International Journal of Electrical Engineering Education
XX(X):2–15

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Reliability score evaluation of continuous assessment tests: a longitudinal study

Jorge E. López de Vergara[1] and Ricardo Olmos[2]

Abstract

In this paper, a longitudinal study is presented about the score reliability in continuous assessment tests of the subject Network Architecture II, which is taught in the Telecommunication Technologies and Services Engineering degree, offered at Universidad Autónoma de Madrid, Spain. It is important to evaluate if scores in tests are reliable, because it shows if the phenomenon under study is measured with precision and low error, which are necessary conditions for a fair assessment.

Thus, an analysis is provided about the scores obtained in the twenty-eight continuous assessment tests taken across seven years. Correlation, corrected correlation and Cronbach's α are used as psychometric indicators. Results show that, in general, there is a high correlation in the students' scores among the different assessment tests every year, which let us confirm that they have been correctly set out. Additionally, the results are compared among different years, showing that they are similar in the different cohorts. However, variations have also been observed over the years, with tests with poor correlation.

Finally, based on the analysed data, a novel method is proposed to early detect problems in the tests' evaluation before the course ends, by correlating their scores in pairs. A low correlation between two tests in the same year, empirically below 0.2, would imply issues in the evaluation.

Keywords

Score correlation, corrected correlation, Cronbach's α, internal consistence, test pair correlation.

Introduction

With the introduction of the European Higher Education Area degrees, subjects have changed to use methodologies where a continuous assessment is applied[1]. These methodologies enable the instructor to follow the students work during the course, which should result in a better education quality. However, it is important to review if such continuous evaluation has been adequate, or, on the contrary, it has had any deviation not easily identifiable. For this, techniques proposed in Psychometry[2] can be leveraged to study the reliability of the students' scores.

Thus, this paper brings up the need to deep analyse the scores obtained by the students in the subject named Network Architecture II, taught in the Telecommunication Technologies and Services Engineering degree, offered at Universidad Autónoma de Madrid since 2012. The subject, of 6 ECTS credits, is taught in the spring semester of the second year. Four topics are covered: Queuing Theory, Data-link Layer, Wireless and Mobile Networks, and Network Security and Management. Since its inception, the subject has followed exactly the same teaching and assessment methodology, with the same professor in charge of it since 2012.

The continuous assessment methodology used in the theoretical part of the subject is based on other existing proposals[3]. Basically, it consists in carrying out four mid-term tests, one per topic, throughout the semester, combined with problem solving activities, which are presented by the students in the classroom. It is possible to pass the subject without taking a final exam if all mid-term tests are also passed, being the final subject a mean of the tests—20% each— and the problem solving activities—another 20%—.

If a student fails one of the tests, then the student has to do the final exam, which covers all the topics of the subject, both theory and problems. However, in this case, students can still follow the continuous assessment, because it can improve the final grade. In this last case, the continuous assessment score is worth a 40% of the final grade. These weights—40%-60%—allow assessing that the student has worked over the year, and that all the topics—both theory and problems—have been adequately studied for the final exam. If the final exam score is higher than this weighing, then the final grade is directly that one obtained in the final exam. If a student has left the continuous assessment, he or she can still take the final exam, which is also worth a 100% of the final grade. All

[1]Dept. Electronics and Communications Technologies, Faculty of Engineering, Universidad Autónoma de Madrid, Spain
[2]Dept. Social Psychology and Methodology, Faculty of Psychology, Universidad Autónoma de Madrid, Spain

Corresponding author:
Jorge E. López de Vergara, Dept. Tecnología Electrónica y de las Comunicaciones, Escuela Politécnica Superior, Universidad Autónoma de Madrid, Calle Francisco Tomás y Valiente, 11, E-28049 Madrid, Spain
Email: jorge.lopez_vergara@uam.es

Prepared using sagej.cls

these weighing rules stress the need to evaluate the reliability scores of the continuous assessment tests, to avoid unfair situations.

To reduce the subjectivity in assessment, mid-term tests are set as multiple-choice quizzes, with four possible options, only one option is valid, with a $^{-1}/_{3}$ penalty for wrong answers, following the common indications for this type of evaluation[4]. The first test is about solving a Queuing Theory problem together with some theoretical questions, lasting 50 minutes. This first test, given that is linked to the problem to solve, has between 12 and 15 questions. The other three tests related to the remaining topics have 20 questions to solve in 30 minutes, about small exercises and theoretical questions. Scores range from 0 to 10. A test is failed if the score is below 5, being 5 a pass and 10 an excellent.

In this paper, four questions are addressed: *Do mid-term tests correlate with the final grade? Do mid-term test results have consistency every year? How is the subject evolving across the years? Has difficulty substantially varied over the years?* To answer these questions, the following points are analysed:

1. The correlation between mid-term tests with respect to the final score, and the score without taking into account each test—namely, corrected correlation—, for those students that have taken all continuous assessment tests. It is important to remark that, with respect to the total number of enrolled students, many of them give up from the continuous assessment itinerary, so it is not possible to count with all students for the study, which will generate some bias in the analysis, underestimating the reliability of the set of tests. Figure 1 shows a comparison between enrolled students and those that completed the continuous assessment itinerary across the seven years of the study. As shown, the number of students that have given up vary every year, which is also going to affect the mean score of each year, as it will be shown in the following sections.
2. Additionally, it is studied if there is consistency in the set of continuous assessment tests taken every year. For this, the Cronbach's α coefficient[5] is used.
3. The evolution of the subject over the years, so it can be seen if the correlations obtained previously are more or less stable among different years or cohorts.
4. In addition, it is analysed if the subject difficulty has substantially varied, by studying the performance in the four tests taken every year.

Based on the results of this analysis, a novel method is proposed to early detect problems in the test evaluation before the years ends, by correlating the scores in test pairs. This method enables to identify if the scores of a test do not correlate well with others, identifying any assessment problem before the course ends. This novel contribution is a major improvement over our previous work[6] on the same topic.

In order to show that this case study can be sufficiently representative to gain insight on the assessment of continuous evaluation, Figure 2 provides data about the performance over the years of the whole 4-year degree and the subject after final exams. Performance is calculated as the number of passed credits divided by the number of total enrolled credits. Their mean values over the years are similar—76.37 for the degree and 78.08 for the subject—. As shown in the figure, the performance of the subject is better than

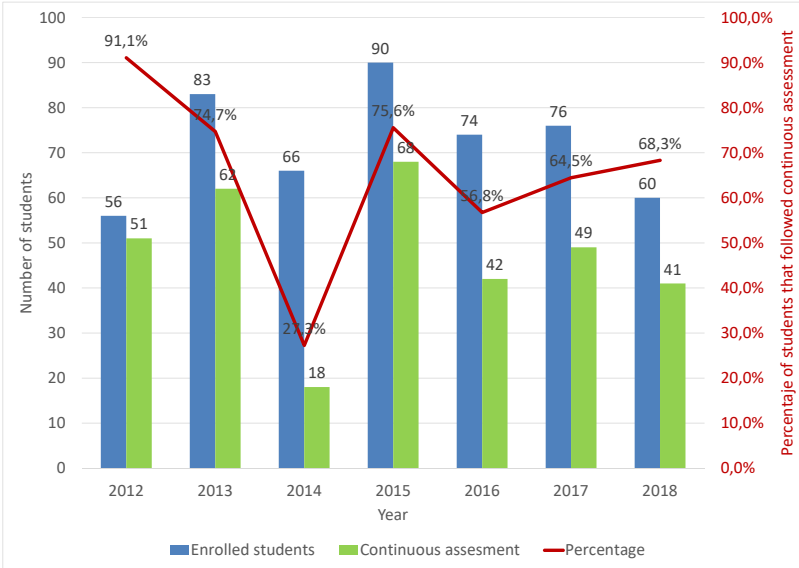


Figure 1. Student population under study and percentage of those that completed the continuous assessment itinerary for the different studied years.

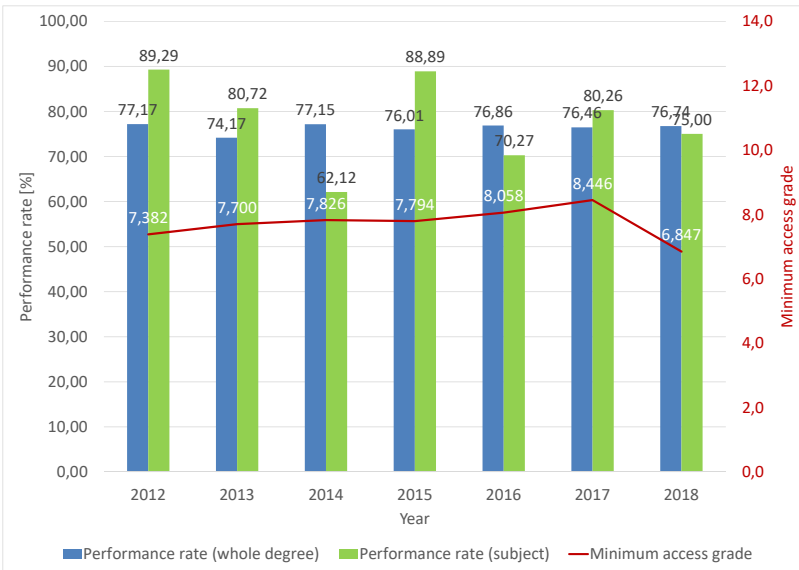


Figure 2. Students' performance in the degree and in the subject under study for the different studied years, and minimum access grade of this degree in the previous year.

the degree in those years in which most students followed the continuous assessment, with a high correlation—0.92—between subject performance and percentage of students completing the continuous assessment. To have a measure of the proficiency of each cohort, Figure 2 also shows the minimum access grade* to enter the degree the year before, this is, the lowest grade of the students that could enrol the degree in each cohort—70 new students access the degree every year—. These grades do not show much correlation with the students' performance in the subject.

Given the high correlation between following continuous assessment and students' performance, evaluating the score reliability of the tests is a must. It is worth mentioning that the score reliability partially shows if the assessment is fair, because it informs if the phenomenon under study is measured with precision and low error. To consider that an assessment is fair, it is necessary that it has high reliability—however, it is not sufficient, because the reliability does not guarantee that the assessed concepts represent well the content universe that have to be evaluated in a subject—. The assessment of a subject with low reliability will have a score variability due to issues not related to the students' knowledge level, which causes an unfairness situation[7].

Thus, the answers to the questions here raised will let us have a formal methodology to study the reliability of the continuous assessment process in this subject, as well as in other ones with similar characteristics. Moreover, the proposed detection method provides a way to identify problems of this type before the course ends, letting the teaching team to correct it during the course. After looking in the bibliography, it has not been found any study similar to this one, or at least, as specific, in the scope of education in Electrical and Electronic Engineering.

As related research, it is important to cite the work from McKenzie and Schweitzer[8], where they looked for factors that could predict the academic performance in university students with a longitudinal study, and where they observed that the academic performance is much correlated. This is, the usual behaviour is that each student has similar scores over the years. With respect to the methodology used in this paper, the work in Galván-Sánchez *et al.*[9] also shows the use of correlations and Cronbach's α. However, they use it to measure the reliability of rubrics to assess oral presentation skills, and not to review the scores in continuous assessment tests.

To accomplish with the outlined work, the paper first presents in next section the psychometric indicators used in this study. Then, the correlation among the scores of the four mid-term tests is analysed, and after this, its evolution over the different years is studied as well. Later, based on the analysed results, the method to early detect tests with low reliability is proposed. Finally, all obtained results are discussed, providing the conclusions of the analysis.

*A scale from 0 to 14 is used in Spain for the grades to access to the university, being a 5.0 the lowest value to pass. Scores from 10 to 14 can be obtained with additional exams of subjects related to the degree that the student wants to study.

Psychometric indicators

Before doing the proposed analysis, and to better understand the fundamentals in which it is based, it is necessary to present the psychometric indicators that are used.

The first of these indicators is the Pearson's correlation coefficient, which is defined as:

$$r_{XY} = \frac{\sum_{i=1}^{N}(x_i - \mu_X)(y_i - \mu_Y)}{N \cdot \sigma_X \cdot \sigma_Y} \tag{1}$$

where X and Y are the random variables to compare—in this case, test scores—, x_i and y_i are individual samples of these variables—here, scores of the student i—, μ_X and μ_Y, their expected values, N the number of elements—students that followed the continuous assessment—, and σ_X and σ_Y are standard deviations of X and Y, respectively.

Pearson's correlation coefficient takes values in the interval $[-1, 1]$. The more its value is close to 1, the higher direct relationship between the random variables, and when closer to 0, less linear relationship is expected between both variables. If the correlation is negative, the relationship between the variables will be inverse.

In this work, the correlation coefficient is used to compare the scores of each continuous assessment test with the final grade, obtained as a mean of the four mid-term tests taken every year. Additionally, the corrected correlation is also used, where each test scores are compared with a mean of the other three test scores. This is, the score of the test to compare is taken out of the global score. From a psychometric viewpoint, it is recommended that the correlation values should be greater than 0.2 [10]. It should be noted that this value is usually used for item discrimination, not for test comparison. Anyways, we think that is a proper threshold also for our study, because it shows a modest correlation, pointing to problems in the test reliability below this value.

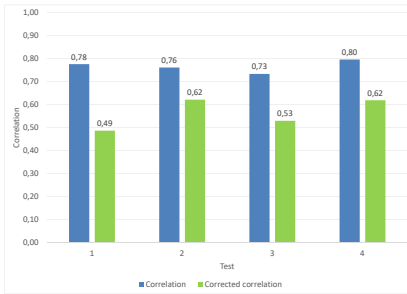
The second indicator is the Cronbach's α [5], defined as follows:

$$\alpha = \frac{J}{J-1}\left(1 - \frac{\sum_{j=1}^{J}\sigma_j^2}{\sigma_X^2}\right) \tag{2}$$

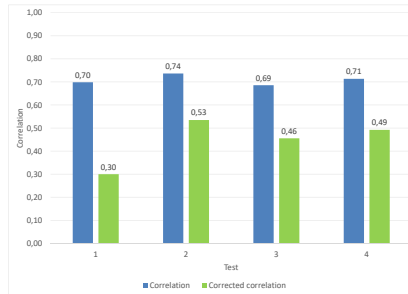
where J maps to the number of assessment tests—in this case study, $J = 4$—and σ_j^2 the variance of the assessment test j and σ_X^2 the variance of the global continuous assessment grade. From a psychometric viewpoint, a value of α greater than 0.7 is necessary to have a good assessment test consistency [11].

Not all experts agree on 0.7 as a good reference point to assess the reliability of a test [12]. Obviously, this threshold will depend on the context, the consequences and the implications that the measurement instrument has for a person—e.g., it is not the same an evaluation to enter a job than another one to evaluate the quality of a product—. A university test can be considered a high-stakes test, that is, an evaluation with high consequences, so the evaluators have to be rigorous when setting the minimum admissible value of α.

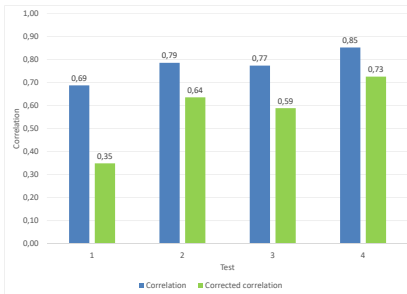
Finally, the third indicator is the difficulty index, which is obtained as the mean of the scores on each test. If the mean is high, the test was easy to solve. On the contrary, a low value means that the test was difficult.



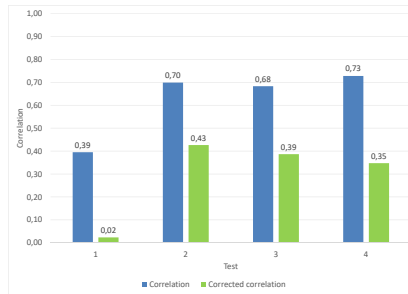
(a) Year 2012.



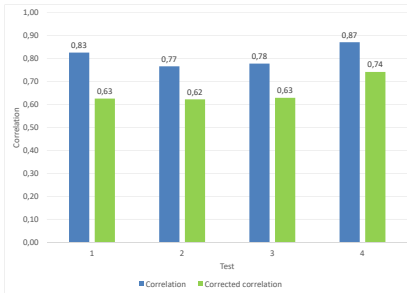
(b) Year 2013.



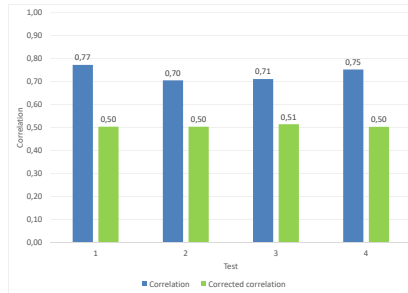
(c) Year 2014.



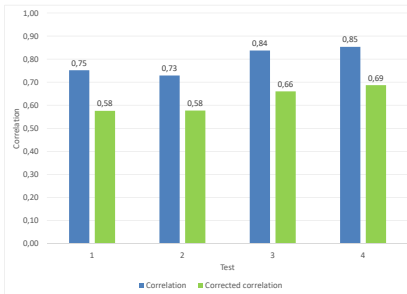
(d) Year 2015.



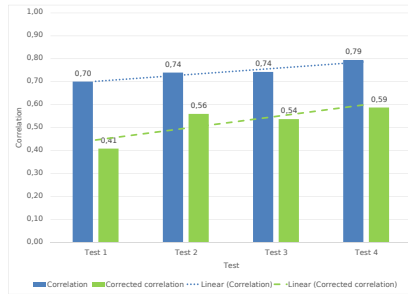
(e) Year 2016.



(f) Year 2017.



(g) Year 2018.



(h) Mean over the years.

Figure 3. Score correlation of each test over the years.

Correlation among the scores of the different tests

To answer the first question raised in the introduction, Figure 3 shows the result of obtaining the correlation and corrected correlation—this is, excluding each test from the mean to calculate its correlation with the other tests—of the four continuous assessment tests over the years, applying Eq. 1. Figure 3(h) shows the mean across the seven years, revealing the trend that the correlation and corrected correlation is lower in the first test and higher in the fourth one. Probably, this is caused because at the end of the course, the students have a clear view on how they are evaluated and the difficulty of the subject, after passing the other tests.

As it can be shown, in general, the scores of the different tests have a high correlation with the mean score. Being this correlation around 0.7, this is a good result from a psychometric viewpoint with respect to how the continuous assessment has been done. Corrected correlation is high as well, around 0.5, although it has higher variation depending on the year.

There is only one case in the twenty-eight analysed mid-term tests, the first one in year 2015, which correlations are low. This fact has brought the need to study this test in depth. Figure 4 shows a scatter plot where this phenomenon can be seen in detail. It represents the scores of this test with respect to the mean grade of all tests, and the mean grade of the other three tests, excluding this test scores.

As it can be inferred from the points with the cross shape, which represent the obtained score in the test with respect to the mean grade of the other three tests, there is not any

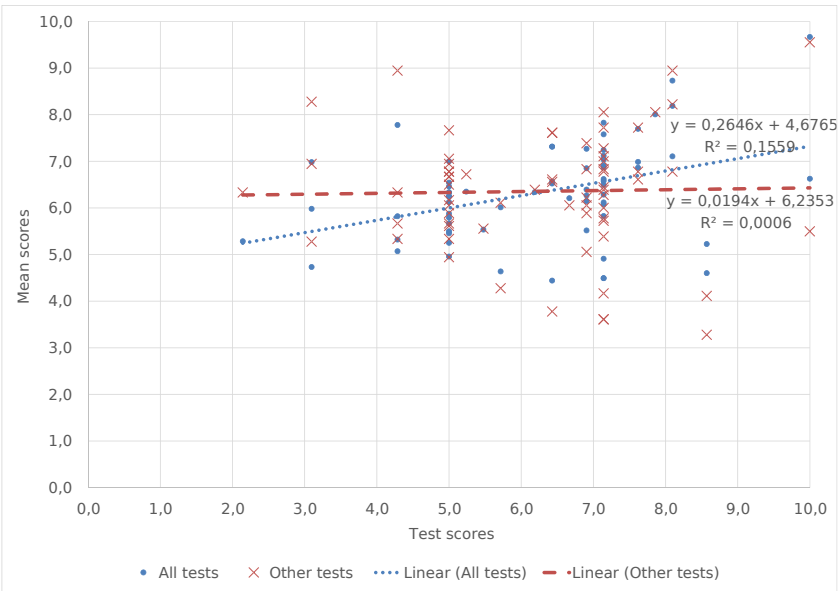


Figure 4. Scatter plot for test 1 in year 2015 with respect to all tests scores and the other test scores.

clear trend. In fact, if a regression is done, a practically flat slope is obtained and a determination coefficient R^2 near to zero, which indicates that the score obtained in this test does not explain the scores of the other continuous assessment tests in that year.

The possible causes of this fact could be the following ones:

- Being a continuous assessment test, it was done at lecture times, and there was another test of other subject an hour before. This probably caused that some students did not study this test sufficiently, unlike the other continuous assessment tests this year.
- In a complementary way, being this test the first one in the year, the students could be still probing the subject difficulty, and they did not valued it appropriately. This fact is also reflected in the rest of the courses, as stated above.

Evolution over the years

To know how the difficulty of the subject has evolved over the years, several measurements have been taken. The first one is that of the mean scores of the continuous assessment tests for the students that have completed all the course tests, and the resulting mean grade. Table 1 details these values. To understand them better, it is also necessary to compare them with the number of students that completed the four mid-term tests each year, as presented before in Figure 1.

As shown, the scores were evolving positively until year 2014. In that year, less students followed the continuous assessment, being then the year with higher bias. Later, scores have been declining in the following years. In the last three years, scores show that several students continued with the continuous assessment despite they failed one or

Table 1. Mean score of the continuous assessment tests. Standard deviation is shown in parenthesis. Minimum access grade is also shown for comparison.

Year	2012	2013	2014	2015	2016	2017	2018
Test 1	6.58	6.90	7.05	6.26	5.75	4.88	5.39
	(2.48)	(2.31)	(1.85)	(1.54)	(2.48)	(2.27)	(1.48)
Test 2	5.54	5.50	5.44	5.86	5.25	5.99	5.45
	(1.44)	(1.51)	(1.22)	(1.49)	(1.69)	(1.65)	(1.26)
Test 3	7.23	7.06	6.98	7.38	5.52	5.73	4.85
	(1.80)	(1.53)	(1.38)	(1.55)	(1.79)	(1.64)	(1.84)
Test 4	5.75	6.82	7.49	5.84	4.83	5.18	4.39
	(1.86)	(1.55)	(1.33)	(1.95)	(2.17)	(2.09)	(1.85)
Mean score	6.27	6.57	6.74	6.33	5.34	5.45	5.02
	(1.45)	(1.22)	(1.11)	(1.03)	(1.66)	(1.41)	(1.29)
Min. access grade	7.382	7.700	7.826	7.794	8.058	8.446	6.847

more of the tests. This fact would explain the fall in the global mean grade, as well as the mean scores under 5.0 in some tests, being reduced also the bias initially raised.

It could be argued that a high test score mean could also point to the cohort being more proficient, but this is not necessarily the case, given that the number of students that enter to study the degree every year is the same, and there are not high fluctuations in their access grades among years. To support this statement, we can analyse the minimum access grades, previously presented in Figure 2, and shown as well in Table 1 for clarity. In the first three years both scores increase, but the minimum access grade cannot explain the decrease in the following years. For instance, in 2017 the students had the highest minimum access grade, but also one of the lowest mean scores in the subject. Hence, the correlation between the mean scores and the minimum access grades is low, with a value of 0.11.

Anyway, the study of the reliability of the test scores does not depend on the proficiency of the cohorts, but on the standard deviation of the scores[13], to know that each test has been able to discriminate among different levels of knowledge. The standard deviation values over the seven years of the study show that all tests fulfil this purpose. Moreover, with respect to mean scores and their standard deviation, it can be observed that they are not able to identify the singularity of year 2015, discovered when using the correlation between the first test and the others.

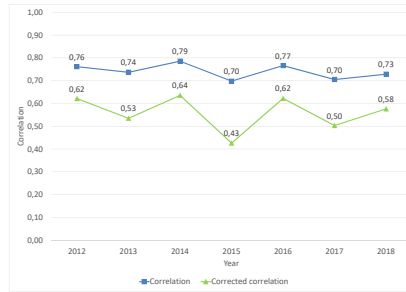
Regarding the temporal evolution of the correlations, Figure 5 shows how they have varied over the years for each test. It can be said that the evolution for all four mid-term tests has been quite similar, existing the singularity explained in previous section in year 2015, where test 1 is uncorrelated with respect to the other tests. Additionally, the fact that the first test is taken at the beginning of the course, and that it is more focused on solving a problem, causes that the corrected correlations in this test are usually lower than the rest.

Nevertheless, in general, there exists a high correlation in the scores. To corroborate this fact and have a measure of the internal consistence among the four mid-term tests, the Cronbach's α, defined in Eq. 2, has been calculated for every year. This result is provided in Table 2 for all years. As reflected in the table, the consistence is high, above 0.7 threshold for five of the seven years under study. Year 2015 has the lower α value, being this fact coherent with the correlation values explained before.

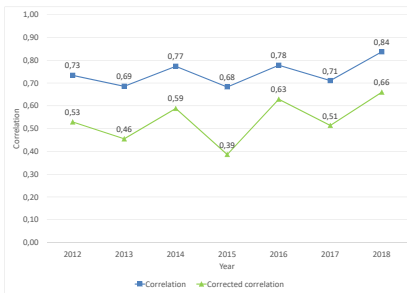
Early detection based on test pairs correlation

Although the longitudinal study provides interesting insights about the reliability of the test scores, the methodology followed until now poses a problem, which is that the instructor has to wait until the course ends to work out the psychometric indicators. Thus, we propose a novel method to early detect the reliability of tests, using again Pearson's correlation, but comparing the results of every pair of test, immediately after their results are available. This method provides instructors a way to detect problems before the course ends, so they can take corrective actions when needed. For instance, to

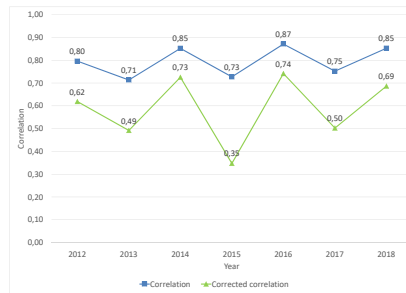
(a) Test 1: Queuing Theory.



(b) Test 2: Data-link Layer.



(c) Test 3: Wireless and Mobile Networks.



(d) Test 4: Network Security and Management.

Figure 5. Evolution of correlation and corrected correlation for all tests over the years.

solve this issue when it is identified, it is possible to do a detailed and more costly item-by-item psychometric assessment[14] of the test with problems, annulling those questions with low reliability, or even taking a second-chance exam if necessary.

This method is based on the high correlation that has been found between different tests in the same year during the seven years of this study. Thus, if a low correlation is found between two tests, at least one of them will have low reliability. To identify it, it is necessary to find another test with low correlation with one of these identified before. In the studied case, given that four tests are taken across the semester, such test will be easily identified before the course ends.

Following this idea of correlating every pair of test results, the final number of correlations C to be done on J tests is:

$$C = \binom{J}{2} = \frac{J(J-1)}{2} \tag{3}$$

This is, the number of possible combinations of a pair of tests results chosen from J. For $J = 4$—the case of this study—, $C = 6$, which is a reasonable number of correlations and comparisons. As an example, Table 2 shows the correlations of every test pair every year in the study.

Table 2. Cronbach's α and correlation values for pairs of tests.

Year	α	Correlation test pair {1,2}	{1,3}	{2,3}	{1,4}	{2,4}	{3,4}
2012	0.75	0.43	0.34	0.50	0.44	0.56	0.49
2013	**0.64**	0.35	**0.17**	0.39	**0.20**	0.43	0.54
2014	0.75	0.33	0.25	0.52	0.33	0.70	0.70
2015	**0.49**	**0.20**	**-0.05**	0.32	**-0.06**	0.30	0.45
2016	0.82	0.49	0.42	0.58	0.65	0.52	0.62
2017	0.71	0.37	0.43	0.36	0.36	0.43	0.38
2018	0.80	0.42	0.51	0.47	0.50	0.54	0.61

* $\alpha < 0,7$ and correlation $\leq 0,2$ are marked in **bold** typeface.

The method follows the algorithm below, which has to be run every time a new test result is obtained, once there are at least two tests to correlate, and until the course ends:

1. Calculate the correlation of the scores with prior available test results.
2. If correlation is less than 0.2 for any pair of tests—following the threshold criterion presented in Mullis *et al.* [10]—, find another test with low correlation with any of this test pair:
   (a) If there are more tests with low correlation with one of the test pair, this test is identified as low reliable.
   (b) If there are no other tests with low correlation, wait for next test results to identify which one is low reliable.
3. Store the correlation results for the next iteration with another test results.

In order to check the algorithm, Table 2 values are assessed. For instance, in year 2015, with a Cronbach's $\alpha = 0.49$, the first correlation pair shows a value of 0.20, so any of these two tests are suspect to be unreliable. However, it is necessary to wait until the third test results are available to identify which test was unreliable. This is clearly identified when the pair of tests 1 and 3 show a correlation of -0.05, which points to test 1 as it has low correlation with the other two tests. Finally, the correlation of tests 1 and 4 also corroborates that the test 1 can be identified as unreliable.

The other year with $\alpha < 0.7$ is 2013. In this case, the first observed low correlation is between the tests 1 and 3, which is later repeated with tests 1 and 4, pointing again to test 1 as low reliable. This case is a bit weird, because the correlation of the tests 1 and 2 is not low, neither it is between tests 2 and 3 nor between tests 2 and 4. However, in previous section, it showed a corrected correlation of 0.3, the second lowest in the longitudinal study, which confirms that this algorithm is working properly.

Conclusions

To finish this study related to the reliability of scores in continuous assessment tests in the subject under study, and in order to answer the questions raised in the introduction, the following findings have been identified:

1. In general, there is a high correlation among mid-term tests with respect to the mean grade, as well as with the corrected grade obtained by excluding each test to calculate the correlation. Only one of the twenty-eight mid-term tests was initially found in which the corrected correlation could not be considered adequate. In general, the last test has a higher correlation with the final grade and with respect to the other three tests. On the other hand, the first test usually has a lower correlation with the rest of the scores, even though most years are above the defined threshold.
2. Cronbach's α has shown that there is internal consistency in the set of tests taken each year for continuous assessment, except in 2015, where a singular case has been identified, reducing α value.
3. In general, apart from the singular case, correlations obtained for the different tests over the years are very similar. It has to be taken into account that correlations—and measurements that depend on them, such as Cronbach's α—have a higher sampling error than other statistical instruments, such as means or proportions [15]. Thus, fluctuations in this indicator can be partly due to sampling effects if the sample is small—i.e., $N < 100$—, like it happens in all studied years for this subject.
4. Additionally, it seems that the difficulty on the continuous assessment tests has risen in the last years. However, this can be caused because many students have taken all tests, despite they have failed one or more of them. This did not happen in 2014, where those with a failed test abandoned the continuous assessment, so all students in the study passed all tests in that year. This explains also that 2014 is the year with best final grade, due to the introduced bias.

This evaluation method was initially performed for the first five years, and then, with the last two years, once new data has been available each year after the initial assessment. With this, it has been checked that the proposed methodology allows a rigorous study that deeply analyses the results of every year for the same subject. At the same time, it is relatively easy to implement with a spreadsheet, without needing sophisticated statistical tools—the spreadsheet used in this work is available on demand—. Moreover, the results provide an additional support to instructors against a possible claim of the students' grades.

However, to obtain the psychometric indicators it was initially necessary that all test results of the year were available, so another method has also been defined without that restriction to early detect tests with low reliability. This method has been tested against available data using the common correlation threshold of 0.2, easily detecting the singularity in 2015, and pointing as well to another test in 2013, the second test with low corrected correlation in the whole dataset. The obtained results show that this method can be useful to evaluate the reliability of continuous assessment test scores.

Acknowledgements

Authors would like to thank Universidad Autónoma de Madrid (UAM) for organising the learning assessment course "Evaluación del Aprendizaje," given in its Teaching Training Program, which led to the development of this study. They also thank the Vice-dean of Teaching Quality in the

Engineering School at UAM for providing performance data of the whole degree. Finally, the authors would also like to thank the anonymous reviewers for their valuable comments, which helped to improve the final version of the paper.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

1. Cano M. Students' involvement in continuous assessment methodologies: A case study for a distributed information systems course. *IEEE Transactions on Education* 2011; 54(3): 442–451. DOI:10.1109/TE.2010.2073708.
2. Crocker L and Algina J. *Introduction to classical and modern test theory*. Orlando, FL: Holt, Rinehart and Winston, 1986.
3. Vaezi-Nejad SM and Olabiran Y. Telematics education II: Teaching, learning and assessment at foundation level. *International Journal of Electrical Engineering Education* 2005; 42(2): 147–163. DOI:10.7227/IJEEE.42.2.3.
4. Burton RF. Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment and Evaluation in Higher Education* 2001; 26: 41–50. DOI:10.1080/02602930020022273.
5. Cronbach LJ. Coefficient alpha and the internal structure of tests. *Psychometrika* 1951; 16(3): 297–334. DOI:10.1007/BF02310555.
6. López de Vergara JE and Olmos R. Estudio longitudinal de las calificaciones de evaluación continua en la asignatura de Arquitectura de Redes II del grado en Ingeniería de Tecnologías y Servicios de Telecomunicación. In *Proceedings XIII Jornadas de Ingeniería Telemática - JITEL'2017*. pp. 333–339. DOI:10.4995/JITEL2017.2017.6496.
7. Wells C and Wollack J. An instructor's guide to understanding test reliability. Technical report, Testing & evaluation Services. Univ. Wisconsin, 2003.
8. McKenzie K and Schweitzer R. Who succeeds at university? Factors predicting academic performance in first year Australian university students. *Higher education research & development* 2001; 20(1): 21–33. DOI:10.1080/07924360120043621.
9. Galván-Sánchez I, Verano-Tacoronte D, González-Betancor SM et al. Assessing oral presentation skills in electrical engineering: Developing a valid and reliable rubric. *International Journal of Electrical Engineering Education* 2017; 54(1): 17–34. DOI:10.1177/0020720916659501.
10. Mullis IVS, Martin MO and Diaconu D. Item analysis and review. Technical report, TIMSS, 2003.
11. Schmitt N. Uses and abuses of coefficient alpha. *Psychological assessment* 1996; 8(4): 350. DOI:10.1037/1040-3590.8.4.350.

12. Lance CE, Butts MM and Michels LC. The sources of four commonly reported cutoff criteria: What did they really say? *Organizational research methods* 2006; 9(2): 202–220. DOI: 10.1177/1094428105284919.
13. Gulliksen H. The relation of item difficulty and inter-item correlation to test variance and reliability. *Psychometrika* 1945; 10(2): 79–91. DOI:10.1007/BF02288877.
14. Brooks GP and Johanson GA. Tap: Test analysis program. *Applied Psychological Measurement* 2003; 27(4): 303–304. DOI:10.1177/0146621603252467.
15. Feldt L, Woodruff DJ and Salih FA. Statistical inference for coefficient alpha. *Applied Psychological Measurement* 1987; 11(1): 93–103. DOI:10.1177/014662168701100107.

Author Biographies

Jorge E. López de Vergara (jorge.lopez_vergara@uam.es) is associate professor at Universidad Autónoma de Madrid (Spain), where he teaches Computer Networks related subjects. He received his M.Sc. and Ph.D. degrees in Telecommunication Engineering from Universidad Politécnica de Madrid (Spain) in 1998 and 2003, respectively. He researches on network and service management and monitoring, having co-authored more than 100 scientific papers about this topic.

Ricardo Olmos (ricardo.olmos@uam.es) is associate professor of Methodology of Behavioural Sciences at Universidad Autónoma de Madrid (Spain), where he teaches Data Analysis. He graduated in Psychology in 1999 and received a Ph.D. in Psychology in 2009 from that university. His research lines are related to computational linguistics, psychometry and statistical linear models.