Self-, peer-, and instructor-assessment from a Bloom's perspective

Abstract:

The education literature supports the belief that higher order skills are essential for the ability of current students to compete globally within the accounting profession. Among the existing skill assessment alternatives offered by virtual learning environments, self and peer assessments are implementable feasible options with low human and materials resources cost. Some inconsistency among self-, peer- and instructor-assessment studies may be due to lack of proper theoretical framework. The objective of this study is to compare the three assessment methods (peer-, self-, and instructor-assessment) using Bloom's Taxonomy of the Cognitive Domain. Through a sample of 98 undergraduate accounting students, enrolled in a research methodology course, we collected data from different tasks using different assessment methods. The five tasks were created using different cognitive levels from our theoretical framework. A free learning management system called Moodle was used to implement the framework. The initial findings show statistical alignment among the three assessment methods. When the further analysis is done, by grouping the tasks using Bloom's taxonomy, differences emerge when comparing lower and higher levels of Bloom. As a result, these findings may help to explain the divergences previous researches. These findings can help in new strategies for assessment as well as teaching and learning especially in courses with high numbers of online students. There is also potential for the development of higher order skills from the experience of peer assessments.

-

Keywords: Assessment, rubrics, peer-assesment, self assessment, Bloom taxonomy

1 Introduction

Assessment is a component of training and education which involves planning, discussion, consensus building, reflection, measurement, analysis and improvement based on a learning objective (Buzzetto-more & Alade, 2006). Assessment relevance emerges from the need to provide feedback information from the teaching required by students, school and society as part of teaching and learning process.

In the e-learning environment, the assessment process may include pre and post tests, diagnostic analysis, student tracking, using line, support project-based learning, and data aggregation and analysis(Buzzetto-more & Alade, 2006).

The assessment process requires the intensive use of human resources that makes it very expensive and long. As the growing number of students, the evaluation process becomes a critical element in the financial viability of the process. One of the new academic initiatives, with large volume of students, are the Massive Open Online Courses (MOOC) where there has been a great expectation of growth in the educational market (Gray, 2013). The assessment process is one of the the key of a successful business plan for MOOC.

The instructor assessment tends to become impracticable in courses with large numbers of students and therefore, there is a need to identify new ways of viable assessment with the same or higher quality.

Rubrics

To improve assessment quality, efforts have become highly time and resources consuming. For this reason, the use of assessment tools that can promote thinking and learning while students are assessed are gaining interest to be used in the Learning Management Systems (LMS). Among the instruments available, the adoption of rubrics proves to be the fastest and most efficient way to evaluate student work. When rubrics design are in accordance with the recommendations in the literature they may also be teaching tools that support student learning and the development of sophisticated thinking skills.

The rubrics are used to assess specific tasks. This valuation method is particularly suited to e-learning because it facilitates the experience of assessments for instructors and students. Instead of writing repetitive comments to students, emerges increased quality when the items incorporate these comments.

The adoption of rubrics can combine the professor assessment to peer- and self-evaluation. In addition, the benefits highlighted by the increase of critical thinking and a self-assessment approach among the students (Isaacson & Stacy, 2009). Thus, the rubrics are a standard applicable not only to the assessment made by the instructor, but also for self- and peer- assessment.

There are two main types of rubrics: evaluative and instructional (Andrade, 2000). Evaluative rubrics are typically used when a judgment of quality is required. They are developed by the evaluators to guide the analysis of the student's efforts in a predefined scheme that transforms subjectivity and involves a rehearsal for a more objective assessment (Isaacson & Stacy, 2009; Moskal, 2000). Instructional rubrics add

to the evaluative ones the purpose of giving feedback on the work in progress as well as give detailed assessments about the final product(Andrade, 2000).

Evaluative and instructional rubrics are operational modalities of summative and formative assessments (Andrade, 2000). To recover the concept the summative evaluation is one that summarizes evidence on one aspect towards a judgement, both self- and peer- assessment have no instructional character because they represent only the measurement of an ending process. Summative assessments design uses standards, targets and criteria. What turns them into formative assessments is the return of feedback to the student indicating the existence of a gap between the current level of work being evaluated and the desired pattern. From this point of view, there is no way to the evaluation be formative without having happened previously, a summative evaluation whether explicitly or not(Taras, 2005).

Feedback

Moreover, the feedback is the information about the gap useful to change the gap to some extent. Thus, since that a change is expected it is supposed that the feedback is provided in such a context that allows the reduction of the gap (Taras, 2005). Therefore, the given feedback requires means to reduce the discrepancy between what is produced and what is desired. The key premise is that students become able to develop the ability to evaluate their work and need to have access to a recognized high quality standard for comparison. This argument implies the development of higher order skills through authentic assessment experiences (Sadler-smith, 2007). Formative rubrics represents this need as they exposure the feedback as the assessment criterion. In this setting, a self-assessment lets the student to compare the product to a standard as part of the task. Hence learning by the evaluation process. As this same process is repeated to evaluate peers tasks, the process expands the experience to another round of assessment and learning on same level student's tasks.

Self-assessment

Sitzmann, Brown, & Bauer (2010) explain the self-assessment construct through the distinction between three approaches: knowledge, cognitive learning and emotional outcomes. The first concerns the students` assessments on their knowledge or increases in their knowledge in a given domain. The self-assessment in the dimension of cognitive learning refers to the information understanding what includes knowledge based both on facts and knowledge. The critical distinction between these two dimensions is the source from which comes the understanding of the students. The first refers to testimonials from their levels of knowledge while the second refers to self-assessments on exams and tasks as assigned by the professor. The third approach - emotional outcomes - includes three possible approaches: reactions, motivation and effectiveness. The reactions reflect student satisfaction with their educational experience(Sitzmann, Brown, Casper, Ely, & Zimmerman, 2008). Motivation refers to the degree to which students have struggled to apply the acquired knowledge. Finally, effectiveness is about students' confidence in their ability to perform tasks associated with the given training (Sitzmann et al., 2008)

When the students are responsible for making the assessment, their ability to meet these requirements should be sufficient to match the importance of evaluation (O'Toole, 2013). As a result of the novelty of this method, this can be a potential

problem because students doubt their own abilities to produce constructive feedback and appropriate suggestions (Bannister & Thorne, 1997 apud Chen, 2010). In fact, students tend to give more value to experts and instructors than themselves. (Chen, 2010) Most of the students agreed that receiving feedback from a instructor is more significant than from another student(Theising, Wu, & Heck Sheehan, 2014). See student work from the perspective of self-positioning as an evaluator can be a powerful learning experience for students since they are not experienced evaluators (O'Toole, 2013). They can build on this experience becoming better evaluators of the quality of student work and as a consequence of their own work. As a result responsibility can grow over time(O'Toole, 2013).

Peer-assessment

The experience of students evaluating each other is of educational value saving teachers time and increasing student learning (Sadler & Good, 2006). Most of the students perceived that the peer review offers a positive impact on self-confidence, improves learning behaviours, and help identify personal strengths and limitations (Theising et al., 2014). The dialogue between pairs is recommended for the development of critical thinking which may include the trial of contributions that each give the other (Bonk & Smith, 1998). The pair can act as an additional contribution to the development of skills (Ng, 2014). It was also observed that the involvement provided by peer review develops more cognitive target skills that are more difficult on a self assessment (Chang, Tseng, & Lou, 2012). The peer review improves the student's abilities to relate instructional objectives with assessment activities, to understand the criteria and procedures to identify the strengths and weaknesses of own performance of the students to improve their understanding and confidence in the subject at hand and improving future performance (Chen, 2010).

The peer-assessment can expose the scene of collaborative learning by offering a way to include suggestions for another student on how to do a better job. In other words, students need not only the course content to produce the work but also to meet the evaluation criteria. This knowledge is necessary to bring elements to offer a suggestion about the work under assessment independently on its own or third parties. In this sense, a social pressure to cooperate may result in better learning.

Previous studies about peer-assessment and self-assessment

A simple investigation of the results of self- ,peer-, compared to the instructor-assessments produced results showing that there is some consistency in varying degrees (Chang, Liang, & Chen, 2013; Chang et al., 2012; Chen, 2010; Jessica Napoles, 2008). However, we observed potential issues regarding anonymity and type of task. These two factors can induce the researcher to obtain inaccurate conclusions.

Anonymity

A survey conducted with students working on the same subject in class showed evidence of differences among self-, peer- and instructor- assessment rankings. It was found that self- and instructor-assessments are consistent with each other but the peer-assessment showed significant differences with the other two methods.(Chang et al., 2013, 2012) A later study tested a new configuration using the same web-based support and found that this new environment the three methods converged to the same results

(Chang et al., 2012). In these two studies, the peer assessments were not anonymous. This maybe an indication that the configuration or something else can influence the convergence of the results obtained.

The collaborative environment can raise questions about the assessment from colleagues specifically in terms of anonymity(Chen, 2010). In addition, students showed up divided on the impact of anonymity and friendship(Theising et al., 2014). Students may feel uncomfortable about criticizing the performance of each other(Chen, 2010). It was observed that even among different groups, students assessing their own group members and others, without restriction about who was the other group being assessed, the tendency to weigh the evaluation positively to their own group over the other is a fact (Sadler & Good, 2006). The anonymity configuration in online courses can set a context that affects positively or negatively the confidence in the assessment. At this particular context it was found that the peer blind assess outperformed the assessment tasks in writing and provided more critical feedback to their peers than students participating in peer identified assess.(Lu & Bol, 2007)

Type of tasks assessed

The analysis of the process shall be prejudiced because any comparison between the groups of tasks has an explanatory sense restricted to task rather than wider and generalized way would be if the grouping were based on a valid framework. Otherwise, the feedback will have utility in the teaching-learning process restricted to summative results.

This distinction between from high items or low cognitive skills observed in a study showed easier and better accuracy self-assessment on low-level items in contrast to difficulty in high-level issues (Sadler & Good, 2006).

Regardless of epistemological beliefs, collaborative argumentation promoted more constructive and interactive environment in questioning activities and helped to build higher quality arguments in the case studies than in collaborative summaries. This points consistent clues to build the hypothesis that the type of task can interfere with the quality of the collaborative environment during the peer-assessment. Therefore, the effects of corresponding tasks and epistemological beliefs varied depending on the types of learning outcomes as understanding is opposed to argumentation in the online peer-assessments(Cho, Lee, & Jonassen, 2011).

Analysis of the results of previous comparatives studies about self, peer- , and instructors' studies without considering peer anonymity and grouping of tasks may explain some of these divergent results in previous studies. A conceptual theoretical framework may help guiding the grouping of tasks.

Theoretical framework using Bloom Taxonomy

The original levels from Bloom' taxonomy of cognitive domain are knowledge, comprehension, application, analysis, synthesis and evaluation. The first ones are called lower level and grows to the last ones that are called highest levels. For each level are associated sample verbs that were used to classify the tasks. The knowledge level is associated to the verb 'label'(Huitt, 2011).

Since there is an apparent difference in the results associated with the task types comes the need to use a framework to provide a basis to analyse this issue. The understanding and reasoning approach refers to cognitive skills taxonomy proposed by Bloom that can be used as a framework to identify whether there are differences among the three assessment methods (self, peer and instructor) by issue and also considering the different cognitive dimensions of Bloom. The same framework can analyse the hypothesis that analyses an eventual significant distinction on peer-assessments by gender. The adoption of the taxonomy of cognitive abilities proposed by Bloom is not new to assessments analysis (Buzzetto-more & Alade, 2006). A study compared timing of issues in a college chemistry and concluded that in the admission exam is predominantly higher levels of skills issues. However, during the course mainly above 99% assessments focus on the first level in the taxonomy. (Karamustafaoğlu, Sevim, Karamustafaoğlu, & Çepni, 2003)

A variant of Bloom's taxonomy proposed by Anderson & Krathwohl (2001) focus on curriculum to determine the level of high-level skills involved in the questions asked to carry out the assessment tasks. This variant detected the presence of higher levels of issues in the first three years of a course (Collier-Reed, 2011).

The modified Bloom's taxonomy level has three levels: Level 1 with knowledge and recall of information, Level 2 with Comprehension and application and understanding and being able to interpret data and the highest level 3 with problem solving and use of knowledge and understanding in new circumstances (Palmer and Devitt, 2007).

Furthermore, the concrete requirement to characterize the formative assessment is the degree of engagement of students with the use of this return to apply learning in future work. Thus, the formative evaluation of the feedback can be used intentionally from conception to develop high-level skills as defined by Bloom.

Previous studies have shown that can occur extreme predominance of the use of questions focused on cognitive abilities of low level even when there are institutional recommendations for the use of assessments in cognitive skills of high level (Karamustafaoğlu et al., 2003).

The learning project can hold an increase of peer-assessment capacity and responsibility to include steps of activity, peer-review, assessment and reflection, learning to be a better advisor, more active and with further evaluation. Review activities may be included in course design specifically to provide more opportunities to develop assessment skills. This is a didactical approach of learning-action. The process can be managed more effectively where the technology can be used to add suggestions on the material evaluated in the assessment area with comments from evaluation points added by others and then reflecting on the comments to be added in the same area

Theoretical framework using Bloom and anonymous rubric

This set of needs suggests the use of rubrics because it improves the performance score especially if they are analytically specific to a topic and complemented by examples and training. Additionally, rubrics potentially promote and enhance learning. (Jonsson & Svingby, 2007)

Besides the aspect of anonymity that is possible through rubrics, there must be a clear criterion in the evaluation process to standardize the evaluation process. Therefore, it is necessary that the criteria are transparent, predictable and relevant to the curriculum and student needs in order to design a reliable system (O'Toole, 2013).

The used rubrics main characteristic is to contribute to the system transparency, reliability and accuracy is that they support the learning and development of skills and understanding. Additionally, they must be easy to use, explain clearly show the expectations of the teacher, and provide students with a more informative feedback about their strengths.(Andrade, 2000)

However, there are no indications on how to produce the rubrics in such a way to meet all these needs. The recommendations are that to make an instructional rubric the professor should look for models, criteria lists, build and deconstruct criteria and articulated levels of quality, create and review the items of design before using it. (Andrade, 2000). Moreover, the need to evaluate becomes less important than the desire to increase the learning process and to develop new capabilities.

The rubrics design depends on how much training or summative purposes are the evaluation objectives. The success of the system depends on the students: engagement in their learning, ability to act in context, ability to apply their resources in a different context of their learning one, ability to apply their skills in an unfamiliar context in future meta-training for develop their capabilities even further meta-ability to help others develop their skills. Well-designed assessment practices may include several or all of these aspects to promote good learning and reliable evaluation (O'Toole, 2013).

On the other hand, previous experience with the use of rubrics is extremely important because if students are simply presented the rubrics and asked to use voluntarily, they end up ignoring the rubrics.(Sadler & Good, 2006) Therefore, there is the need to pre-teach students about how to use the rubrics through a training in this tool (Jonsson & Svingby, 2007).

The adoption of a comprehensive framework facilitates the design of valid rubrics (Jonsson & Svingby, 2007). However, it can be stated that the rubrics should measure skills according to the dimension you want to measure and may be knowledge, understanding or higher level capabilities. Thus, we use in this study the levels of Bloom's taxonomy associated to the tasks developed by the students.

In this work, self assessment concerns the second approach to cognitive learning in which the object being measured is knowledge based on facts and knowledge tasks. This definition excludes the self perception of knowledge and emotional outcomes.

The objective of this study is to do a comparative analysis of the difference among instructor-assessment students self-assessment , and anonymous peer-assessment using Bloom's Taxonomy of the Cognitive Domain. Thus, we used a conceptual theoretical framework for categorizing the tasks according to the skills represented in this framework. The research questions are as follow:

1.1 Are there significant differences among the three assessment methods by task?

1.2 Are there significant among the three assessment methods by task and by gender?

2.1 Are there significant among the three assessment methods by Bloom's taxonomy level 1 and two?

2.2 Are there significant among the three assessment methods by Bloom's taxonomy level 1 and two by gender?

3.1 Are there significant the three assessment methods ?

3.2 Are there significant among the three assessment methods by gender?

2 Methods

2.1 Participants

The participants were 101 undergraduate students in accounting at a public university in Brazil. The research methodology course introduces students in the complex task of writing scientific articles and research report. From the sample, three students were excluded for not having submitted their assignment. The final sample was 98.

The genre has been tested in peer assessment methods which is randomly assigned by Learning Management System used (Moodle). As there was the occurrence of evaluators of the same gender interspersed with pairs of evaluators with different genres, the test was performed considering three possible evaluators for each of the papers submitted by students: male-male (MM), female-female (FF) and mixed (FM).

2.2 Learning Management System

The Learning Management System (Moodle) was used to assess the three assessment methods using the module called 'workshop'.

In this particular approach, students assess their activity compared to a criterion understood as correct and which is the same used by the instructor to evaluate the task. This gives the student the feedback on what is expected with the potential to promote or enhance learning (Jonsson & Svingby, 2007).

2.3 Experimental procedure

The study employed some assessment activities comprising a training session (Unit 1), formal task (Unit 2) and assessment task (Unit 3).

2.3.1 Unit 1: Training session

Week 1:

Students were trained on the use of rubrics in a simple task asking a question about sports that is very popular here in Brazil. They should answer who was a soccer player present in an international event 44 years ago, goals made, and from which team he belonged. This was a typical hands-on activity with the support of the instructor to enhance their motivation and confidence using this peer assessment system.

Due to the institutional needs and alignment to the ethical questions about researches, using human beings was exposed to the students that the final course grade

used performance measures and that the material produced was part of a survey in which there would be no individual identification. The effective use of their assessments as part of the final course grade was the argument encourage sincere and responsible assessment.

Week 2 -8:

Students had normal classes about all the sections of the academic articles present in the academic abstract.

2.3.2 Unit 2; The formal task

Week 9:

Each students received a copy of same academic article inside of the class and had to accomplish the following tasks present in Figure 1

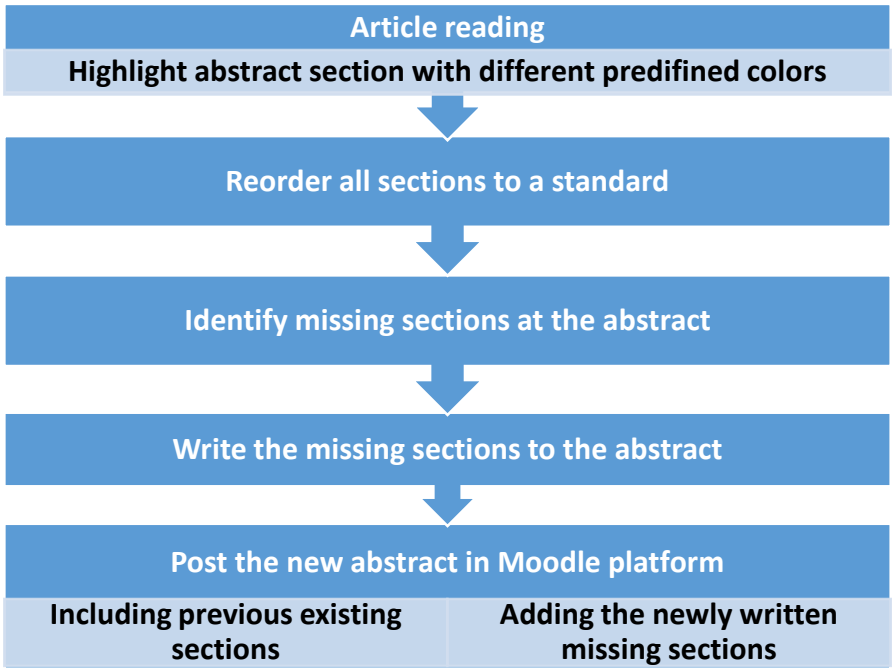

Figure 1 – In class activity schedule

In the class activity (Figure 1), the student acted in the first two cognitive levels of Bloom when he had the opportunity to show the ability to know and understand the basic concepts regarding the mandatory and optional sections of a scientific abstract. In this case, the student received an abstract of an article published in a scientific journal and a copy of the full article as well. His first task was to identify the abstract of the sections that are present. These sections are context, gap, objective, materials and methods, results/discussion and conclusion. The next task is to identify the section or missing sections and produce a phrase to complete the abstract of the six sections. The next task is to place the abstract on the standardized sequence as the abstract provided was in a different order.

At the end of the class, the system was shifted to an assessment phase where students assessed the submitted work by their peers.

2.3.3 Unit 3; The assessment task

Week 10:

After submitting the task online, the produced abstract were anonymously evaluated online by other two anonymous colleagues in addition to the evaluation of their own work in a double blind peer- and self-assessment configuration. The instructor performed the same evaluation for all students using the same rubric designed using the proposed theoretical framework within Moodle Learning Management System.

A survey carried to all students after they completed the task to verify the degree of protection of anonymity. Despite the warnings not to identify their own work with their personal information, four students left indications of their names in the works. It is likely that there is some kind of influence assessment for breach of anonymity. This can be a threat to validity and resulted in 4% of students who identified others' works. This value was considered acceptable and without implications for this research.

The tasks that make up the activity were associated with the skills in the Bloom cognitive domain as shown in Table 1:

Task	Domain
T1 – Identify abstract sections - highlight	Knowledge – Level 1
T2 – Reorder sections as a standard	
T3 – Show the ability to identify missing sections -	Comprehension – Level 2
T4 – Produce the first missing section (context) - summarize	
T5 – Produce the second missing section (gap) - summarize	

Table 1- Tasks and Bloom's domains

2.4 Development of assessment rubrics

The rubrics were created with reference to the literature using five tasks about abstract creation with value for each task ranging from one to five (Table 2). The idea is to assess if student has knowledge and comprehension about academic article abstract. The minimum score is 5 and the maximum is 22.

Rubrics header	Bloom's level	Rubrics options	Value
T1 - Correctly highlighted abstract existing sections as of template	1	Did not.	1
		Wrong three sections or more.	2
		Wrong two sections.	3
		Wrong one section.	4
		Correctly marked all sections	5
	1	Did not.	1

Rubrics header	Bloom's level	Rubrics options	Value
T2 - Reorder the abstract		Wrong three sections or more.	2
		Wrong two sections.	3
		Wrong one section.	4
		Correctly reordered all sections	5
T3 - Show the missing parts – context and gap	2	Did not.	1
		Identified more than 2 missing sections.	2
		Identified one missing section (not correct).	3
		Identified the 2 missing sections (correct).	4
T4 - Quality of the missing section produced - context	2	Not produced the text.	1
		Produced the text, but is not correct.	2
		Produced correctly, however the text is not clear	3
		The text is clear and correct.	4
T5 - Quality of the missing section produced - gap	2	Not produced the text (gap).	1
		Produced the text, but is not correct.	2
		Produced correctly, however the text is not clear	3
		The text is clear and correct.	4

Table 2 - Rubrics

3 Results

We collected several data regarding gender, the different tasks, and different tasks grouping options. Data was analysed comparing the three methods (instructor-, peer-, and self-assessment) matched and the influence of variable gender (factor) by the method called Analysis of Variance (ANOVA) Repeated Measures Model Mixt, when there was difference between the assessment methods, we used the Bonferroni post hoc test. We adopted the significance level $p \leq 0.05$.

Tasks and group of tasks	Bloom level	F(2,190)	p	Differences	Differences
T1	1	24.87	<0.001	Instructor-assessment is significantly lower than Peer-assessment and Self-assessment method (p <0.001 for both). post hoc Bonferroni	Peer-assessment and self-assessment do does not have significant difference (p = 0.92)
T2	1	25.12	<0.001	Instructor-assessment is significantly lower than Peer-assessment and self-assessment method (p <0.001 for both). post hoc Bonferroni	Peer-assessment and self-assessment does not have significant difference (p = 0.37).
T3	2	4.02	= 0.02	Peer-assessment is significantly lower than those of self –	Instructors-, peer-, and self-assessment

				assessment (p = 0.03). post hoc Bonferroni	have no significant differences.
T4	2	48.43	<0.001	Instructor-assessment is significantly lower than those of peer-assessment and self-assessment (p <0.001 for both). post hoc Bonferroni	Peer-assessment and self-assessment does not have significant difference (p = 0.46).
T5	2	11.40	<0.001	Instructor-assessment is significantly lower than those of the groups and peer self (p <0.001 for both)	Peer-assessment and self-assessment does not have significant difference (p = 0.41).
T1 + T2	1	32.86	<0.001	Instructor-assessment is significantly lower than peer-assessment and self-assessment (p <0.001 for both).	Peer-assessment and self-assessment do does not have significant difference (p = 0.40)
T3+T4+ T5	2	30.05	<0.001	Instructor-assessment is significantly lower than peer-assessment and self-assessment (p <0.001 for both)	Peer-assessment is significantly lower than self-assessment (p = 0.05).
T1+T2+ T3+T4+ T5	1+2	42.09	<0.001	Instructor-assessment is significantly lower than peer-assessment and self-assessment (p <0.001 for both)	The peer-assessment method has a strong tendency to have significantly lower than self-assessment method (p = 0.07).

Table 3 - Assessment methods

Tasks and group of tasks	Bloom level	Gender
T1	1	F (4, 190) = 0.44 p = 0.78.No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.
T2	1	F (4, 190) = 0.39 p = 0.82. No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender
T3	2	F (4, 190) = 0.11 p = 0.02. No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.
T4	2	F (4, 190) = 2.99 p = 0.02. There is a significant interaction, so the difference found between the assessment methods varies according to the combination of gender, so it is necessary to conduct a specific analysis for each gender via ANOVA repeated Measures. -Gender MM (male-male) - Assessment Methods: F (2, 80) = 33.00 p <0.001 * There is a significant difference. By post hoc Bonferroni the instructor-assessment is significantly lower

		than peer-assessment e self-assessment ($p < 0.001$ for both) . The peer-assessment has significantly below the self-assessment ($p = 0.02$). -Gender FM (female, female) - Assessment Methods: $F(2, 84) = 16.03$ $p < 0.001$ * There is a significant difference. By post hoc Bonferroni the instructor-assessment is significantly lower than those of the peer-assessment and self-assessment ($p < 0.001$ for both). Peer-assessment and self-assessment do not have significant differences ($p = 0.94$). -Gender FF (female, female) -Assessment Methods: $F(2, 26) = 11.57$ $p < 0.001$ * There is a significant difference. By post hoc Bonferroni the instructor-assessment is significantly lower than peer-assessment and self-assessment ($p < 0.001$ for both). Peer-assessment and self-assessment do not have significant differences ($p = 1.00$).
T5	2	$F(4, 190) = 0.32$ $p = 0.87$. No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.
T1 + T2	1	$F(4, 190) = 0.31$ $p = 0.87$. No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.
T3+T4+T5	2	$F(4, 190) = 0.97$ $p = 0.43$.No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.
T1+T2+T3+ T4+T5	1 + 2	$F(4, 190) = 0.78$ $p = 0.54$. No significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.

Table 4 - Gender and assessment methods

The possibility of errors of judgment was also analysed. Confronting ratings with those made by the instructor a set of possible errors was listed:

I. Assign notes to an activity not performed
II. Not assign a ranking to a complete activity (correct or not)
III. Give full marks to activity not fulfilled correctly
IV. Not give full marks to properly fulfilled activity

The following frequencies by error type were obtained, which will be analysed grouped by Bloom levels (Table 5):

Error type	Assessment method	Knowledge (T1 + T2) – Bloom 1		Comprehension (T3+T4+T5) – Bloom 2		
		T1	T2	T3	T4	T5
I	Peer	-	-	5	38	20
	Self	-	-	6	23	8
II	Peer	1	-	-	-	-
	Self	1	-	-	-	-
III	Peer	4	4	-	19	5
	Self	17	25	12	46	24
IV	Peer	-	-	9	-	-
	Self	-	-	-	-	-

Table 5 - Errors X Bloom levels

The first level of Bloom can be seen that there is little bias in the peer reviews and relatively large self-assessment with error III. This result is consistent with the observations noted in previously exposed statistical numbers.

The second level of Bloom does not present the same uniformity error frequencies. In type I error, the frequency of favouritism is higher in peer reviews in task 4 and 5 (produce the missing abstract section). On error type III was observed that favouritism in self-assessments occur more frequently. It should be noted that the bias did not occur consistently when comparing the issues with each other as seen in the previous level of Bloom even the latter two relatively similar issues. However, the third issue is one of the following two different task but showed a pattern in which the self-assessment is superior in multiple (2 and 4, respectively) evaluation errors. This confirms the cohesion between the activities on the second level of Bloom. These observations suggest that there are different behaviours in peer- and self-assessments between the two levels of Bloom's taxonomy.

3.1 Are there significant differences among the three assessment methods by task?

Table 3 demonstrate that the instructor-assessment is significantly lower than those of peer-assessment and self-assessment for all tasks except for task T3 were peer-assessment has significantly lower than those of self–assessment. We expected the instructor more rigorous than the students. Peer-assessment and self-assessment do does not have significant difference for all task except T3. In task 3 Instructors-, peer-, and self-assessment have no significant differences.

Anonymity and the assessment before receiving feedback from peers or instructor may have contributed to the higher level of peer- and self-assessments compared to instructor-assessment. These findings are in the opposition to previous research without double-blind review where student's assessments was done after feedback. (Chang et al., 2013; J. Napoles, 2008).

3.2 Are there significant differences among the three assessment methods by task and by gender?

The results illustrated in Table 4 demonstrate no significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender except by T4 were the instructor-assessment have significantly lower than peer-assessment e self-assessment for all combination of gender. Only for the combination of MM (male-male) the peer-assessment is significantly lower than self-assessment. For FM (female-male) and FF(Female-female) peer-assessment and self-assessment do not have significant differences.

It was not found previous research on gender differences when assessing different tasks.

3.3 Are there significant differences among the three assessment methods by Bloom's taxonomy level one (lower) and two (higher)?

Table 3 demonstrate that the instructor-assessment is significantly lower than peer-assessment self-assessment when considering lower and higher level of Bloom's taxonomy. Peer-assessment and self-assessment do not have significant difference.

Peer-assessment and self-assessment do not have significant difference for lower level task and peer-assessment is significantly lower than self-assessment for higher level task. The difference is due to the level of the task.

Anonymity and the assessment before receiving feedback from peers or instructor may have contributed to the higher level of peer- and self-assessments compared to instructor-assessment. These findings are in the opposition to previous research without double-blind review where student's assessments was done after feedback. (Chang et al., 2013; J. Napoles, 2008).

3.4 Are there significant differences among the three assessment methods by Bloom's taxonomy level 1 and two by gender?

According to Table 4 there is no significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender for lower and higher level of tasks.

It was not found previous research on gender differences when assessing different tasks.

3.5 Are there significant differences among the three assessment methods?

Table 3 demonstrate instructor-assessment is significantly lower than peer-assessment and self-assessment. The peer-assessment method has a strong tendency to have significantly lower than self-assessment method.

Anonimity and the assessment before receiving feedback from peers or instructor may have contributed to the higher level of peer- and self-assessments compared to instructor-assessment. This findings are not consistent to previous research without double-blind review where students assessments was done after feedback. (Chang et al., 2013; J. Napoles, 2008).

3.6 Are there significant differences among the three assessment methods by gender?

The results illustrated in Table 4 demonstrate no significant interaction, so the difference found between the assessment methods is the same for the three combinations of gender.

It was not found previous research on gender differences when assessing different tasks.

4 Discussion

In all tasks (combined or not), significant differences were identified and carried out the Bonferroni post hoc test as shown in Table 6:

Task	Post hoc Bonferroni test
T1	Instructor-assessment < peers-assessment and self – assessment. These last two have no significant
T2	Instructor-assessment < peers-assessment and self – assessment. These last two have no significant difference
T3	Peer-assessment has significantly lower than the self-assessment. These last two have no significant difference
T4	Instructor-assessment < peers-assessment and self – assessment. These last two have no significant difference
T5	Instructor-assessment < peers-assessment and self – assessment. These last two have no significant difference
Knowledge (T1+T2)	Instructor-assessment < peers-assessment and self – assessment. These last two have no significant difference
Compreehension (T3+T4+T5)	Instructor-assessment < peers-assessment and self – assessment. Peer-assessment has significantly lower values than self-assessment
Global (T1+T2+T3+T4+T5)	Instructor-assessment < peers-assessment and self – assessment. Peer-assessment has a strong tendency to have significantly lower values than self-assessment

Table 6 – Research question testing

Based on the result showed in Table 6, it is conclusive that assessments ratings made by the students were significantly higher than those done by the instructor, with the exception of task 3 in which there was convergence between the peer ratings with the instructor keeping the superiority of the ratings given in peer-assessments.

In all research questions except one related to the task four, there was no significant interaction between the assessments methods and gender, so the difference found between the assessment method is the same for the three combinations of genres peer evaluators. In the fourth task, the pairs of male-female and female-female pairs evaluators follow the same group feature is that teacher evaluations is significantly lower than the other two, without peer and self assessment methods presenting significant difference. However, the assessment methods of male-male peer evaluators showed the difference that the marks awarded by peers have significantly lower values than the self-assessments. This feature was considered sufficient to consider that there may be some difference in behaviour between genders when grading able to double

blind peer. This fact was not identified in previous research. However, this effect is in the statistical limit to be considered as factual.

Analysis of Variance (ANOVA) was also carried out gathering the questions by Bloom dimensions. Significant differences in both dimensions were identified and carried out, the Bonferroni post hoc test as shown in Table 6:

When considering Bloom's dimensions Table 6 demonstrate that the Knowledge dimension does not follow the general characteristic that self-assessments are higher than those made by peers-assessment.

From the perspective of whether any favouritism in granting student ratings higher than that given by the instructor, can be noted evidence that it is a fact.

However, the favouritism that is convergent in the dimension of knowledge presents itself differently in the comprehension level from Bloom's taxonomy. This leads to consider the assumption that favouritism is affected by the type of task, corroborating what said Cho, Lee & Jonassen (2011), but that did not associate this with any framework.

On the other hand, this does not ensure that the cause of this discrepancy also may not have been the rubric used. However, this assumption has no basis in own set of used rubrics, since the last two are identical and also feature favouring differences. Thus, it is possible to rule out the effect of the rubric as a factor that might interfere with the statement made by Cho, Lee & Jonassen (2011) and thus confirm the indication that the result of the items is affected only by the type of task.

5 Conclusion and Implication

It can be said that there is evidence of convergence between self and peer assessments compared with those made by the instructor. However, it should be noted that the fourth and fifth rubrics are identical varying only the proposed task was to produce a text referring to the context and gap, respectively. So, difference is attributed to the lack of homogeneity for the activity that was applied to the students, that is, in the Bloom dimension of comprehension may occur in convergence distortion, and therefore the accuracy, evaluations made by the students.

On the other hand, the relatively uniform convergence between self and peer assessments on the Bloom's dimension of knowledge demonstrates the trend towards greater favour in which the student evaluates himself than that made by peers and both superior to that made by the instructor as already identified in previous research. But this configuration is not repeated consistently in the dimension of comprehension and that had not been identified previously.

The reported result corroborates previous research in general, but the taxonomy of Bloom presents a new dimension to interpret the data.

It was found that by offering activities that involve comprehension as a high level of Bloom's taxonomy the discrepancies between the assessments made by the students of their own work and received by their peers may not have uniform behaviour.

This survey was carried in a classroom situation using a Learning Management System, and so it is easily replicable in an online course using the resources already

available in similar systems. Therefore, the research is independent of the number of students since met a few conditions: double blind reviews by peers, and adequately produced rubrics.

The insertion of Bloom's taxonomy to analyse the accuracy of the assessments made by the students advanced knowledge in relation to previous results by the fact show that there are indications of a higher accuracy in assessments in activities that require a higher level of cognitive skills.

The set of findings in this work includes providing an indication that the use of a comprehensive framework can serve as a basis for effective analysis items. Your summative efficiency was confirmed in the first cognitive dimension of Bloom (knowledge) and evidence were identified that in the second dimension (comprehension) there are factors that indicate the need for further research. There is a clear indication that these findings are not directly linked to how the rubric was designed, but the need to produce them considering a framework cannot be ignored.

6 References

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing.* New York-USA: Longman.

Andrade, H. G. (2000). Using Rubrics to Promote Thinking and Learning. *Educational Leadership*, *57*(5), 13–18.

Bonk, C. J., & Smith, G. S. (1998). Alternative instructional strategies for creative and critical thinking in the accounting curriculum. *JOurnal of Accounting Education*, *16*(2), 261–293.

Buzzetto-more, N. A., & Alade, A. J. (2006). Best Practices in e-Assessment. *Journal of Information Techology*, *5*, 251–269.

Chang, C.-C., Liang, C., & Chen, Y.-H. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? *Computers & Education*, *60*(1), 325–334. doi:10.1016/j.compedu.2012.05.012

Chang, C.-C., Tseng, K.-H., & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education*, *58*(1), 303–320. doi:10.1016/j.compedu.2011.08.005

Chen, C. (2010). The implementation and evaluation of a mobile self- and peer-assessment system. *Computers & Education*, *55*(1), 229–236. doi:10.1016/j.compedu.2010.01.008

Cho, Y. H., Lee, J., & Jonassen, D. H. (2011). The role of tasks and epistemological beliefs in online peer questioning. *Computers & Education*, *56*(1), 112–126. doi:10.1016/j.compedu.2010.08.014

Collier-Reed, B. (Ed. . (2011). *Proceedings of the First Biennial Conference of the South African Society for Engineering Education Proceedings of the First Biennial Conference of the South African Society for Engineering Education*.

Gray, D. (2013). Barriers To Online Postsecondary Education Crumble: Enrollment In Traditional Face-To-Face Course Declines As Enrollment In Online Courses Increases. *Contemporary Issues In Education Research*, *6*(3), 345–348.

Huitt, W. (2011). Bloom et al.'s Taxonomy of the Cognitive Domain. *Educational Psychology Interactive*. Retrieved from file:///C:/Users/William Huitt/Documents/Misc/bloom.pdf

Isaacson, J. J., & Stacy, A. S. (2009). Rubrics for clinical evaluation : Objectifying the subjective experience. *Nurse Education in Practice*, *9*(2), 134–140. doi:10.1016/j.nepr.2008.10.015

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics : Reliability , validity and educational consequences, *2*, 130–144. doi:10.1016/j.edurev.2007.05.002

Karamustafaoğlu, S., Sevim, S., Karamustafaoğlu, O., & Çepni, S. (2003). Analysis of Turkish High-School Chemistry-Examination Questions According To Bloom's Taxonomy. *Chemistry Education Research and Practice*, *4*(1), 25. doi:10.1039/b2rp90034c

Lu, R., & Bol, L. (2007). A Comparison of Anonymous Versus Identifiable e-Peer Review on College Student Writing Performance and the Extent of Critical Feedback. *Journal of Interactive Online Learning*, *6*(2), 100–115.

Moskal, B. M. (2000). Scoring Rubrics : What , When and How ? *Practical Assessment, Research & Evaluation*, *7*(3), 1–7. Retrieved from http://pareonline.net/getvn.asp?v=7&n=3

Napoles, J. (2008). Relationships Among Instructor, Peer, and Self-Evaluations of Undergraduate Music Education Majors' Micro-Teaching Experiences. *Journal of Research in Music Education*, *56*(1), 82–91. doi:10.1177/0022429408323071

Napoles, J. (2008). Relationships Among Instructor, Peer, and Self-Evaluations of Undergraduate Music Education Majors'Micro-Teaching Experiences. *Journal of Research in Music Education*, *56*(1), 82–91.

Ng, E. M. W. (2014). Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring. *Computers & Education*, *73*, 141–148. doi:10.1016/j.compedu.2013.12.016

O'Toole, R. (2013). Pedagogical strategies and technologies for peer assessment in Massively Open Online Courses (MOOCs).

Sadler, P. M., & Good, E. (2006). The Impact of Self- and Peer-Grading on Student Learning. *Educational Assessment2*, *11*(1), 1–31. doi:10.1207/s15326977ea1101_1

Sadler-smith, E. (2007). in Management Education, *6*(2), 186–205.

Sitzmann, T., Brown, K. G., & Bauer, K. N. (2010). Self-Assessment of Knowledge : A Cognitive Learning or Affective Measure ?, *9*(2), 169–191.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *The Journal of Applied Psychology*, *93*(2), 280–95. doi:10.1037/0021-9010.93.2.280

Taras, M. (2005). Assessment – Summative and Formative – Some Theoretical Reflections. *British Journal of Educational Studies*, *53*(4), 466–478.

Theising, K., Wu, K., & Heck Sheehan, A. (2014). Impact of peer assessment on student pharmacists' behaviors and self-confidence. *Currents in Pharmacy Teaching and Learning*, *6*(1), 10–14. doi:10.1016/j.cptl.2013.09.020