and instructor-assessments: The competitive effect. *Computers in Human Behavior*, 60;112-120.
- Dunning, D., Heath, C., Suls, J.M. (2004). Flawed Self-Assessment: Implications for Health, Education, and the Workplace. *Psychological science in the public interest: a journal of the American Psychological Society*, 5(3):69-106.
- Ekman, P., O'Sullivan, M. (2006). From Flawed Self-Assessment to Blatant Whoppers: The utility of voluntary and involuntary behavior in detecting deception. *Behavioral Sciences & the Law*, 24:673-686.
- Evans, A.W., Leeson, R.M.A., Newton John, T.R.O., Petrie, A. (2005). The influence of self-deception and impression management upon self-assessment in oral surgery. *British Dental Journal*, 198:765-769.
- Falchikov, N., Boud, D. (1989). Student Self-Assessment In Higher-Education - A Meta-Analysis. *Review of Educational Research*, 59(4):395-430.
- Falchikov, N., Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research*, 70(3) :287-322.
- Gregg, A.P. (2007) When vying reveals lying: The Timed Antagonistic Response Alethiometer. *Applied Cognitive Psychology*, 21(5): 621-647.
- Hamade, R.F. (2009). Profiling the desirable CAD trainee: technical background, personality attributes, and learning preferences. *Journal of Mechanical Design*, 131(12):121009.
- Hamade, R.F., Artail, H.A. (2008). A study of the influence of technical attributes of beginner CAD users on their performance. *Computer-Aided Design*, 40(2):262-272. doi:10.1016/j.cad.2007.11.001
- Holtgraves, T. (2004). Social desirability and self-reports: Testing models of socially desirable responding. *Personality and Social Psychology Bulletin*, 30(2):161-172.
- Jonsson, A., Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, 2:130-144.
- Luchins, A.S. (1942). Mechanization in problem solving: The effect of Einstellung. *Psychological monographs*, 54(6): 1-95, ISSN 0096-9753
- McManus, I.C., Lissauer, T., Williams, S.E. (2005). *Learning in practice. Detecting cheating in written medical examinations by statistical analysis of similarity of answers: pilot study*. *BMJ*. 330:1064
- Morris, J.P. (2019). The Necessity of Autonomous Evaluation of Parametric Modeling and Drafting Instruction. In: *Proceedings of the ASEE Annual Conference & Exposition*, Tampa, Florida. <https://peer.asee.org/33409>
- Myers, D.G., Ridl, J. (1979). Can we all be better than average. *Psychology Today*, 13(3):89.
- Nikou, S.A., Economides, AA. (2016). The impact of paper-based, computer-based and mobile-based self-assessment on students's cience motivation and achievement. *Computers in Human Behavior*, 55:1241-1248
- Parke, C.S. (2001). An approach that examines sources of misfit to improve performance assessment items and rubrics. *Educational Assessment* , 7(3):201-25.

- Paulhus, D.L. (1984). 2-Component Models of Socially Desirable Responding. *Journal of Personality and Social Psychology*, 46(3):598-609
- Paulhus, D.L., John, O.P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, 66(6):1025-1060.
- Piegl, L.A. (2005). Ten challenges in computer-aided design. *Computer-Aided Design*, 37(4):461-470.
- Popham, W.J. (1997). What's wrong and what's right with rubrics?. *Educational Leadership*, 55(2):72-75.
- Reddy, M., Andrade, H. (2010). A review of rubrics use in higher education. *Assessment & Evaluation in Higher Education*, 35(4): 435-448.
- Reinhard, M.A., Dickhäuser, O., Marksteiner, T., Sporer, S.L. (2011). The case of Pinocchio: teachers' ability to detect deception. *Social Psychology of Education*, 14:299-318.
- Rossignac, J. (2004). Education-driven research in CAD. *Computer-Aided Design*, 36(14):1461-1469. doi:10.1016/j.cad.2003.10.008
- Schumacher, C., Ifenthaler, D. (2018). Features students really expect from learning analytics. *Computers in Human Behavior*, 78:397-407.
- Sundström, A. (2005). *Self-assessment of knowledge and abilities. A literature study*. Educational Measurement, 54. Umea University. ISSN: 1103-2685.
- Walczyk, J.J., Mahoney, K.T., Doverspiek, D., Griffith-Ross, D.A. (2009). Cognitive Lie Detection: Response Time and Consistency of Answers as Cues to Deception. *Journal of Business and Psychology*, 24(1):33-49
- Willard, G., Gramzow, R.H. (2009). Beyond oversights, lies, and pies in the sky: exaggeration as goal projection. *Personality and Social Psychology Bulletin*, 35(4):477-792.
- Xu, W., Galloway, R. (2005). Using behavioral modeling technology to capture designer's intent. *Computers in Human Behavior*, 21:395-405.
- Ye, X., Peng, W., Chen, Z., Cai, Y. (2004). Today's students, tomorrow's engineers: An industrial perspective on CAD education. *Computer-Aided Design*, 36(14):1451-1460. doi:10.1016/j.cad.2003.11.006