Proposal of statistical analysis to support the assessment method of a Brazilian Industrial Engineering course

Juliana Helena Daroz Gaudêncio[a]*, Anna Paula Galvão Scheidegger[b], Camila Pereira Pinto[a], João Batista Turrioni[a], Ana Maria Silveira Turrioni[c]

[a]Universidade Federal de Itajubá, Itajubá, MG, Brazil

[b]Texas A&M University, College Station, TX, USA

[c]Pontífica Universidade Católica de São Paulo, São Paulo, SP, Brazil

*juliana_hdg@yahoo.com.br

Abstract

The requirement to adapt the university education to demands of society and professional market is improving the teaching methodologies that try to develop skills and prepare the students to deal with day-to-day situations of a business environment. Therefore, this research presents the assessment method utilized in an Industrial Engineering course carried out by a Brazilian public university in partnership with a multinational company. The course had 37 participants, consisting of: 28 students divided into four groups; 1 teacher; 4 university tutors; and 4 company tutors. The main objective is to assess the consistency of grades assigned to students and their work groups using agreement, variance and correlation analyses. As conclusion, the analyses indicated a possible deficiency in the assessment method application since the values of agreement and correlation coefficients were lower than expected but also provided a positive contribution to the improvement of assessment and the course as a whole.

Keywords

Assessment method. Active learning. Active teaching. Higher education. Engineering education.

How to cite this article: Gaudêncio, J. H. D., Scheidegger, A. P. G., Pinto, C. P., Turrioni, J. B., & Turrioni, A. M. S. (2017). Proposal of statistical analysis to support the assessment method of a Brazilian Industrial Engineering course. *Production, 27*(spe), e20162189. http://dx.doi.org/10.1590/0103-6513.218916

1. Introduction

In the last years, the need to adapt education to the demands of society and, in particular, of the professional market has attracted considerable attention of educators and researchers. Currently, the universities shall be able not only to provide knowledge to students, but also to develop skills and prepare them to deal with day-to-day situations of a business environment. The challenge of education in modern society is even greater within engineering education, since it should focus primarily on the development of students' ability to solve problems and provide innovative solutions and for this reason the application of practices that strengthen self-learning and cooperative learning is necessary, if not essential (Chua et al., 2014).

An integrated economy demands workers with insight into international issues that should be developed preferably from their academic education. Higher education that is global in its proposals for courses and forms of management can greatly contribute to the formation of these workers (Mückenbergera et al., 2013). Meanwhile, the companies invest in corporate education for unprepared workers, showing that higher education is failing in some aspects and thus demanding study (Vieira & Francisco, 2012).

Recognizing the importance of changing higher education, new teaching methods are developing in which students are no longer passive recipients of information but become the center of attention. The practice of this type of learning is a culture shock for new students and for educators who use this technique in their disciplines. Besides the challenge of changing the way of conducting classes, teachers are faced with the task of determining how to properly assess student learning. Traditional assessment techniques which are only concerned with the assignment of grades don't correctly evaluate the evolution of students (Waters & Mccracken, 1997; Eva, 2001; Kritikos et al., 2011).

According to Norton et al. (2013), due to recent changes at higher education to focus on student learning, much has been written about the students' assessment and its effects on learning. However, the integration of assessment with the learning and the innovations in how to assess remain a challenge (Struyf et al., 2001). Moreover, the pressure on institutions to establish more effective ways to evaluate, led the assessment modes to the center of engineering education discussions (Raud, 2010).

Considering this context, it was created a course in a Brazilian public university named "Project Semester in Industrial Engineering" that aimed to use an active approach of education where students were divided into groups to solve real problems of a multinational company. During the course, the use of this methodology provided difficulties in the way of how students would be assessed. Thus, this research aims to assess the consistency of grades assigned to students and their work groups using statistical analysis together with a qualitative approach which assistance in define the better assessment mode considering the context of a course based on active teaching. Thereby, an agreement, variance, and covariance analyses have been proposed in order to assist results' interpreting of implemented evaluations.

The work proposal described in this paper comes in conjunction with the active teaching methodology described by Shekar (2007) that relates an active learning course in product development as part of the engineering undergraduate education in New Zealand, and by Donohue (2014) that relates how group-based audience response systems quizzes have been integrated into a civil engineering course in foundation design at the United Kingdom. These works show that the active teaching has been diffused at engineering in several countries.

2. Learning assessment

Recent efforts to improve education have been challenging instructors to shift their focus from what they teach to what their students learn (Agrawal & Khan, 2008). Along with changing course content and instructional modes, instructors have placed greater emphasis on assessment. Assessment is an important item that teachers should include when planning their teaching method because it gives students' stimulus to their learning. More recently, modes such as self-assessment, peer assessment, simulations and other innovative techniques have been introduced in higher education; however, it is not a simple task choosing the most appropriate mode (Shahadan et al., 2012).

It is worth emphasizing the importance of distinguishing two types of assessment: formative and summative as described by Phillips (2005). The formative has a development proposal, helping students learn more effectively and providing them ongoing feedback on their learning. While summative is compared to the traditional approach, measuring what students learned at the end of a set of activities and thereby giving them a final grade (Norton et al., 2013). To ensure that the assessment is part of the learning process, the evaluation must be focused on the students and the educational institution must build a curriculum centered on their professional and personal development. Assessment instruments and their approaches need to be developed so that the students can demonstrate their acquired knowledge instead of only exposing memorized information (Brown, 2004).

In a case study that analyzed the students' perceptions of summative and formative assessments, the summative evaluation has been seen as a necessary evil, once it simply appeared as a mode of measuring memory or the ability to list facts. On the other hand, the formative assessment was considered fairer because it allowed measurement and development of qualities, skills and competencies that would be valuable in other life situations. A means of conducting the formative assessment includes developing activities aimed to simulate a real-life context in such a way that students can understand more clearly the relevance of their academic work to broader situations, i.e. outside the university environment. In sum, the case study showed that perceptions of poor learning, lack of control and irrelevant tasks in relation to summative assessment contrasted sharply with the perception of high quality learning, active participation of students, opportunities for feedback and meaningful tasks in relation to the formative assessment (Sambell et al., 1997).

Focusing on the formative assessment, peer and self-assessment would appear to be an ideal tool to facilitate learning-oriented assessments. The formative assessment has the capacity to encourage students to take more responsibility for their own learning by requiring them to provide their own feedback, contribute to their own assessment and to the assessment of their peers. Having students to provide feedback improves their judgement, assessment ability and critical evaluation skills (Willey & Gardner, 2010).

Measuring what students know is one of the most complex and controversial aspects of education and because of this, it is important to conduct studies to explore issues related to measurement, techniques and technologies available for the assessment process, once the evaluations' results can be used as a relatively efficient and objective source of information in the search to understand the best way of students' learning (Heubert & Hauser, 1999).

3. Method

The "Project Semester in Industrial Engineering" course was held through a partnership between a public educational institution and a multinational company in Brazil which involved students of the eighth semester of Industrial Engineering. For this, some company's real problems were presented to students who had the task of proposing different and innovative solutions in order to be prepared for real and challenging business situations. The course had 37 participants, consisting of: 28 students divided into four groups; 1 teacher; 4 university tutors; and 4 company tutors. Therefore, two tutors, one from the company and other form the university monitored each group of students. Figure 1 shows the structure described herein.

This research followed a qualitative approach with regard to the explanation of the assessment method adopted in the course and in analyzing whether the results were consistent with perceptions of people involved in its conduction. On the other hand, it was essentially a quantitative research since, through statistical analyses, it sought to interpret the grades assigned to students and their work groups as well as the quality of the assessment method applied. Al-Nashash et al. (2009) adopted a similar approach that used several qualitative

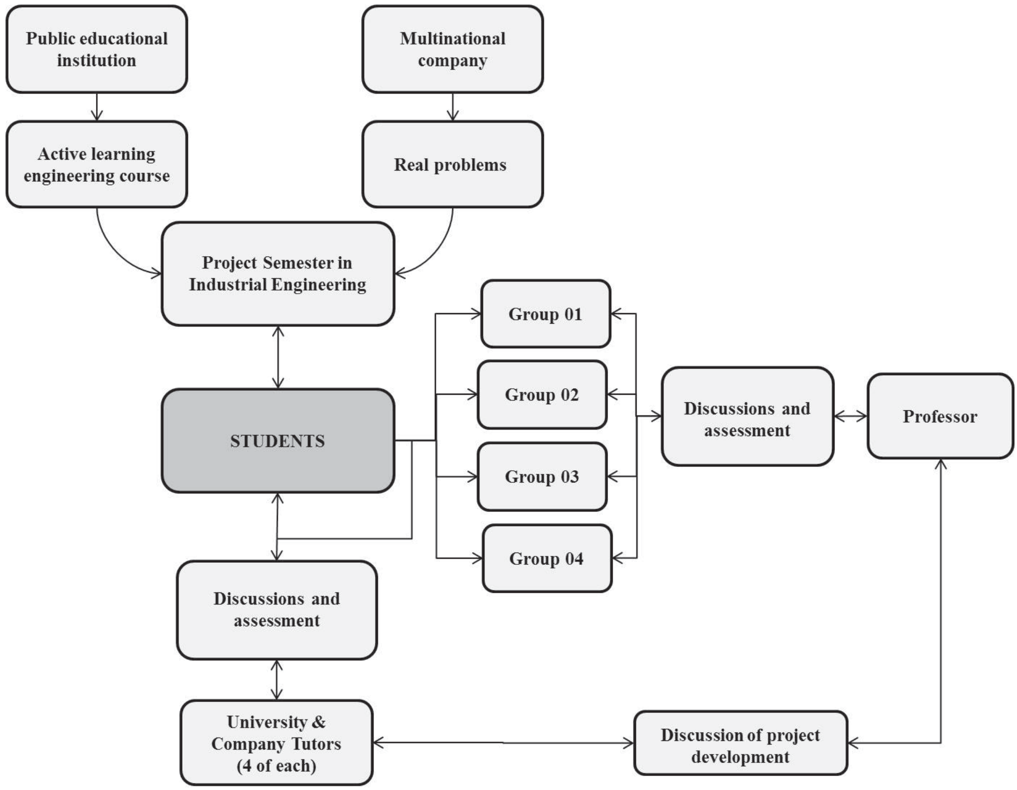
Public educational institution
Multinational company
Active learning engineering course
Real problems
Project Semester in Industrial Engineering
Group 01
Group 02
Group 03
Group 04
STUDENTS
Discussions and assessment
Professor
Discussions and assessment
University & Company Tutors (4 of each)
Discussion of project development

Figure 1. Project Semester in Industrial Engineering course structure.

and quantitative assessment tools to achieve the educational program objectives and outcomes of the electrical engineering course at the American University of Sharjah, the United Arab Emirates.

To carry out the statistical analyses, data were collected through questionnaires and participant observation by university tutors who accompanied the students on their activities.

The statistical analyses were performed through the assistance of the software Minitab® following a three-step approach. Initially, the agreement analysis technique was applied in order to attest the level of agreement between the grades assigned by tutors. Following, an analysis of variance (ANOVA) was performed to verify differences between the grades regard to the students or groups' participation in the project as well as the assessment performed by tutors or by classmates. Finally, a data correlation analysis was performed in order to examine the relationships between the grades given by tutors to the groups and also between the grades obtained by students in each group.

3.1. Ethical consideration

By involving real problems of a multinational company, the process of data access was carefully considered in this study. Therefore, an agreement was signed by the company, the university, students and tutors to ensure the confidentiality of the company's data and the right to publish the results of the project.

4. Description of the assessment method

The teacher in partnership with the university tutors thought in the assessment method adopted in the discipline in order to give better grades those students who were dedicated from the beginning and strove to learn. Starting from these premises, the periodic assessments must be undertaken throughout the semester in order to monitor the participation of each student and the knowledge acquired by him/her. For this, students were evaluated in four different ways: peer assessment (student-student), university tutor-student assessment, professor-student assessment, and, finally, company tutors-group assessment. These assessments are detailed below and Table 1 summarizes their main characteristics.

Table 1. The assessment structure adopted in the course.

Assessment	Method of assessment	Periodicity	Grade composition
Peer	1. Questionnaire with the quantitative scale	Unique	20%
University tutor-student	1. Unstructured and non-standardized	Weekly	20%
	2. Questionnaire with the quantitative and the qualitative scale about the presentation	Unique	
Professor-student	1. Evaluation of presence and participation in class	Weekly	30%
	2. Evaluation of the submitted project reports	Twice	
Company tutor-group	1. Unstructured and non-standardized	Monthly	30%
	2. Questionnaire with the quantitative and the qualitative scale about the presentation	Unique	

- Peer assessment: each student evaluated other members of your group considering 18 criteria of competence with grades from 0 to10 without repetition. This restriction was imposed to avoid possible friendship interference;
- University tutor-student assessment: (1) the tutor weekly graded each student with grades from 0 to 10 without repetition to differentiate each student's participation throughout the project; and (2) evaluation the groups' final presentation where each tutor assessed the four groups considering 16 criteria, with grades from 1 to 6, without repetition;
- Professor-student assessment: (1) periodic assessments where the teacher graded the presence and participation of students; and (2) assessment of the project progress through two reports in order to measure whether the groups were fulfilling the goals set by the company tutors;
- Company tutor-group assessment: (1) the tutors monthly evaluated the evolution of the solutions presented by the group; and (2) assessment the groups' final presentation, considering the same 16 criteria and rating scale (1-6) used by university tutors.

5. Analysis of students' grades obtained through questionnaires

The final presentation grades given by university and company tutors were initially analyzed using agreement analysis to attest the level of agreement between the grades assigned by Tutor # 04 (T4), who is the manager of one of the company's departments, to the other tutors (evaluators). Then, ANOVA one-way was done to prove, statistically, the differences between the mean scores assigned to students individually, through peer review, and to groups by evaluating the final presentation. Finally, it was conducted a correlation analysis of the grades given by tutors to groups and also of those obtained by students in each group as also used by Timossi et al. (2010).

5.1. Agreement analysis: groups' evaluation held by university and company tutors

This analysis was applied to the final presentation assessment of each group which included 16 criteria, such as: whether the presentation of the problem and of the proposed solutions was exposed clearly; whether the main idea of the work was mentioned; or even if the pace of presentation was appropriate; among others as described in the Table 2. This analysis certifies the level of agreement of grades assigned by Tutor T4 who is the most experienced tutor in the problem area in relation to the grades assigned by the remaining. Tutors T1, T2, T3, and T4 are from the company and Tutors T5, T6, and T7 are from the university. Table 3 illustrates this comparison of agreement among Tutor T4 and the remaining.

Table 2. The description of the criteria used in the assessment of oral presentation.

Criteria
1. Was the exhibition structured clearly?
2. Did the presenter know to conduct the sequence between the topics of slides?
3. Was the main idea of the work presented clearly and understandable?
4. Were examples presented relevant?
5. Were examples presented interest?
6. Was the pace of the presentation adequate?
7. Was the volume of information in the presentation adequate?
8. Was the tone of the presenter's voice adequate?
9. Were the slides efficiently used?
10. Was the presentation well prepared?
11. Was the presentation interesting and motivating?
12. Was the presenter's eye contact with the audience right and proper?
13. Did the presenter manage to hold the audience's attention?
14. Did the presenter show / summarize the main points of the presentation?
15. Did the presenter demonstrate confidence and knowledge about the topic presented?
16. Was the time spent in the presentation well managed?

Table 3. The percentage of the agreement for all groups evaluated.

Appraiser	Inspected	Matched	Percent	95% CI[a]
T1	64	14	21.88	(12.51; 33.97)
T2	64	29	45.31	(32.82; 58.25)
T3	64	18	28.13	(17.60; 40.76)
T5	64	21	32.81	(21.59; 45.69)
T6	64	14	21.88	(12.51; 33.97)
T7	64	19	29.69	(18.91; 42.42)

[a] Confidence interval (CI).

As a result, this analysis shows that Tutor T2 was the one that reached the highest percentage of agreement with Tutor T4, totaling 45.31% among the four completed assessments of the final presentation. The Tutor T2 is a company tutor with plenty of experience in his work and thus taking a closer point of view of T4. As mentioned, the assessment was composed of 16 criteria for each of the four groups and thus a total of 64 grades were compared.

The percentage of agreement is a possible method of summarizing the agreement between pairs of observations. However, this method can be misleading, once the proportion of individuals for which there is no agreement does not tell anything. Therefore, it is suggested to use Cohen's Kappa method of analysis which is a more robust measure (Cohen, 1960). Some Kappa's coefficient agreement analysis can be viewed in works developed by Ayatollahi et al. (2013), Lee et al. (2011), Shankland et al. (2010), and Feletti & Sanson-Fisher (1983). In order to maintain a consistent nomenclature to describe the relative strength of the agreement associated with Kappa statistics, labels are assigned to the corresponding time intervals of Kappa's coefficient as described by Landis & Koch (1977) and are indicated in Table 4.

Table 4. The interpretation of Kappa's coefficient.

Kappa Statistic	Strength of Agreement
< 0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

Source: Landis & Koch (1977).

Table 5 illustrates the degree of agreement through Kappa's coefficient by the same analysis previously shown in Table 2. It can be noted that Tutor T2, who has obtained the highest percentage of agreement in the previous analysis, has a slight agreement relationship with T4, since his Kappa's coefficient is approximately 0.21. But, in general, the assessments performed by tutors are not consistent with the standard assessment performed by T4, once the Kappa's coefficients are very close to zero or negative. So, it is concluded that the tutors didn't essentially apply the same pattern when evaluating the groups' performance and, therefore, it is necessary a better alignment between tutors on the meaning of each criterion and how to punctuate them.

Table 5. The degree of agreement obtained by the Kappa's coefficient for all groups evaluated.

Appraiser	Response	Kappa
T1	Overall	-0.0329
T2	Overall	0.2092
T3	Overall	0.0403
T5	Overall	0.1485
T6	Overall	-0.0147
T7	Overall	0.0890

It is performed another agreement analysis to the groups individually, in order to obtain the robustness of the agreement that exists between the assessments performed by T4 compared to other evaluators for each group separately. The results were very close to the previous analysis, that is, again a weak agreement was found between T4 and other tutors within each group. This reinforces the conclusion that there is a lack of standardization on how to evaluate the groups' performance, i.e., how to fairly and homogeneously assign grades according to the criteria. Therefore, it is noted a deficiency of the assessment method adopted in the discipline.

5.2. Analysis of variance (ANOVA)

ANOVA is a collection of statistical models to analyze differences between the means of data set and its associated procedures. This technique allows multiple groups be compared at the same time (Roberts & Russo, 1999). This comparison is given in relation to the hypothesis test applied to each data set and analyzed through two hypotheses: (1) null hypothesis (H_0) that the population means are equal when p-value > 0.05; and (2) alternative hypothesis (H_1) that the population means are different when p-value ≤ 0.05. Practical examples of the use of ANOVA can be found elsewhere (see, for example, Montgomery, 2012).

ANOVA one-way was applied in two different ways: (1) it was applied to grades of the seven students (S1 – S7) obtained through peer assessment as described in Table 6; this analysis verify the degree of student's

Table 6. The description of the criteria used in the peer assessment.

Criteria	
1. Empathy	2. Sanity
3. Communication	4. Initiative
5. Proactivity	6. Flexibility
7. Organization	8. Decision-making
9. Self-development	10. Ethic
11. Committal	12. Responsibility
13. Ecological awareness	14. Interpersonal relationship
15. Team spirit	16. Customer focus
17. Focus on results	18. Guidance for quality

participation inside their work group; and (2) it was applied to verify the participation among the groups in the engineering course.

5.2.1. Analysis of students' participation within the group

The first step is to define the normality of the data (students' grades) through the test of Anderson-Darling. As a result, the p-value > 0.05 for all students' grades within each group. In sequence, the second step is to verify the variance homogeneity: this step is conduct by the Levene's test that considering the F statistic. The result of Levene's test is analyzed through two hypotheses: (1) null hypothesis (H_0) that all variances are equal when p-value > 0.05; and (2) alternative hypothesis (H_1) that at least one variance is different when p-value ≤ 0.05.

The Levene's test indicated that the variances of students' grades are equal when the groups G1, G2, and G4 are analyzed once p-value > 0.05. However, the variance obtained by the students' grades for G3 aren't equal; therefore, the ANOVA isn't applied in this scenario. In order to conduct the ANOVA, the outliers were removed from the data since there indicate divergences of opinion among the students and in this way, characterizing a personal evaluation of teammates instead of considering the real participation of the member within the group.

Table 7 shows the ANOVA of the grades obtained through the peer assessment conducted by all groups.

The first analysis performed for Group 1 shows that the average grades among seven students is statistically different. However, after a brief analysis of the data through the box plot graph shown in Figure 2, it was observed that Student 4 (S4) received, mostly, lower grades in relation to the six other members of the group. Thus, ANOVA analysis was redone excluding this student and the average between the grades became statistically equal and proved the homogeneity of students' participation with the exception of S4.

Table 7. ANOVA one-way of peer assessment.

Group	Students	Pvalue	Valid hypothesis	Population's average
1	S1, S2, S3, S4, S5, S6, S7	0.000	H_1	Different
1[b]	S1, S2, S3, S5, S6, S7	0.574	H_0	Equal
2	S1, S2, S3, S4, S5, S6, S7	0.996	H_0	Equal
3[a]	S1, S2, S3, S4, S5, S6, S7	0.000	H_1	Different
3[b]	S3, S4, S5, S6, S7	0.000	H_1	Different
4	S1, S2, S3, S4, S5, S6, S7	0.013	H_1	Different
4[b]	S1, S2, S3, S5, S7	0.411	H_0	Equal

[a]Analysis without the presence of outliers; [b]Analysis without students with lower or higher grades in relation to the others.

The analysis conducted for Group 2 showed that the average grades among the seven students were statistically equal, unlike the analysis performed for Group 3 which showed that the average grades among the students were statistically different. This last analysis was done after the identification and elimination of outliers, as shown in Figure 3. The analysis was redone once again, disregarding the Student 1 (S1) who obtained a set of grades lower than the other students, as well as the Student 2 (S2) who has obtained a set of grades greater in relation to others, as can be seen in Figure 3. However, again the result was maintained with the average grades remaining statistically different and suggesting that there was no homogeneity in students' participation in this

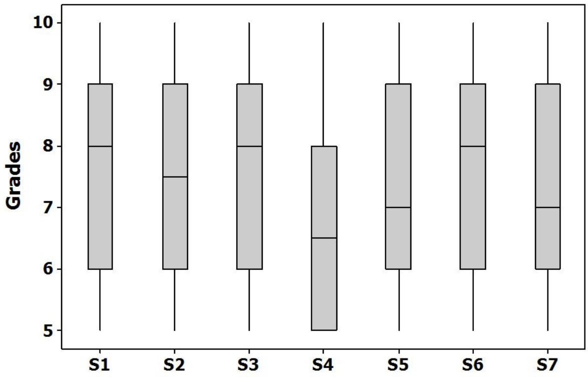

Figure 2. Boxplot of the Group 1students' grades.

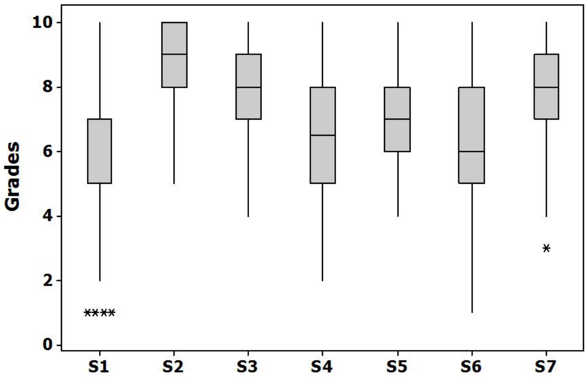

Figure 3. Boxplot of Group 3 students' grades.

group. This analysis was aligned with the university tutor opinion who directly accompanied the group and related that some students were extremely dedicated while others were disinterested. Also, this may have happened by the work method chosen by Group 3. This group chose to split their project into two teams, each running a subproject and this division made it difficult for students assessing colleagues on the other part of the group.

Finally, the analysis carried out for Group 4 showed that the average grades among the students were statistically different. However, excluding S4 who had received mostly lower grades, and also S6 who had received mostly greater grades (Figure 4), analysis was remade and the average grades became statistically equal, confirming the homogeneity of the group, except for S4 and S6.

All the outliers described in this section were identified by the boxplot which is a useful graphical display for describing the behavior of the data in the middle as well as at the ends of the distributions. The boxplot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range. If the data stays above the Q3+1.5*(Q3-Q1) or below the Q1-1.5*(Q3-Q1) it is defined as an outlier and it could be represented by an asterisk.

5.2.2. Analysis of the group's participation during the course

According to the same steps stablished to the students' grades through the peer assessment; the first step defines the normality of the data (groups' grades) by means of Anderson-Darling test and, as a result, p-value > 0.05 for all groups. The second step performs Levine's test that indicated non-homogeneity variance among the groups' grades once p-value = 0.02. Thus, the ANOVA isn't applied in this scenario. In order to conduct the ANOVA, the outliers shown in Figure 5 were removed from the data since there indicate the lack of standardization among the tutors at the moment of the grades' definition as related previously by the agreement analysis. Figure 5

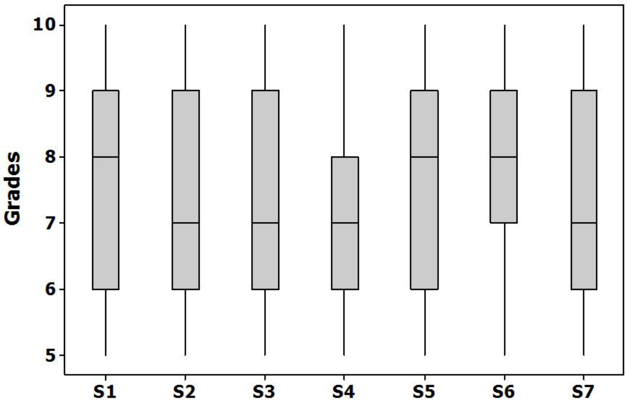

Figure 4. Boxplot of Group 4 students' grades.

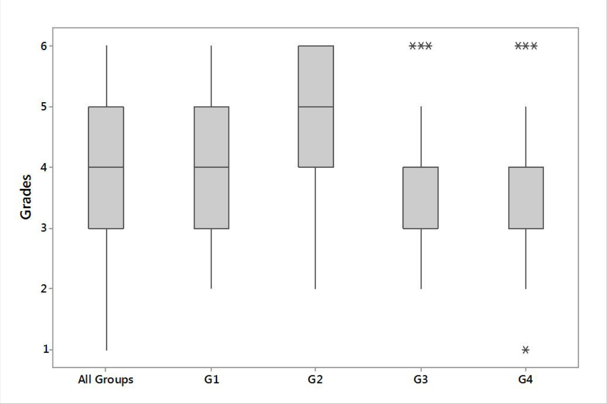

Figure 5. Boxplot of groups' grades.

shown the groups' grade distribution thought the boxplot chart. It is possible to note by a visual analysis that the groups G3 and G4 appeared to have the same grades' distribution.

After the outliers removing, the ANOVA was applied to the grades obtained by groups through the final presentation assessment as described in Table 2. The hypothesis test was applied according to the information contained in Table 8.

Table 8. ANOVA of groups' final presentation assessment.

Groups	Pvalue	Valid hypothesis	Population average
1, 2, 3, 4	0.000	H_1	Different
3 and 4	0.197	H_0	Equal

The analysis was performed between all groups and showed that the average grades between them was statistically different. The second analysis was performed between the groups G3 and G4; which, after a visual analysis through the box plot graph shown in Figure 5, appeared to have the same grades' distribution. The ANOVA confirmed this fact which showed that the grades' average was statistically equal.

At a 0.05 level of significance the intervals for each group's average are obtained and showed on Figure 6. If the intervals do not overlap, it identifies means that differ from each other. Otherwise, the means are considered equal according it happens with groups G3 e G4. Therefore, it is possible to conclude at the 0.05 level of significance that the group G2 obtains the higher mean of grades when compare to the remaining groups.

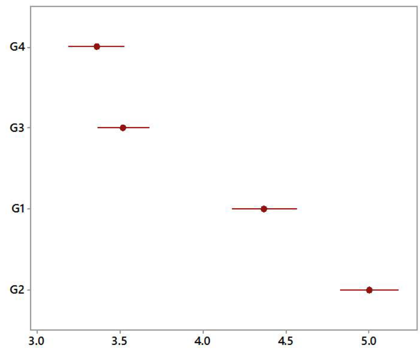

Figure 6. Means comparison chart of G1, G2, G3, and G4.

5.3. Correlation analysis

The correlation ‘is a measure of the linear relationship between random variables, wherein the correlation coefficient “r” is a quantitative measure of the strength linear relationship between two random variables and is also called Pearson’s coefficient. If two variables are perfectly linearly related with a positive slope, r = 1; and if they are perfectly linearly related with a negative slope, r = -1; and, if there is no linear relationship, r = 0. Correlations below |0.5| are generally considered weak while above |0.8| are strong (Montgomery & Runger, 2013). Correlation analyses between scales of learning and its influence factors in the higher education network are detailed by Goodyear et al. (2005).

5.3.1. Analysis of grades assigned to students by their peers

The correlation analysis is a parametric test, so the first step is to define the normality of the data through the test of Anderson-Darling. As a result obtained previously, the p-value > 0.05 for all grades obtained within and between groups. Therefore, analysis was performed in G1, as an example, considering S4 who presented a set of lower grades compared to the other students of the same group.

It is noted in Table 9 that the coefficients indicate only a moderate correlation once, in most cases, they have values less than |0.5|. The grades assigned by the pair S2-S3 are the most strongly correlated, i.e., these students likely attributed grades to S4.

Table 9. Pearson’s correlation coefficient of grades received by S4 in G1.

	S1	S2	S3	S5	S6
S2	0.253				
S3	0.220	0.629*			
S5	0.175	0.207	0.160		
S6	0.308	0.413	0.190	0.462	
S7	0.056	-0.115	-0.137	-0.159	-0.034

*Significant p-value (pvalue < 0.05).

It wasn’t performed the correlation analysis for G2 due to the homogeneity between the grades received by all students in this group and, thus, there isn’t a student who stood out within the group. Although, the correlation analysis performed in G3 was conducted for S1 who obtained a set of lower grades, and also for S2 who obtained a set of greater grades.

It is noted in Table 10 that mostly correlations are moderate with coefficients less than |0.5|, except for the pair S2-S6 who attributed grades very similar to S1 and S3-S6, S3-S7 pairs that likely attributed grades to S2.

Finally, the correlation analysis of G4 was conducted for S4 who obtained a set of lower grades and also for S6 who, inversely, obtained a set of greater grades. Similar to the previous analysis, it is observed in Table 11

Table 10. Pearson's correlation coefficient of grades received by Student 1 (a) and Student 2 (b) of Group 3.

(a)	S2	S3	S4	S5	S6	(b)	S1	S3	S4	S5	S6
S3	0.295					S3	0.250				
S4	0.441	0.187				S4	0.159	0.443			
S5	0.178	-0.056	-0.055			S5	0.335	0.393	0.283		
S6	0.520*	0.064	0.441	0.112		S6	-0.028	0.618*	0.292	0.163	
S7	0.272	-0.147	0.226	0.053	0.316	S7	0.082	0.612*	0.222	0.483	0.284

*Significant p-value (pvalue < 0.05).

Table 11. Pearson's correlation coefficient of grades received by Student 4 (a) and Student 6 (b) of Group 4.

(a)	S1	S2	S3	S5	S6	(b)	S1	S2	S3	S4	S5
S2	-0.008					S2	-0.388				
S3	-0.333	-0.332				S3	-0.483	0.262			
S5	-0.120	0.286	-0.263			S4	0.182	-0.194	-0.267		
S6	0.246	-0.063	0.038	-0.306		S5	0.090	-0.279	-0.160	-0.069	
S7	0.279	-0.102	-0.035	-0.193	0.976*	S7	0.055	0.208	0.228	0.069	-0.037

*Significant p-value (pvalue < 0.05).

that the correlation coefficients were moderate with values less than |0.5|. In this case, the exception was the pair S6-S7 that attributed grades much similar to S4.

Through the correlation analysis carried out, it was obtained a similar result of agreement analysis. That is, the students' grades who have detached within each group, in a positive or negative way, are not strongly correlated because they had differing opinions regarding their teammates in each criterion. Although, as indicated by ANOVA, they agreed among themselves about the average grades of their colleagues. This fact emphasizes, once again, that or the criteria to be evaluated were not well defined or they did not use the same parameter for grading.

5.3.2. Analysis of the grades assigned to groups by university and company tutors

According to Tables 12-15, the grades assigned by tutors to groups' final presentation were also performed by correlation analysis. It is noted in Table 12 that the correlation coefficients indicated a moderate correlation once, in most cases, the coefficients had values less than |0.5|, except for the pairs T2-T5 and T4-T7 who attributed grades very similar to G1.

In Table 13 is noted that the correlation coefficients were again not very strong except for pairs T1-T3 and T5-T7 who had attributed grades very similar to G2.

Table 12. Pearson's coefficient of grades assigned by tutors to G1.

	T1	T2	T3	T4	T5	T6
T2	0.343					
T3	0.365	0.346				
T4	0.414	0.290	0.305			
T5	0.394	0.510*	0.372	0.405		
T6	-0.288	0.237	-0.107	0.010	0.226	
T7	0.139	0.357	0.152	0.521*	0.461	0.342

*Significant p-value (p-value < 0.05).

Table 13. Pearson's coefficient of grades assigned by tutors to G2.

	T1	T2	T3	T4	T5	T6
T2	0.390					
T3	0.554*	0.192				
T4	0.221	0.232	0.031			
T5	0.074	0.259	0.133	-0.124		
T6	0.048	0.218	-0.374	-0.313	0.427	
T7	0.222	0.170	0.336	-0.174	0.540*	0.229

*Significant p-value (pvalue < 0.05).

Table 14. Pearson's coefficient of grades assigned by tutors to Group 3.

	T1	T2	T3	T4	T5	T6
T2	0.654*					
T3	0.123	-0.338				
T4	0.563*	0.363	0.267			
T5	0.330	0.324	-0.122	-0.226		
T6	0.523*	0.692*	-0.048	0.678*	0.238	
T7	-0.170	-0.145	-0.147	0.021	-0.347	-0.100

*Significant p-value (pvalue < 0.05).

Table 15. Pearson's correlation coefficient of grades assigned by tutors to G4.

	T1	T2	T3	T4	T5	T6
T2	-0.115					
T3	-0.241	0.139				
T4	0.423	0.422	0.232			
T5	-0.159	-0.092	0.362	-0.071		
T6	0.524*	-0.207	0.255	0.377	0.300	
T7	0.124	0.645*	-0.050	0.119	0.099	-0.051

*Significant p-value (pvalue < 0.05).

It is noted again in Table 14, a greater presence of correlation coefficients with values less than |0.5|. On the other hand, there were five pairs of tutors with a moderate correlation between the grades assigned to G3 which featured that some tutors had a similar point of view regarding the presentation of this group.

Finally, in Table 15 it is observed that the correlation coefficients were mostly smaller than |0.5|, and thus did not represent a strong correlation, except for pairs T1-T6 and T2-T7.

As seen, only a few tutors attributed very similar grades on all criteria for groups and this fact mainly occurred in the grades assigned to G3. The fact that the grades assigned by tutors are not closely related means that tutors do not equally assessed groups within each of the criteria, showing a misalignment between them on understanding the criteria or on the parameter for diminishing the grades.

During the project conducting, the groups were followed by a total of seven tutors: three from the university and four from the company. The university's tutors had a greater alignment in the moment to assigning the grades to the groups which didn't happen with the company's tutors. This fact is attributed by the proximity maintained by the university's tutors since they were responsible for closer students' monitoring and they knew about the progress and difficulties of all four groups. On the other hand, the four company's tutors carried out a monthly monitoring only with their respective group.

6. Limitations

This research arose from the difficulty of measuring student learning in a course based on active learning method. Thus, statistical analyses are performed to evaluate whether the grades received by students were homogeneous and if there was consistency in the assessment method adopted. The assessment method showed herein was specifically defined for the course "Project Semester in Industrial Engineering". Also, the number of participating groups in the course was set according to the number of projects proposed by the multinational company that were four. Therefore, were chosen four company tutors and more three university tutors along with the teacher to assistance the groups.

7. Discussion of the results obtained by statistical analyses

Figure 7 summarizes the statistical analyses performed herein. The agreement analysis of grades attributed to groups' final presentation attest the agreement degree of assessment performed by T4 compared to the remaining tutors. As a result, the Kappa's coefficients provided negative values or tending to zero. This result provides an interpretation that there is a lack in tutors' evaluation standardization, i.e., tutors did not quantify their perceptions of students' skills in the same way.

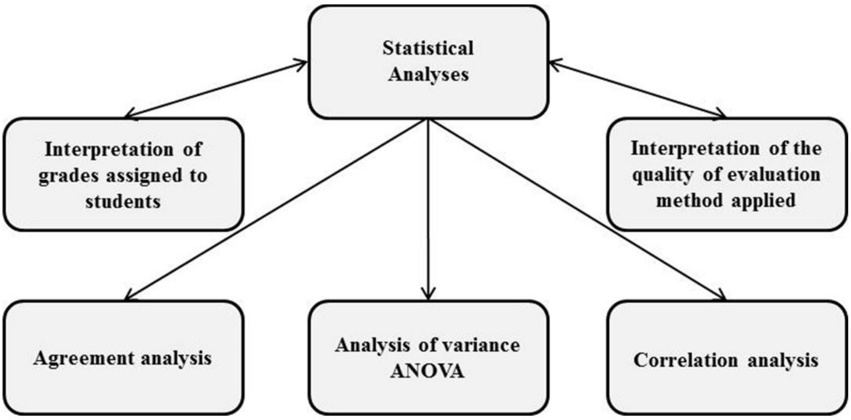
Statistical Analyses
Interpretation of grades assigned to students
Interpretation of the quality of evaluation method applied
Agreement analysis
Analysis of variance ANOVA
Correlation analysis

Figure 7. Statistical analyses schematic.

Subsequently, ANOVA one-way was performed to verify statistically whether the average grades received by students and groups were equal or not, and also to check if the average given by each tutor or each team members were equal. Thus, it was possible to examine whether students and groups participated in a similar manner during the project and whether tutors or teammates similarly evaluated students in relation to their peers.

The ANOVA performed among the four groups by assessing the final presentation indicated a differentiation between the grades obtained by groups and proved the difference between their performances, where G2 had higher grades and also there was the closeness between the grades received by G3 and G4, whose performances were considered by tutors inferior to others. These results were the same described qualitatively by tutors to researchers and therefore well portray the tutors' perceptions. Thus, for this purpose, the assessment method described herein was satisfactory. Also, regarding to ANOVA, it was observed that tutors gave statistically different average grades for all groups and therefore did not have a same view to evaluating the groups' performance. The same gap was reported by the agreement analysis which showed that the tutors have different points of view and they were not minimized by a standardized way to assess the students.

Lastly, correlation analysis of the student's peer assessment was performed. By analyses were observed moderate correlations between the grades, obtaining a similar result for the agreement analysis. That is, the fact that students' grades who have detached within each group were not strongly correlated, could be explained by the reason that students had differing opinions regarding their teammates in each criterion. Although, as indicated by ANOVA, they agreed among themselves about to the average grades of their teammates. This indicates that either the criteria to be assessed were not well defined or students did not use the same parameter for grading. Also, correlation analysis was performed to the final presentation assessment performed by tutors and according to this analysis, only a few tutors attributed very similar grades for groups and this fact was mainly attributed to G3.

7.1. Analysis of the formative assessment

The peer assessment has the potential to provide accurate and valid assessment information; several factors will influence the quality of the results. Specifically, they are reliability, relationships, stakes and equivalence as detailed by Norcini (2003).

Considering the reliability, the literature shows clearly that even experienced evaluators differ when observing exactly the same events. Consequently, evaluations from several colleagues are needed to achieve reliable results. This scenario was conducting during the peer assessment application since each student was evaluated by six colleagues.

The nature of the students' relationships that were being evaluated can demonstrate difficulties since students who compete with each other or who are personal friends may be motivated in grading by more than relevant performance. This specific concern was not evidenced in the course. The competition between the students was healthy and this did not affect the grades. It was detected by university tutors that the students who got the best grades within the group actually excelled in their activities.

When peer assessment is used in a high stakes' setting, it results in inflated estimates of performance and few below average evaluations. Although this evaluation was part of the students' final grade; it didn't influence the decision taken by the students once it composed only 20% of the final grade.

Finally, one of the fundamental issues in peer assessment is the evaluation of a student is equivalent to that of his or her colleagues. However, the self-assessment wasn't considered in this course, therefore the comparison between the student evaluation and the remaining wasn't analyzed. The analysis performed is in the comparison of the evaluation received by each student and the conclusions obtained by tutors.

7.2. Analysis of the groups' presentation assessment by tutors

According to the previous description, the grades attributed to the four students' groups determine that G2 was the group that stood out among the others. Even the evaluation's correlation degree among tutors wasn't extremely strong due to the lack of standardization in answering the questionnaire; it was possible to detect the difference between the grades received by each group. The groups G3 and G4 had exactly the same evaluation due to their shortcomings in the project conducting.

The lack of standardization in the questionnaire's answering was evidenced by the agreement analysis which uses the Kappa's coefficient in order to define the strength of agreement among attributed grades by tutors. The tutor T4 was considered the standard tutor since who is the most experienced in the problem area; however, its grades attributed to the 16 criteria of evaluation stablished a slight agreement with the grades attributed to the same 16 criteria by the remaining tutors. This lack of standardization is also visualized by the correlation analysis performed for scores attributed to those same 16 criteria.

8. Conclusion

After the discussion of results, research leads to the following conclusions:

(i) The statistical analysis indicated a possible deficiency of the assessment method adopted since the evaluation criteria weren't adequately explained to the evaluators - tutors and students. The poor agreement between tutors on assessing students and between students on evaluating their peers, as well as the moderate an low correlation between the grades assigned by tutors to groups and by students to their teammates, indicate that tutors and students have not equally evaluated groups within each criterion. This difference shows a misalignment between tutors and students in understanding the criteria or the reasons for diminishing the grades;

(ii) The statistical analysis was also useful to understand the grades obtained by each student and group. The information from analysis can be used as feedback to students and their groups;

(iii) The research contributes positively to the improvement of courses based on active learning, since the statistical analysis allow outline some conclusions about the procedures for assigning grades to students and, thus, to propose improvements in the assessment method adopted. Furthermore, it is believed that a consistent and equitable method of assessment may stimulate interest and participation of students and hence their learning;

(iv) The peer mode proposed as one of the course mode provided the vision that the students had about their teammates. The students' view obtained through the grades coincided with the same perception that tutors had even with a look outside.

Although the questionnaires adopted by the assessment method has provided data that representing the real participation of students in the course, the way that they were applied left to be desired. A possible solution to the deficiency of the assessment method adopted application is to conduct a prior discussion between the evaluators about the meaning of each criterion used in the assessment in order to allow an equal understanding among everyone about: (1) what is expected from the student according to that criterion; (2) what would be the maximum that the student could achieve; and, (3) what are reasons to discount some points of the student's grade, according to the questionnaire scale. This would avoid each tutor using their own understanding of each evaluation criterion on the assessment as well as the manner of grading students.

Considering a future research, this course in partnership with the multinational company signed an agreement so that the projects are carried out over two years and can be extended every two years. Thus, each year new projects and new teams will be part of course. As the projects will be new, the professionals linked directly to the company's areas will be the new tutors and the number of groups can be changed according to the number of projects previously defined by the company. The university's tutors will also be changed in order to several professors of Industrial Engineering course have contact with this project. Thereby, the proposed improvements to the assessment method can be carried out, and if new problems arise, they can be improved in future editions of the partnership.

Acknowledgements

The authors thank FAPEMIG, CAPES, and CNPq for their support to this research. Also, it is important to emphasize gratitude to the University and to the company for the opportunity to participate in the design and implementation of active learning course.

References

Agrawal, D. K., & Khan, Q. M. (2008). A quantitative assessment of classroom teaching and learning in engineering education. *European Journal of Engineering Education*, *33*(1), 85-103. http://dx.doi.org/10.1080/03043790701746389.

Al-Nashash, H., Khaliq, A., Qaddoumi, N., Al-Assaf, Y., Assaleh, K., Dhaouadi, R., & El-Tarhuni, M. (2009). Improving electrical engineering education at the American University of Sharjah through continuous assessment. *European Journal of Engineering Education*, *34*(1), 15-28. http://dx.doi.org/10.1080/03043790802710169.

Ayatollahi, S. M. T., Bagheri, Z., & Heydari, S. T. (2013). Agreement analysis among measures of thinness and obesity assessment in Iranian school children and adolescents. *Asian Journal of Sports Medicine*, *4*(4), 272-280. PMid:24800002. http://dx.doi.org/10.5812/asjsm.34247.

Brown, S. (2004). Assessment for Learning. *Learning and Teaching in Higher Education*, *1*, 81-89. Retrieved in 28 August 2016, from http://www2.glos.ac.uk/offload/tli/lets/lathe/issue1/articles/brown.pdf

Chua, K. J., Yang, W. M., & Leo, H. L. (2014). Enhanced and conventional project-based learning in an engineering design module. *International Journal of Technology and Design Education*, *24*(4), 437-458. http://dx.doi.org/10.1007/s10798-013-9255-7.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37-46. http://dx.doi.org/10.1177/001316446002000104.

Donohue, S. (2014). Supporting active learning in an undergraduate geotechnical engineering course using group-based audience response systems quizzes. *European Journal of Engineering Education*, *39*(1), 45-54. http://dx.doi.org/10.1080/03043797.2013.833169.

Eva, K. W. (2001). Assessing tutorial-based assessment. *Advances in Health Sciences Education: Theory and Practice*, *6*(3), 243-257. PMid:11709638. http://dx.doi.org/10.1023/A:1012743830638.

Feletti, G. I., & Sanson-Fisher, R. W. (1983). Measuring tutor ratings in relation to curriculum implementation. *Higher Education*, *12*(2), 145-154. http://dx.doi.org/10.1007/BF00136633.

Goodyear, P., Jones, C., Asensio, M., Hodgson, V., & Steeples, C. (2005). Networked learning in higher education: student's expectations and experiences. *Higher Education*, *50*(3), 473-508. http://dx.doi.org/10.1007/s10734-004-6364-y.

Heubert, J. P., & Hauser, R. M. (1999). *High stakes: testing for tracking, promotion, and graduation*. Washington: National Academy Press.

Kritikos, V. S., Woulfe, J., Sukkar, M. B., & Saini, B. (2011). Intergroup peer assessment in problem-based learning tutorials for undergraduate pharmacy students. *American Journal of Pharmaceutical Education*, *75*(4), 73. PMid:21769149. http://dx.doi.org/10.5688/ajpe75473.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174. PMid:843571. http://dx.doi.org/10.2307/2529310.

Lee, J., Imanaka, Y., Sekimoto, M., Nishikawa, H., Ikai, H., & Motohashi, T. (2011). Validation of a novel method to identify healthcare-associated infections. *The Journal of Hospital Infection*, *77*(4), 316-320. PMid:21277647. http://dx.doi.org/10.1016/j.jhin.2010.11.013.

Montgomery, D. C. (2012). *Design and analysis of experiments*. United States of America: John Wiley & Sons.

Montgomery, D. C., & Runger, G. C. (2013). *Applied statistics and probability for engineers*. United States of America: John Wiley & Sons.

Mückenbergera, E., Togashib, G. B., Páduac, S. I. D., & Miurad, I. K. (2013). Process management applied to the establishment of international bilateral agreements in a Brazilian public institution of high education. *Production*, *23*(3), 637-651. http://dx.doi.org/10.1590/S0103-65132012005000076.

Norcini, J. J. (2003). Peer assessment of competence. *The Metric of Medical Education*, *37*(6), 539-543. PMid:12787377. http://dx.doi.org/10.1046/j.1365-2923.2003.01536.x.

Norton, L., Norton, B., & Shannon, L. (2013). Revitalizing assessment design: what is holding new lecturers back? *Higher Education*, *66*(2), 233-251. http://dx.doi.org/10.1007/s10734-012-9601-9.

Phillips, R. (2005). Challenging the primacy of lectures: the dissonance between theory and practice in university teaching. *Journal of University Teaching and Learning Practice*, *2*(1), 1-12. Retrieved in 28 August 2016, from http://ro.uow.edu.au/jutlp/vol2/iss1/2

Raud, Z. (2010). Active learning power electronics: a new assessment methodology. In *Proceedings of the International Power Electronics and Motion Control Conference*, Ohrid, Macedonia.

Roberts, M. J., & Russo, R. (1999). *A student's guide to analysis of variance*. New York: Routledge.

Sambell, K., Mcdowell, L., & Brown, S. (1997). But is it fair? An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*(4), 349-371. http://dx.doi.org/10.1016/S0191-491X(97)86215-3.

Shahadan, T. N. T., Shafie, N., & Liew, M. S. (2012). Study on subject's assessment for students' active learning in a private institution. *Procedia: Social and Behavioral Sciences*, *69*, 2124-2130. http://dx.doi.org/10.1016/j.sbspro.2012.12.176.

Shankland, R., Genolini, C., França, L. R., Guelfi, J. D., & Ionescu, S. (2010). Student adjustment to higher education: the role of alternative educational pathways in coping with the demands of student life. *Higher Education*, *59*(3), 353-366. http://dx.doi.org/10.1007/s10734-009-9252-7.

Shekar, A. (2007). Active learning and reflection in product development engineering education. *European Journal of Engineering Education*, *32*(2), 125-133. http://dx.doi.org/10.1080/03043790601118705.

Struyf, E., Vandenberghe, R., & Lens, W. (2001). The evaluation practice of teachers as a learning opportunity for students. *Studies in Educational Evaluation*, *27*(3), 215-238. http://dx.doi.org/10.1016/S0191-491X(01)00027-X.

Timossi, L. S., Francisco, A. C., Santos Junior, G., & Xavier, A. A. P. (2010). Analyis on quality of work life of employess with different levels of education through an analysis of correlation. *Production*, *20*(3), 471-480. http://dx.doi.org/10.1590/S0103-65132010005000031.

Vieira, F. H. A., & Francisco, A. C. (2012). Corporate education implementation stages and their impacts on Brazilian companies: a multi-case study. *Production*, *22*(2), 296-308. http://dx.doi.org/10.1590/S0103-65132012005000018.

Waters, R., & Mccracken, M. (1997). Assessment and evaluation in problem-based learning. *IEEE Xplore Digital Library*, *2*, 689-693. http://dx.doi.org/10.1109/FIE.1997.635894.

Willey, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education*, *35*(4), 429-443. http://dx.doi.org/10.1080/03043797.2010.490577.

Received: Aug. 28, 2016
Accepted: June 5, 2017