

Spanish university assessment practices: Examination tradition with diversity by faculty

Ernesto Panadero<sup>1</sup>, Juan Fraile<sup>2</sup>, Javier Fernández Ruiz<sup>1</sup>, David Castilla-Estévez<sup>3</sup> & Miguel Ruiz<sup>3</sup>

<sup>1</sup> Departamento de Psicología Evolutiva y de la Educación, Universidad Autónoma de Madrid, Spain.

<sup>2</sup> Facultad de Educación y Humanidades, Universidad Francisco de Vitoria, Madrid, Spain.

<sup>3</sup> Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, Spain.

**Recommended citation:**

Panadero, Fraile, Fernández-Ruiz, Castilla-Estévez & Ruiz-Díaz (2018). Spanish university assessment practices. *Assessment and Evaluation in Higher Education*.

doi:10.1080/02602938.2018.1512553

**Check the final year of publication as 2018 refers to online first publication.**

Funding:

Research funded by: personal grant to first author under Ramón y Cajal framework (RYC-2013-13469); Fundación BBVA call Investigadores y Creadores Culturales 2015 (project Transición a la educación superior id. 122500); and by Spanish Ministry of Economy and Competitiveness (Ministerio de Economía y Competitividad) National I+D Call (Convocatoria Excelencia) project reference EDU2016-79714-P.

Correspondence concerning this manuscript should be addressed to: Ernesto Panadero. Despacho 109, Facultad de Psicología. Universidad Autónoma de Madrid, Cantoblanco. 28049. Spain. E-mail: [ernesto.panadero@uam.es](mailto:ernesto.panadero@uam.es).

**Notes on contributors**

Ernesto Panadero is a Ramón y Cajal researcher at the Developmental & Educational Psychology Department, at Universidad Autónoma de Madrid and honorary professor at Deakin University (Australia) at the Centre for Research in Assessment and Digital Learning. <http://orcid.org/0000-0003-0859-3616>

Juan Fraile is a lecturer and researcher at the Faculty of Education, at Universidad Francisco de Vitoria, where also coordinates the IDEA research group.

Javier Fernández Ruiz is a second-year PhD student at Universidad Autónoma de Madrid. His main research topics are Formative Assessment and teachers' decision making processes. <https://orcid.org/0000-0001-5419-7687>

David Castilla-Estevez is a doctoral researcher at the Social Psychology & Methodology Department, at Universidad Autónoma de Madrid. <http://orcid.org/0000-0002-0224-3618>

Miguel A Ruiz is associate professor at the Department of Social Psychology and Methodology at the Universidad Autónoma de Madrid since 1996. He is specialized in multivariate data analysis and structural equation modelling in the fields of behavioral sciences and education.

### **Abstract**

Classroom assessment is crucial to understand how students approach course materials, even more in a competitive environment such as the one in higher education. Our aim was to explore the current situation of assessment in higher education to consider further institutional and training actions. Every syllabus from all public universities in Spain was entered into a database, from which 1,693 syllabi were selected completely at random for a content analysis. It was found that (1) university teachers use a greater variety and number of assessment instruments than did their counterparts of decades ago, (2) final exam score is still the highest-weighted source of information for the final grade, (3) the cluster of assessment practices show that traditional approaches are still the most prevalent ones, (4) formal peer and self-assessment practices are still extremely rare in the classroom, (5) assessment practices barely change between first- and fourth-year courses, and (6) most variations in assessment are explained by differences on faculty/academic divisions. This research has implications for European legislation, university regulation and university teacher training programmes.

*Keywords:* higher education assessment, assessment evidence, assessment instrument design, different educational levels assessment, different disciplines assessment.

**Comparing nationwide: assessment instruments, grade weighting, student involvement, faculty/academic division, first and fourth year**

Assessment is crucial for student learning in the classroom. As stated by Wiliam (2011): ‘It is only through assessment that we can find out whether a particular sequence of instructional activities has resulted in the intended learning outcomes’ (p. 3). This becomes especially relevant in higher education as students are supposed to be more autonomous and have less frequent occasions for feedback than at other educational levels (Koivuniemi, Panadero, Malmberg, & Järvelä, 2017; Wold, 2013). It is a complex educational environment in which students are supposed to be able to navigate in bigger classroom groups, with fewer opportunities to receive individualized attention, and to identify the learning goals while achieving them. Throughout these endeavours, students must identify the needed strategies to comply with different assessment practices, since each course teacher usually has freedom to choose between assessment methods unless these are established at the institutional level (e.g., department). For this reason, it is important to explore the current situation of assessment in higher education in order to consider further actions, and this is the aim of the present study.

If there is a key official document where the assessment method is outlined, that would be the syllabus. This document serves as a contract between the teacher and the students on aspects such as content of the course, communication method, etc., and it includes all the relevant information about how students’ performance in the course will be assessed: type of assessment instruments (e.g., exams, exercises), weight of each of the instruments, how the final grade will be calculated, etc. This article will explore a nationwide sample of syllabi to extract conclusions about the most-used assessment instruments, how the typical exam is set up, how the final grade is calculated, what the assessment profiles are and whether formal peer and self-assessment practices are implemented. Answering these questions, we

can better understand how assessment is being conducted in a country within the European Higher Education Area.

### **Assessment in higher education and the Spanish context**

Spain participated in the Bologna Process which aimed to ensure similar standards and quality of higher-education credentials across the European Union (Wächter, 2004). One of the pillars of this process is the implementation of continuous evaluation. This presupposed a change in the teaching and, especially, the assessment paradigms which, in Spain before the Bologna process, were highly dependent on lectures and final exams as the majority methods with a summative orientation (Ibarra & Rodríguez, 2010; Zabalza, 2003).

Spanish university regulations only cover general assessment aspects such as the number of times a student can try to pass a course, the general procedures (usually referring to exams), etc. (Rodríguez-Gómez, Ibarra Sáiz, & García Jiménez, 2013). Rodríguez-Gómez, Ibarra Sáiz, and García Jiménez also concluded that university regulations locate the responsibility for assessment decisions at the departmental level. At the same time, departments rely on teachers as they are entitled to academic freedom. Importantly, Spanish university teachers do not receive formal pedagogical training at the entry level; therefore, they mostly draw upon their experiences as students or work within the faculty/department/area instructional style. Later on, they can gain access to the voluntary pedagogical training programmes that some universities have implemented. Importantly, all teachers are evaluated by their students via surveys. The results are returned to the teachers but are usually just informative, without any repercussions in the short term. As Spanish private universities use different schemas, it was decided to explore only universities within the public system, which are the majority in Spain.

### **Characteristics of assessment in higher education**

This study covers four assessment components. First, it is important to explore what types of *assessment instruments* are being used to evaluate higher education achievement. These refer to instructional activities such as exams, practices and portfolios. The choice of assessment instruments provides different insights into how students are learning (Brookhart & Nitko, 2015). For example, if one course assessment relates exclusively to the final exam, then assessment is not fulfilling the purpose of tracking the development of the students' learning. In addition, we will also explore two other assessment instruments' features. First, what number of instruments are used to conclude whether the students perform different assessment activities (e.g., final exam and practices) or whether they just go through a final and unique activity. Second, we will analyse what type of questions are used in the final exams in order to further reflect on the nature of this instrument practice.

Secondly, how teachers *weight the different assessment instruments to calculate the final course grade* will be explored. In higher education the summative function, based on the certification of students' achievement (Brown, 2002), is clearly represented because the vast majority of university courses generate a final grade calculated as the sum of the activities performed by the student. Based on that final grade, students' performances can be easily compared and organized hierarchically to select the 'best' students. Due to the importance of such grades, it is then relevant to explore on what information the students' performances are calculated.

Obviously, in a context as competitive as the university, the final grade is of vital relevance to the students. Research has found students' grading experiences and perceptions largely influence their approaches to learning (Tippin, Lafreniere, & Page, 2012). Therefore, it is important to analyse how the final grade is calculated and what assessment instruments it is based on. For example, if only the final exam is used to calculate 100% of the final grade, there are no chances for the students to build a trajectory of how well they are doing during

the semester. In such a case, students' grades for their semester's work and learning depends on a 'one shot' assessment.

Thirdly, we will explore whether different *assessment profiles* can be identified. The purpose here is to analyse the data to create clusters based on the different assessment practices. There is a large literature on teachers' conceptions about assessments and how they shape what they do in classrooms (e.g., Brown, 2004). This research goes from the teachers' conceptions to their practices. Here we go from the practice to try to identify the patterns and create assessment profiles.

Fourthly, the presence of formative assessment practices via *student involvement in assessment activities – i.e., peer and self-assessment* – is also explored because these practices have a positive effect on students' learning (Brown & Harris, 2013; van Zundert, Sluijsmans, & van Merriënboer, 2010). Unfortunately, the formative function cannot be taken for granted in the university. Rodríguez-Gómez et al. (2013) pointed out the absolute lack of regulation on formative assessment practices (e.g., self- and peer assessment) in Spain except for the case of one university. Importantly, these authors also found that teachers might choose to implement these practices on their own, but they might face a struggle because of the lack of regulatory support and their low instructional training. This underscores the importance of exploring whether formative assessment practices are actually included in the syllabi in the form of self/peer assessment or grading. Unfortunately, previous research has found that university teachers involve their students in assessment significantly less than teachers at other educational levels (e.g., Panadero & Brown, 2017; Panadero, Brown & Courtney, 2014). This can be perceived as a contradiction, as it is at this level that students are supposed to be at their highest autonomy, yet they seem to have the least to say in their assessment. Hence the importance of contrasting these previous results as they are based on online surveys with more objective measures of teachers' assessment practices.

### **Comparing assessments in two areas: university-year level and knowledge areas**

Exploring the effects of the university-year level, year is important because it may be assumed that students in the first year of a programme would not be assessed the same way as those in the last year (Brown, 2011; Hirschfeld & Brown, 2009). Students who have been longer in a programme are assumed to have more knowledge and developed skills and, therefore, it is possible that teachers would use different assessment practices with them. The differences between the beginning and the end of the programme –1st vs. 4th – were analysed to cover the extreme range of changes. A previous study using the same syllabi methodology and exploring just one particular degree did not find any differences based on university-year level (Fraile, Pardo & Panadero, in press).

In a similar fashion, it may be expected that knowledge areas, arising from different academic and even teaching traditions, might differ (e.g., mathematics vs. history) (Pellegrino, 2002). Here, we explored faculty and academic division as grouping variables. Some analyses were run at the faculty level because they are more specific and provide more information, but due to the high number of formal disciplinary fields – 11 – for some analyses, we fused them together into five academic divisions, since it was not feasible to run intelligible analyses with them (see Appendixes 1 and 2 for more information).

### **Syllabus as unit of analysis**

A syllabus is an official document in which the teacher(s) of a course provide(s) pedagogical information, expectations and responsibilities to students (Stanny, Gonzalez, & McGowan, 2015). Being an official document, it is compulsory for all subjects in Spanish universities to define the assessment procedure (ANECA [Agencia Nacional de Evaluación de la Calidad y Acreditación – National Agency for Quality Assessment and Accreditation], 2013), including information such as types of assessment instruments (e.g., exams, portfolio), weight of each instrument in the final course grade, etc. Syllabus analysis provides an

interesting view into the instructional environment teachers create in their university courses, as it is an ‘unobtrusive but powerful indicator of what takes place in classrooms’ (Bers, Davis, & Taylor, 2000 p. 7).

There has been plenty of successful research done using this methodology in different areas, such as exploring the alignment of the syllabi with student learning outcomes (e.g., Cashwell & Young, 2004) and the effects of academic teacher development programmes (Rathbun, Leatherman, & Jensen, 2017). Importantly, though a syllabus must comply with certain institutional requirements (e.g., including compulsory official information), it still contains a large portion of information that is specific for that course. This is because teachers have a considerable level of freedom in determining the content and structure of their courses, how to design their instructional activities, how to assess students’ learning, etc. (Stanny et al., 2015). Therefore, analysis of syllabi will provide an interesting picture of assessment practices.

### **Aim and research questions**

Our aim is to explore how assessment is conducted in higher education through a nationwide sample of syllabi, which is a distinctly more objective measure than directly asking the university teachers about their practices via surveys or interviews. Four research questions were posed, each of them with two parts: a descriptive one, in which we report general characteristics (e.g., frequencies), and a comparative one, using the moderating variables (first year vs. fourth year, faculty/division).

RQ1: Assessment instruments: what type, how many are used and how are exams constructed?

RQ2: What percentage weight is each instrument given in the calculation of the final grade?

RQ3: What is the most common assessment profile?

RQ4: Are students involved in assessment (e.g., peer and self-assessment)?

## Method

### Data collection and description

A database was created including every university syllabus ( $N = 78,094$ ) for all the public universities in the country with undergraduate studies ( $N = 48$ ). Every syllabus was assigned its own case code. A random selection of cases was performed, resulting in the first sample of 1,450 cases. In the second round, the random selection was expanded to ensure reaching a minimum of 30 cases from each university, for a total sample of 1,645 cases. In the third round, we added cases from three universities that initially had been mistakenly classified as private. Then, the syllabi were downloaded from the different universities' websites and, after elimination of the ones with no information about the assessment methods, we concluded with a final sample of 1,693 cases.

The distribution of the syllabi by university-year level was as follows: 441 from the first year, 434 from the second year, 402 from the third year and 355 from the fourth year. Additionally, 42 syllabi correspond to optional courses that can be taught in the third or fourth year. The most represented faculty are in Engineering & Architecture (29.2% of the total syllabi), followed at some distance by Medical & Health Science (11.2%), Economics and Hard Sciences (10.8% each), Law & Administration (7.5%), Human Studies (7.4%) and Philology & Languages (7.1%). The less represented faculties are in Education (5.6%), Psychology & Social Labour (3.9%), Arts (3.5%) and Communication (3.0%).

### Data coding

The selected cases were coded by following the categories in Appendix 1. Additionally, in Appendix 2, the categories are explained in more detail. Three coders participated in the data collection with different levels of implication (coder 1 processed 55% of the cases, coder 2 processed 40%, and coder 3 processed 5%). Krippendorff's alpha was calculated for the main categories after a training/discussion process among the three coders.

The level of agreement on categories was above or close to .70, as follows: Final Written Exam, 0.69; Practical Exam – i.e., real-world task to perform –, 0.86; Partial Exam –i.e., mid-term exams-, 0.82; Portfolio, 0.73; Assignment (combining Individual and Group), 0.68; Practices, 0.67; and Attendance, 0.72. Therefore, the agreement among the three coders was above the recommendation of a minimum of 0.60 (De Swert, 2012).

### **Methodology and data analysis**

To answer RQ1, frequencies taken from a multi-response set were calculated for assessment instruments and types of exam. Differences between frequencies of use were examined using chi-square exams for independence, including bootstrap when chi-square assumptions could not be met, and joint distributions of use were analysed using multivariate Correspondence Analysis. Statistics and *p*-values, together with Cramer's V effect size value (or phi value for 2×2 tables), and adjusted standardized Z residuals were calculated. Three exceptions were made. Firstly, to evaluate the difference between the number of assessment instruments used in first- and fourth-year courses, the non-parametric Mann-Whitney exam was performed (due to non-normal distribution). Secondly, to evaluate the differences between the number of assessment instruments used among faculties, universities and regions, several non-parametric Kruskal-Wallis exams were carried out. Lastly, in order to find to what extent differences in usage of assessment instruments across cases was explained by universities and regions, two logistic regressions were carried out. Assessment instruments used were mapped using multiple-correspondence analysis.

For RQ2, descriptive analyses were calculated for absolute frequency, means and standard errors. For comparisons, again, non-parametric Mann-Whitney and Kruskal-Wallis exams were computed.

Regarding RQ3, several latent class analyses were executed, since this approach has better properties than traditional k-means clustering (Magidson & Vermunt, 2002). As data were scarce due to the many different possible distributions of use of assessment practices, bootstrap analysis (500-replications) was calculated to assess  $p$ -values for competing solutions.

To answer RQ4, we computed the cases reporting self-grading, self-assessment, peer grading and/or peer assessment. Some recodification was made to compute one overall formative assessment variable, which was included in several chi-square independence exams, analogously to the exams explained before. A new dichotomous variable (yes vs. no) labelled as *Student involvement in assessment* was computed. It referred to all the cases where student involvement in assessment is present. Three chi-square independence exams were carried out in order to know under which circumstances student involvement in assessment is more likely to occur.

All analyses were carried out using IBM SPSS 23 and Latent Gold 5.1 software.

## Results

The results of the four research questions are presented following the same structure. First, descriptive data is presented – i.e. labelled as ‘Overall picture’ – in which we provide information about frequencies among the different categories. Secondly, the comparisons in the different grouping variables (first vs. fourth year and faculty/academic division) are presented.

### **RQ1: Assessment instruments: what type, how many are used and how are exams constructed?**

#### **Overall picture**

*Types of assessment instruments.* Analysing the distribution of the types of assessment instruments, the two most widely used were Final Written Exam (70.2%) and Practices

(59.4%). Partial Exam (38.8%), Individual Assignment (33.9%), Attendance (27.3%) and Group Assignment (25.9%) were less frequent. Lastly, Practical Exam (15.3%) and Portfolio (4.6%) were seldom used. Importantly, each syllabus could contain more than one assessment instrument. When assessing the joint use of instruments, we found patterns in their frequency of use. Instruments could be plotted in two dimensions: the first one related to the use of partial exam (right) vs. the use of final written exams (left); and the second one related to the use practices (top) vs. the use of individual assignment (bottom). It is interesting to observe that partial exams were used more when no final written exam was given, and final written exams were related more to attendance and group assignment. Practical exams were more frequently observed when practices were used and no individual assignment was preferred.

This mapping of assessment instruments reveals some patterns which are rather common in our educational system; nevertheless, they should be considered as a first descriptive approach, since the eigenvalues for the dimensions were not very large ( $D_1 = 1.15$ ,  $D_2 = 1.33$ ), explaining 35% of the existing inertia, and the associated reliabilities were also mild ( $\alpha_1 = 0.35$ ,  $\alpha_2 = 0.28$ ).

<<Insert figure 1 around here>>

*Number of assessment instruments utilized.* Six cases were excluded (0.6%) because they did not report any assessment instrument. Most of the syllabi used two ( $N = 564$ , 33.3%) or three ( $N = 660$ , 39.0%) assessment instruments. At a lower level, 284 syllabi (16.8%) used four evidences, and 122 syllabi (7.2%) used only one evidence. Virtually no syllabi used more than four evidences ( $N = 47$ , 2.7% using five evidences, and  $N = 5$ , 0.2% using six).

*Types of final written exams.* More than half of final written exams used closed response multiple choice (52.2%). Next came short-answer questions (39.0%), problems

(30.3%) and essay questions (28.5%). Notice that a final exam can use more than one type of exercise (e.g. multiple choice and essay questions).

### **Comparisons between first- and fourth-year courses**

As final projects and practicums take place in the fourth year and have a distinct assessment, two analyses were made: including and excluding them.

*Types of assessment instruments utilized.* Results of the chi-square values ( $\chi^2$ ), Phi effect sizes ( $\phi$ ) and adjusted standardized residual values ( $Z_E$ ), are presented in Table 1 when significant. Optional courses did not show any significant difference, so they were not included in the following comparisons. Final written exams (70.8%), partial exams (48.4%) and practices (61.5%) were used more in first-year courses. Individual assignments (46.8%), such as reflective essays or literature reviews, were used more in fourth-year subjects. As can be seen in Table 1, both projects and practicums have relevant influences in some of the instruments' use, since differences between first- and fourth-year courses are less significant (or even become non-significant) when these two instruments are not considered.

Including the university-year level in the perceptual map previously obtained (Figure 2), first-year courses were located near practices, practical exam, and partial exams, while fourth-year courses were located relatively closer to using individual assignments, as well as portfolios, attendance and group assignments. The eigenvalues for the dimensions extracted were not very large ( $D_1 = 1.55$ ,  $D_2 = 1.47$ ), explaining 34% of the existing inertia, and the associated reliabilities were slightly higher ( $\alpha_1 = 0.40$ ,  $\alpha_2 = 0.36$ ).

**<<Insert Table 1 around here>>**

**<<Insert figure 2 around here>>**

*Number of assessment instruments utilized.* The Mann-Whitney exam showed that the number of assessment instruments used in first-year courses was not significantly different from those used in fourth-year courses ( $U = 29,718$ ,  $p = .615$ ). Therefore, it can be assumed

that first- and fourth-year students experience a similar number of assessment instruments.

Discarding final projects and practicums, this difference remained non-significant ( $U = 58931.5, p = .662$ ), which means that the influence of these types of subjects was not high in our sample.

*Final written exam types.* As a less-than-five count was expected in over 20% of the cells, a Monte Carlo method with 10,000 samples was included in the chi-square exam. No significant difference was found, meaning that the final written exam types are similar in the first and fourth years.

### **Comparisons among faculties**

*Type of assessment instruments utilized.* Table 2 presents a brief summary of the results and Figure 3 helps visualize them. Including the faculty in the correspondence analysis additionally shows that final written exams are more frequent in Economics and Law & Administration; practical exams are most used in Medical & Health Science and Hard Sciences; partial exams in Hard Sciences and Engineering & Architecture; portfolio in Education and Philology & Languages, with the latter also close to individual assignment; and Communication, Human Studies, Philology & Languages and Education are closer to attendance, group assignment and individual assignment. Practices are located closer to Economics, Medical & Health Science, Hard Sciences and Engineering & Architecture. The eigenvalues for the two dimensions extracted are slightly larger ( $D_1 = 1.68, D_2 = 1.47$ ), explaining 35% of the existing inertia, and the associated reliabilities are slightly higher ( $\alpha_1 = 0.46, \alpha_2 = 0.36$ ), showing an increase in variability and relations.

<<Insert Table 2 around here>>

<<Insert figure 3 around here>>

*Number of assessment instruments.* The Kruskal-Wallis exam shows that the number of assessment instruments used differs among faculties ( $\chi^2 = 25.414, df = 10, p = .005$ ). To

explore this further, two dichotomous variables (occurrence vs. absence) were created: one for the cases with only one instrument used and another for cases with more than three instruments reported. There were two reasons to create these variables. First, since normality assumptions were not met, no parametric comparisons could be run. Second, two or three instruments were used in most of the cases. Thus, using one instrument could be seen as ‘low number of instruments’ utilized, whereas a usage above three instruments could be perceived as ‘large number of different instruments used’. Thus, chi-square analyses were computed to study the behaviour of faculties on these two variables.

For one instrument used, data points towards a weak dependence ( $\chi^2 = 20.936$ ,  $df = 10$ ,  $p = .022$ ,  $V = .115$ ). Only significant residuals are found for Engineering & Architecture ( $Z_E = +2.8$ ) and Arts ( $Z_E = +2.1$ ), meaning that only one instrument is used in significantly more cases than the rest. With respect to presenting more than three, again, a weak dependence ( $\chi^2 = 21.476$ ,  $df = 10$ ,  $p = .018$ ,  $V = .116$ ) was found. More than three instruments are significantly used in more cases only in Education ( $Z_E = +2.7$ ) and less used in Economics ( $Z_E = -2.4$ ) and Communication ( $Z_E = -2.0$ ).

*Types of final written exams.* Multiple choice exams are significantly most used in Medical & Health Science ( $Z_E = +9.8$ ) and Psychology and Social Labour ( $Z_E = +4.8$ ); short answer also in Medical & Health Science ( $Z_E = +3.3$ ) and Hard Sciences ( $Z_E = +2.2$ ); essay questions are used equally except for Engineering & Architecture where they are used the least ( $Z_E = -2.6$ ); and problems are used more in Engineering & Architecture ( $Z_E = +6.1$ ) and Hard Sciences ( $Z_E = +2.6$ ).

## **RQ2: What percentage weight is each instrument given in the calculation of the final grade?**

The final grade is calculated here by a weighted average of assessment instruments by number of cases with available percentage. However, zero-percentage values for each

instrument are very numerous (their numbers range from 35% for final written exams to 96.1% for portfolios). That is why it was necessary to fuse some variables before calculating the weighting. Thus, Final Written Exam and Practical Exam were computed together as Final Exam; Practices and Portfolio were fused into Practical Ex. and Portfolio. Finally, Individual Assignment and Group Assignment were fused into a new Assignments instrument variable. Attendance and Partial Exams were not modified because their assessment meanings are clearly different.

### **Overall picture**

Final Exam is the type of instrument that is weighted more, with 59.46% of the final grade. At a lower level, we find Assignments (31.41%), Practical Exams and Portfolio (27.58%) and Partial Exams (27.47%). Lastly, Attendance is weighted with 12.26% of the final grade.

### **Comparisons between first and fourth years**

For this occasion, practicums and final projects were not included because these two types of assessment have distinctive ways of grading. As all the Kolmogorov-Smirnov exams were significant, non-parametric Mann-Whitney exams were calculated. Results show no significant differences between first- and fourth-year courses regarding how the weight of the instruments is used to calculate the course grade for the Final Exam ( $U = 26,483$ ,  $p = .159$ ), Partial Exams ( $U = 766$ ,  $p = .583$ ), Practical Exams & Portfolio ( $U = 15,779.5$ ,  $p = .054$ ), or Attendance ( $U = 3,023$ ,  $p = .826$ ). Only Assignments ( $U = 2,346.5$ ,  $p = .007$ ) showed a significant difference, as it was found that the mean rank (MR) for fourth-year courses (MR = 88.88) was higher than for first-year courses (MR = 69.38). This indicates that assignments in the compulsory/basic fourth-year courses have greater weight in the final grading than in first-year courses. Unfortunately, due to the lack of precision in the syllabi description of these

assignments, we could not explore whether these are of different natures in the first and fourth years.

### **Comparisons for faculties**

Non-parametric Kruskal-Wallis exams were made for every instrument. Three of the categories showed significant differences. Final exam is used more in Hard Sciences and lowest in Communication, but Assignments and Practical Exam & Portfolio were the other way around: the highest use is in Communication and the lowest in Hard Sciences.

### **RQ3: What is the most common assessment profile?**

#### **Overall assessment profiles**

For the general assessment profile, the 3-class model was the most parsimonious one (BIC = 9,949.0), but it did not fit. Bootstrap confirmed this, so the 4-class model was chosen because of its BIC level (second lowest, BIC = 9,956.8) and its almost non-significant  $p$ -value after bootstrapping ( $p = .02$ ). Profiles for the four clusters are shown in Figure 4. Each line represents one profile, indicating the distributions of probabilities for each assessment practice to be used in that cluster. Cluster 1 includes 38.0% of the cases and presents a *Traditional* approach, based mostly on Final Exam and Practical Exam & Portfolio. Cluster 2 (14.6% of cases) is labelled as *Traditional & Hands-on*, characterized by Assignments and Final Exam, with the greatest weight given to Attendance. Cluster 3 (29.4% of cases) is labelled as *Exam Combination* because of the high presence of Partial Exams and, to a lesser extent, Final Exam. Finally, Cluster 4 (18.0% of cases) is labelled as *Hands-on* because of the high prevalence of Assignments, combined with Practical Exam & Portfolio and Partial Exams.

<<Insert figure 4 around here>>

#### **Comparison of main profiles for first- and fourth-year courses**

As practicums and final projects have distinct assessments, such cases were not included in the comparison. As can be seen in Figure 5, fourth-year assessment is highly based on final exam and assignments, while first-year assessment presents more partial exams. For first-year courses, the best model is the 3-class model, with the second lowest BIC = 2,597.4 and a bootstrapped  $p = .02$ , almost non-significant. For fourth-year courses, the best one is the 2-class model, with BIC = 1,825.2 and  $p = .05$ , which increases to .08 after bootstrapping. Only the most common profiles (Cluster 1) are displayed in Figure 5. For first-year courses, 50.3% of cases are situated in Cluster 1, while 67.9% of cases for fourth-year courses are in Cluster 1.

<<Insert figure 5 around here>>

#### **Comparison of main profiles for courses in each academic division**

Lastly, the same analyses were made for each division. Data for each best model are summarized in Table 3, and profiles are presented in Figure 6. As can be seen, the divisions present very distinct profiles. Arts and Humanities are characterized by an absolute presence of final exams and assignments, and they are also the subject requesting higher attendance at lectures. Engineering & Architecture rely on partial exams and depend the least on attendance and final exams, which is logical since they have partial exams. Social and Hard sciences have similar profiles except for final exams which are used more in social sciences. Lastly, Health Sciences rely on final exams in combination with practical exams and portfolios.

<<Insert Table 3 around here>>

<<Insert figure 6 around here>>

#### **RQ4: What is the presence of student involvement in assessment?**

Here, two different aspects needed to be explored. First, when students are involved in the assessment of their work without grading it, both Peer Assessment and Self-Assessment

variables are considered. When grading is present, studied variables are Peer Grading and Self-Grading.

### **Overall student involvement in the assessment picture**

The occurrence of student involvement in assessments is extremely scarce: Self-Assessment,  $N = 92$  (5.7%); Peer Assessment,  $N = 37$  (2.3%); Self-Grading,  $N = 41$  (2.6%); and Peer Grading,  $N = 26$ , (1.6%).

*To what extent do self- and peer assessment and self- and peer grading, appear together?* Chi-square exams were made, resulting in a moderate link between self-assessment and self-grading ( $\chi^2 = 178.596$ ,  $df = 1$ ,  $p < .005$ ,  $\phi = .334$ ). In contrast, a much stronger link was found between peer assessment and peer grading ( $\chi^2 = 652.529$ ,  $df = 1$ ,  $p < .005$ ,  $\phi = .638$ ). In fact, whereas both variables are present in only 22 cases (23.9% of self-assessment cases, 53.7% of the self-grading cases), for the peer variables, both variables appear in 20 cases (54.1% of peer-assessment cases and 76.9% of peer-grading cases).

*How much student involvement in assessment is present in each instrument?* As can be seen in Table 4, larger percentages of student involvement in assessment correspond to both individual and group assignments and less to portfolio.

<<Insert Table 4 around here>>

### **Comparisons between first- and fourth-year courses**

Without considering final projects and practicums, no significant differences between first- and fourth-year courses were found, either for self-assessment ( $\chi^2 = 2.112$ ,  $df = 1$ ,  $p = .146$ ) or for peer-assessment ( $\chi^2 = 0.935$ ,  $df = 1$ ,  $p = .334$ ).

### **Comparisons among faculties**

Formative assessment does not show significant variability across faculties ( $\chi^2 = 16.819$ ,  $df = 10$ ,  $p = .078$ ).

## **Discussion**

Our aim was to explore what the current situation of assessment in higher education is regarding four different assessment components.

### **Type and number of assessment instruments used and type of final exam**

Regarding our RQ1, it was found that Spanish university teachers mostly use two or three types of assessment instruments and that these instruments are mostly final written exams, practices and, to a lesser extent, partial exams, individual and group assignments and classroom attendance. The use of practical exams and, especially, portfolios is scarce. Additionally, final written exams usually contain multiple choice and short-answer items and, to a lesser extent, problems and essay questions.

Previous research done in the 1980s and 1990s found that Spanish university courses mostly relied on a final exam and a summative approach to assessment (Zabalza, 2003). The landscape has changed since then, according to our results. Additionally, it was also previously found that there were important absences of pedagogical training of Spanish lecturers in assessment and that teachers recognize that they lack knowledge of assessment instrument design (e.g., Panadero et al., 2014; Quesada-Serra, Rodríguez-Gómez, & Ibarra Sáiz, 2017). That was not explored here, and therefore future research needs to analyse whether the exams aim at the level of memory and comprehension of factual material and not at the higher levels of Bloom's taxonomy (Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956) because creating better assessment items requires practice and training (Brookhart & Nitko, 2015). This is a crucial aspect because exam design has a crucial effect on students' learning. For example, a student who has prepared shallowly for an exam might still get a high grade if the instrument uses items with low cognitive levels (Asikainen, Parpala, Virtanen, & Lindblom-Ylänne, 2013). Nevertheless, there is some good news in this landscape: teachers do implement multiple assessments and combination of instruments which should provide students with several occasions for reflection and knowing where they are in relationship to

their performance. This will provide teachers and students with different sources of information that might be closer to continuous evaluation. Notwithstanding, there seems to be considerable room for improvement.

Regarding our moderating variables, it was found that first- and fourth-year assessments do not differ much because they use similar assessment instruments and numbers of them and the same types of final exam questions. The only aspect that varies is that there is a higher number of partial exams for the first year vs. more individual assignments in the fourth year. This seems to emphasize that the alternative source of information for the final exam varies, but that is all that changes. Secondly, the comparison among faculties clearly shows that there are large discrepancies among them in the three aspects explored in RQ1: (a) types of instruments (e.g., Economics uses a large number of final written exams compared to Engineering & Architecture, which has a low number); (b) type of final exam (e.g., massive use of multiple-choice items in Medical & Health Sciences vs. a low number in Engineering & Architecture); and (c) number of instruments used (e.g., Education has a larger number of instruments).

To summarize, the RQ1 results show that there are a variety of assessment instrument types, variations in their numbers and in final exam item design and that the vast majority of the variations occur at the faculty and university levels. This can be interpreted as the strength of certain assessment traditions within areas of knowledge, i.e., faculties. Additionally, our results show that the situation in Spanish higher education assessment seems to have changed from a more traditional approach based on lectures and a final exam (Ibarra & Rodríguez, 2010; Zabalza, 2003) to the current scenario found in our data. The fact that assessment does not change much from the first to the fourth year throws a negative light on the matter. It can be assumed that fourth-year students are more expert, yet they still undergo similar types of assessment.

**Assessment instrument weight in the final grade**

The RQ2 explored how the assessment instruments were weighted in the final grade. The general picture shows that the final written exam is still the instrument that counts the most for the final grade at almost 60%. This emphasizes the need, mentioned above, to explore how those exams are designed and what type of Bloom taxonomy levels they aim for. Additionally, the final written exam is combined with others, such as assignments, partial exams and practices. This again shows that currently, Spanish university teachers are giving more importance to multiple data sources to assess their students' performance. Nevertheless, it is somehow surprising that attendance at lectures is weighted up to 12% of the final grade because, without a strong participation methodology, teachers would be measuring passive attendance, which is not necessarily an accurate predictor of students' learning.

Regarding the comparisons, it was found again that there are barely any changes between first- and fourth-year courses and that most of the variance is explained by differences among faculties. An interesting observation is that hard sciences and communication show opposite results, with the former putting the highest weight on the final written exam and the latter on assignments and practices, which is probably in line with the common assumption that one is about learning the 'right' answers and the other about developing talents. Previous research has found that teachers within the same department tend to score similarly (Beenstock & Feldman, 2018), supporting our results that a large part of the variance is at the institutional level by proximity to colleagues' assessment practices. The interpretation of these results is similar to the one in the previous section: the Spanish situation has leaned towards a more continuous evaluation as intended by the Bologna process, but there is still a long way to go.

**Most common assessment profiles**

RQ3 explored what the most common assessment profiles are, based on the data from the two previous questions. Our first latent class model identified four assessment profiles labelled as traditional, traditional & hands-on, exam combination and hands-on. Looking at the distribution of cases, the results show that a traditional approach to assessment based on exams is still the prevalent approach, as found in previous research (Quesada-Serra, Rodríguez-Gómez & Ibarra-Sáiz, 2016).

The second latent class model shows that the main difference between first and fourth year is the large use of partial exams in first-year courses. Interestingly, this type of instrument is extremely scarce in the fourth year because it is 'compensated' for by a greater use of assignments (significant difference) and final exams (non-significant). Therefore, and due to the fact that only assignments are weighted differently in the first versus the fourth year, one viable hypothesis is that the first-year partial exams are exchanged for assignments in the fourth year. It is important to remember that, unfortunately, the syllabi's descriptions of assignments were not precise enough to code that data, and we could not explore whether the assignments in the first and the fourth years are of different natures. Importantly, these two latent class models continue to show that assessment in higher education is still largely traditional.

The third and last latent class model contains very interesting insights. This model allows us to clearly identify patterns in the five academic divisions. While Health, Social and Hard Sciences show similar patterns (i.e., use of final exams and practices and, to a lesser extent, assignments), Arts & Humanities and Engineering & Architecture have unique features. Arts & humanities has the highest prevalence of final exams, assignments and attendance. Engineering & Architecture has the highest prevalence of partial exams while final exams and attendance are the lowest. These results might help explain why faculty/academic division is a powerful modulating variable: coming from different

knowledge traditions, they might need different instructional strategies that develop into different assessment instruments and practices. Additionally, and as mentioned above, because the Spanish university regulations leave specific assessment decisions to the department/faculty level, a significant part of the variability is explained at that level (Quesada-Serra et al, 2016).

### **Student involvement in assessment**

Finally, RQ4 explored the presence of self/peer assessment and self/peer grading. As might have been expected due to previous research (Panadero et al., 2014; Panadero & Brown, 2017; Quesada-Serra et al., 2016), these assessment practices are extremely scarce; for example, Fraile et al. (in press) found 7.55% of the syllabi to include self-assessment, 9.43% self-grading and 1.99% both. Two interesting findings here are: (a) while there is a moderate link between self-assessment and self-grading, there is a much stronger one for peer assessment and peer grading, i.e., these are usually implemented together; (b) student involvement in assessment is not explained by university-year level or faculties (except for higher values in the education field). Regarding the last one, a conclusion is that university policies and training courses might be the best way to increase formative assessment practices. A recent analysis of Spanish universities' regulations (Rodríguez-Gómez et al., 2013) found only one university's regulations that included self-assessment. Therefore, when the university's regulations do not include self- and peer assessment as grading and instructional options can be a severe impediment to their implementation, along with the lack of training as shown in previous research (Panadero et al., 2014; Panadero & Brown, 2017).

### **Limitations**

This study has several limitations. First, our data come from official documents and not from direct observation of what teachers do. Nevertheless, the syllabus is a contract between the teacher and students, and therefore it should significantly resemble what happens

in the classroom. Significantly, the selection of our sample is extraordinarily strong, coming from a database including all syllabi from all public universities and selecting them completely at random.

Secondly, our study did not include analyses of the assessment instruments' content. It cannot be concluded here that exams in the first and fourth years contain the same type of Bloom-level items. Future research needs to explore the exams' content.

Thirdly, there are two potent variables that, due to our procedure, we could not control for: teaching and training course experience. These have been shown to predict assessment practices. Nevertheless, again, due to our strong data selection method, it can be assumed that the distribution is faithful to the one in the real population.

### **Implications**

Firstly, there has been an improvement in Spanish higher education assessment practices. Currently the final course grade contains information from more than just a final exam because other types of instruments weigh into the grade. These different measures provide occasions for students to receive information about their level of performance. However, more improvement is needed. Teachers report lack of experience in the design of assessment instruments regardless of whether their use is summative or formative (Quesada-Serra et al., 2016). Additionally, activities that are central to students' self-regulated learning, such as their involvement in self- and peer assessments, are practically non-existent. This might be due to two reasons. First is the lack of regulation and a legal framework to implement actions such as self- and peer grading. The second explanation is lack of teacher training and security in how to perform these practices (Quesada-Serra et al., 2016). This is extremely important as training and previous experience have been found to be the highest predictors of formative assessment practices (Panadero et al., 2014; Panadero & Brown, 2017).

Therefore, the solutions for achieving higher levels of assessment expertise are institutional on one side: we need to transcend the faculty and departmental constraints and influences in the way feedback is given (Bailey & Garner, 2010). The university assessment regulations need to be revisited and updated while increasing and, even in some cases, starting the training teachers get in all topics: design of assessment instruments, use of different instructional techniques, implementation of formative assessment, etc. One key result in our study is the lack of differences between the assessments performed in first- and fourth-year courses: students at such different levels should have different knowledge and skills and should be assessed correspondingly.

All of these reforms have to keep teachers' workloads in mind because it might be one of the reasons why teachers still use final exams: it can imply less work than other assessment instruments. Therefore, if teachers cannot be given smaller classes, lighter loads of administrative duties and less pressure to publish, universities should at least make sure the teachers get training in how to improve their assessment instrument designs.

## **Conclusion**

Our data show that there has been an improvement over the last two decades in the Spanish assessment panorama: teachers now use a greater variety and number of instruments and they are weighted into the final grade. However, as our data shows, the prevalent assessment profiles are still traditional, with barely any student involvement in assessment. Furthermore, there are hardly any changes between how assessment is conducted in the first year and the last year, and a large proportion of the variance in assessment can be explained by the faculty and division, which points to different academic traditions.

Hence, there are still actions to be taken. Regulations should be more supportive of assessment practices such as self- and peer assessment to ensure that students leave higher education with these crucial workplace abilities. Additionally, a time for reflection should be

introduced, especially if we consider that the students about to finish the undergraduate programmes in their fourth year are basically assessed in a similar fashion as those entering the university. While some might call this assessment consistent, it is probably more a lack of adjustment to how students develop their knowledge and skills in the programmes. A large part of this might be due to the limited expertise in assessment design that Spanish university teachers reported in previous research, and it might be expected that assessment instruments may be poorly designed in the first as well as in the fourth year. Therefore, these results should be a clear indication for universities to make some profound changes while providing a more positive outcome at European levels. Things have improved since the Bologna process. Time for a new one?

### References

- ANECA (2013). Guía de apoyo para la redacción, puesta en práctica y evaluación de los resultados del aprendizaje. Madrid: ANECA [Agencia Nacional de Evaluación de la Calidad y Acreditación - National Agency for Quality Assessment and Accreditation].
- Asikainen, H., Parpala, A., Virtanen, V., & Lindblom-Ylänne, S. (2013). The relationship between student learning process, study success and the nature of assessment: A qualitative study. *Studies in Educational Evaluation*, 39(4), 211-217.
- Bailey, R., & Garner, M. (2010). Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in Higher Education*, 15(2), 187-198.
- Beenstock, M., & Feldman, D. (2018). Decomposing university grades: a longitudinal study of students and their instructors. *Studies in Higher Education*, 43(1), 114-133.  
doi:10.1080/03075079.2016.1157858
- Bers, T. H., Davis, B. D., & Taylor, B. (2000). The use of syllabi in assessments: Unobtrusive indicators and tools for faculty development. *Assessment Update*, 12(3), 4-7.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956).

*Taxonomy of educational objectives, handbook I: The cognitive domain* (Vol. 19, p. 56). New York: David McKay Co Inc.

Brookhart, S. M., & Nitko, A. J. (2015). Providing formative feedback. *Educational assessment of students* (7th ed., pp. 153-165). Boston, MA: Pearson Education.

Brown, G. T. L. (2002) Teachers' conceptions of assessment. (Tesis doctoral). University of Auckland, Auckland, NZ

Brown, G. T. L. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice*, 11(3), 301-318.

Brown, G. T. L. (2011). Self-regulation of assessment beliefs and attitudes: a review of the Students' Conceptions of Assessment inventory. *Educational Psychology*, 31(6), 731-748. <http://doi.org/10.1080/01443410.2011.599836>

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. McMillan (Ed.), *The SAGE handbook of research on classroom assessment* (pp. 367-393). Thousand Oaks, CA: SAGE.

Cashwell, C. S., & Young, J. S. (2004). Spirituality in counselor training: A content analysis of syllabi from introductory spirituality courses. *Counseling and Values*, 48(2), 96-109.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of educational research*, 58(4), 438-481.

Fraile, J., Pardo, R., & Panadero, E. (in press). Autoevaluación y autocalificación en el Grado en Ciencias de la Actividad Física y del Deporte: estudio censal de las guías docentes. *Revista de profesorado*.

- Hirschfeld, G.H.F., & Brown, G.T.L. (2009). Students' conceptions of assessment: Factorial and structural invariance of the SCoA across sex, age, and ethnicity. *European Journal of Psychological Assessment, 25*(1), 30–38.
- Ibarra, M. S., & Rodríguez, G. (2010). An approach to the dominant discourse of learning assessment in higher education. *Revista de Educación, 351*, 385–407.
- Koivuniemi, M., Panadero, E., Malmberg, J., & Järvelä, S. (2017). Higher education students' learning challenges and regulatory skills in different learning situations/Desafíos de aprendizaje y habilidades de regulación en distintas situaciones de aprendizaje en estudiantes de educación superior. *Infancia y Aprendizaje, 40*(1), 19-55.
- Magidson, J., & Vermunt, J. K. (2002). Latent class modeling as a probabilistic extension of K-means clustering. *Quirk's Marketing Research Review, 20*, 77-80.
- Panadero, E., Brown, G. T. L., & Courtney, M. G. R. (2014). Teachers' reasons for using self-assessment: A survey self-report of Spanish teachers. *Assessment in Education: Principles, Policy & Practice, 21*(3), 365-383. doi:10.1080/0969594X.2014.919247
- Panadero, E., & Brown, G. T. L. (2017). Teachers' reasons for using peer assessment: Positive experience predicts use. *European Journal of Psychology of Education, 32*(1), 133-156. doi:10.1007/s10212-015-0282-5
- Pellegrino, J. W. (2002, February). Understanding how students learn and inferring what they know: Implications for the design of curriculum, instruction and assessment. In *NSF K-12 Mathematics and science curriculum and implementation centers conference proceedings* (pp. 76-92). National Science Foundation and American Geological Institute, Washington, DC.
- Quesada-Serra, V., Rodríguez-Gómez, G., & Ibarra-Sáiz, M. S. (2016). What are we missing? Spanish lecturers' perceptions of their assessment practices. *Innovations in Education and Teaching International, 53*(1), 48-59.

- Quesada-Serra, V. Q., Rodríguez-Gómez, G. R., & Ibarra-Sáiz, M. S. I. (2017). Planificación e innovación de la evaluación en educación superior: la perspectiva del profesorado. *Revista de Investigación Educativa*, 35(1), 53-70.
- Rathbun, G. A., Leatherman, J., & Jensen, R. (2017). Evaluating the impact of an academic teacher development program: practical realities of an evidence-based study. *Assessment & Evaluation in Higher Education*, 42(4), 548-563.
- Rodríguez-Gómez, G. R., Ibarra-Sáiz, M. S. I., & García-Jiménez, E. G. (2013). Autoevaluación, evaluación entre iguales y coevaluación: conceptualización y práctica en las universidades españolas. *Revista de investigación en educación*, 2(11), 198-210.
- Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the culture of teaching and learning through a syllabus review. *Assessment & Evaluation in Higher Education*, 40(7), 898-913.
- De Swert, K. (2012). Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication*, 1-15.
- Tippin, G. K., Lafreniere, K. D., & Page, S. (2012). Student perception of academic grading: Personality, academic orientation, and effort. *Active Learning in Higher Education*, 13(1), 51-61. doi:doi:10.1177/1469787411429187
- van Zundert, M., Sluijsmans, D., & van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, 20(4), 270-279. doi:10.1016/j.learninstruc.2009.08.004
- Wächter, B. (2004). The Bologna Process: developments and prospects. *European Journal of Education*, 39(3), 266-273.
- Wiliam, D. (2011). What is assessment for learning? *Studies In Educational Evaluation*, 37(1), 3-14. doi:10.1016/j.stueduc.2011.03.001

Wold, K. (2013). Collaborative Inquiry: Expert Analysis of Blended Learning in Higher Education. *International Journal on E-Learning*, 12(2), 221-238.

Zabalza, M.A. (2003). *Competencias docentes del profesorado universitario. Calidad y desarrollo profesional*. Madrid: Narcea.

Table 1  
*Assessment instrument used in first vs. fourth-year*

Instrument	1 <sup>st</sup>	4 <sup>th</sup>	Including <i>Final Project</i> and			Excluding <i>Final Project</i> and		
	Year	Year	<i>Practicum</i> (df=1)			<i>Practicum</i> (df=1)		
	%	%	$\chi^2$	$\phi$	$Z_E$	$\chi^2$	$\phi$	$Z_E$
Final Written Exam	70.8	62.3	6.526*	-.090	2.6	.000	†	†
Practical Exam	14.0	12.7	.309	†	†	.081	†	†
Partial Exam	48.4	27.9	34.790**	-.209	5.9	22.432**	-.174	4.7
Portfolio	3.2	5.9	3.541	†	†	1.489	†	†
Individual Assignment	29.2	46.8	26.084**	.181	-5.1	10.681**	.120	-3.2
Group Assignment	20.6	23.4	.899	†	†	3.660	†	†
Practices	61.5	50.4	9.903 **	-.111	3.1	2.759	†	†
Attendance	27.4	26.2	.131	†	†	.025	†	†

\*Significance at .05 level. \*\* Significance at .01 level. Positive  $Z_E$  values mean higher use of that instrument for first-year course. Negative  $Z_E$  values imply higher instrument use in fourth-year course. †Not-shown V are non-significantly different from zero. Not-shown  $Z_E$  values do not reach significance level.

Table 2  
*Assessment instrument used comparing faculties*

Instrument	$\chi^2$	V	$Z_E$
Final Written Exam	64.751**	.196	Ec(+4.1); L(+3.5); EA(-4.8); A(-4.1)
Practical Exam	54.808**	.182	M(+3.7); HS(+2.9); Ec(-3.0); A(-2.6); H(-2.9); Ps(-2.1)
Partial Exam	169.425**	.316	EA(+8.8); HS(+5.9); Pl, Ps, L(-4.2); H(-3.3); C(-2.8); A(-2.7); Ed(-2.5)
Portfolio	31.838**	.137	Ed(+2.4); Pl(+2.5); EA(-3.0); L(-2.6)
Individual Assignment	45.995**	.165	H(+3.3); A(+3.4); Pl(+2.6); Ec, L(-2.5)
Group Assignment	56.708**	.183	Ed(+5.0); C(+2.6); EA(-3.7); S(-3.2)
Practices	61.882**	.191	EA(+4.4); Pl(-5.5); A(-2.4); C(+2.1)
Attendance	59.360**	.187	Pl(+3.2); H(+3.2); A(+2.7); Ed(+2.1); EA(-5.8)

\*Significance at .05 level. \*\* Significance at .01 level.  $Z_E$  values equal or greater than 2 in absolute value are shown between brackets after the faculty code. A positive value refers to a significantly larger usage of the instrument. A negative value refers to a significantly less usage of the instrument. Faculties levels' codes are: M=Medical & Health Science; C=Communication, Ec=Economics, L=Law & Administration, A=Arts, EA=Engineering & Architecture, HS=Hard Sciences, Ed=Education, Ps=Psychology & Social Labour, H=Human Studies, Pl=Philology & Languages.

Table 3

*Best latent class model main properties for each academic division*

Division	N	Classes	BIC	p*	%
Arts & Humanities	315	3	2036.3216	.0560	41.57
Hard Sciences	173	2	1010.5309	.0380	59.11
Health Sciences	159	2	1021.0864	.1920	56.52
Social Sciences	554	4	3320.5307	.0400	38.72
Engineering & Architecture	492	3	3035.8749	.2100	50.82

*N* is the sample size of each branch level. *Classes* refers to the number of clusters found in the best model of each branch level. \* Refers to a bootstrapped p. % refers to first cluster size.

Table 4

*Descriptive analysis of Self- and Peer grading in every assessment instrument.*

		FWE	PE	PaE	Po	IA	GA	Pr	Att
Self-grading	N	5	1	7	1	6	5	11	6
	%	11.9	2.4	16.7	2.4	14.3	11.9	26.2	14.3
Peer grading	N	1	2	1	1	7	9	7	3
	%	3.7	7.4	3.7	3.7	25.9	33.3	25.9	11.1

*FWE=Final Written Exam, PE=Practical Exam, PaE=Partial Exam, Po=Portfolio, IA=Individual Assignment, GA=Group Assignment, Pr=Practices, and Att=Attendance.*

Table 5

*Use of assessment instruments in each cluster*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7
<b>Instrument</b>	N (%)	N (%)					
Final Written Exam	215 (100)	253 (76.0)	35 (12.5)	185 (73.7)	173 (100)	53 (32.5)	266 (100)
Practical Exam	0 (0)	244 (73.3)	0 (0)	0 (0)	0 (0)	14 (8.6)	0 (0)
Partial Exam	0 (0)	95 (28.5)	248 (88.3)	0 (0)	173 (100)	137 (84.0)	0 (0)
Portfolio	0 (0)	74 (22.2)	0 (0)	0 (0)	0 (0)	3 (1.8)	0 (0)
Individual Assignment	67 (31.2)	79 (23.7)	76 (27.0)	251 (100)	33 (19.1)	64 (39.3)	0 (0)
Group Assignment	215 (100)	68 (20.4)	107 (38.1)	0 (0)	0 (0)	45 (27.6)	0 (0)
Practices	111 (51.6)	137 (41.1)	178 (63.3)	116 (46.2)	105 (60.7)	86 (52.8)	266 (100)
Attendance	85 (39.5)	61 (18.3)	0 (0)	68 (27.1)	0 (0)	163 (100)	82 (30.8)
Total	215 (100)	333 (100)	281 (100)	251 (100)	173 (100)	163 (100)	266 (100)

Table 6

*Presence of clusters by academic division*

Cluster	Academic division										Total N (%)
	Arts and Humanities		Hard Sciences		Health Sciences		Social Sciences		Engineering and architecture		
	N (%)	$Z_E$	N (%)	$Z_E$	N (%)	$Z_E$	N (%)	$Z_E$	N (%)	$Z_E$	
1	40 (12.7)	0	11 (6.4)	-2,6	19 (11.9)	-0,3	114 (20.6)	6,8	31 (6.3)	-5,1	215 (12.7)
2	64 (20.3)	0,3	39 (22.7)	1	47 (29.6)	3,3	85 (15.3)	-3,1	98 (19.9)	0,2	333 (19.7)
3	39 (12.4)	-2,5	32 (18.6)	0,5	15 (9.4)	-2,7	73 (13.2)	-3,1	132 (26.8)	6,7	291 (17.2)
4	73 (23.2)	4,6	15 (8.7)	-2,4	28 (17.6)	1	75 (13.5)	-1	60 (12.2)	-2	251 (14.8)
5	15 (4.8)	-3,5	29 (16.9)	3	9 (5.7)	-2	46 (8.3)	-1,8	74 (15.0)	4,2	173 (10.2)
6	41 (13.0)	2,3	26 (15.1)	2,6	16 (10.1)	0,2	41 (7.4)	-2,2	39 (7.9)	-1,5	163 (9.6)
7	43 (13.7)	-1,1	20 (11.6)	-1,6	25 (15.7)	0	120 (21.7)	4,7	58 (11.8)	-2,8	266 (15.7)
Total	315 (100)		172 (100)		159 (100)		554 (100)		492 (100)		1692 (100)





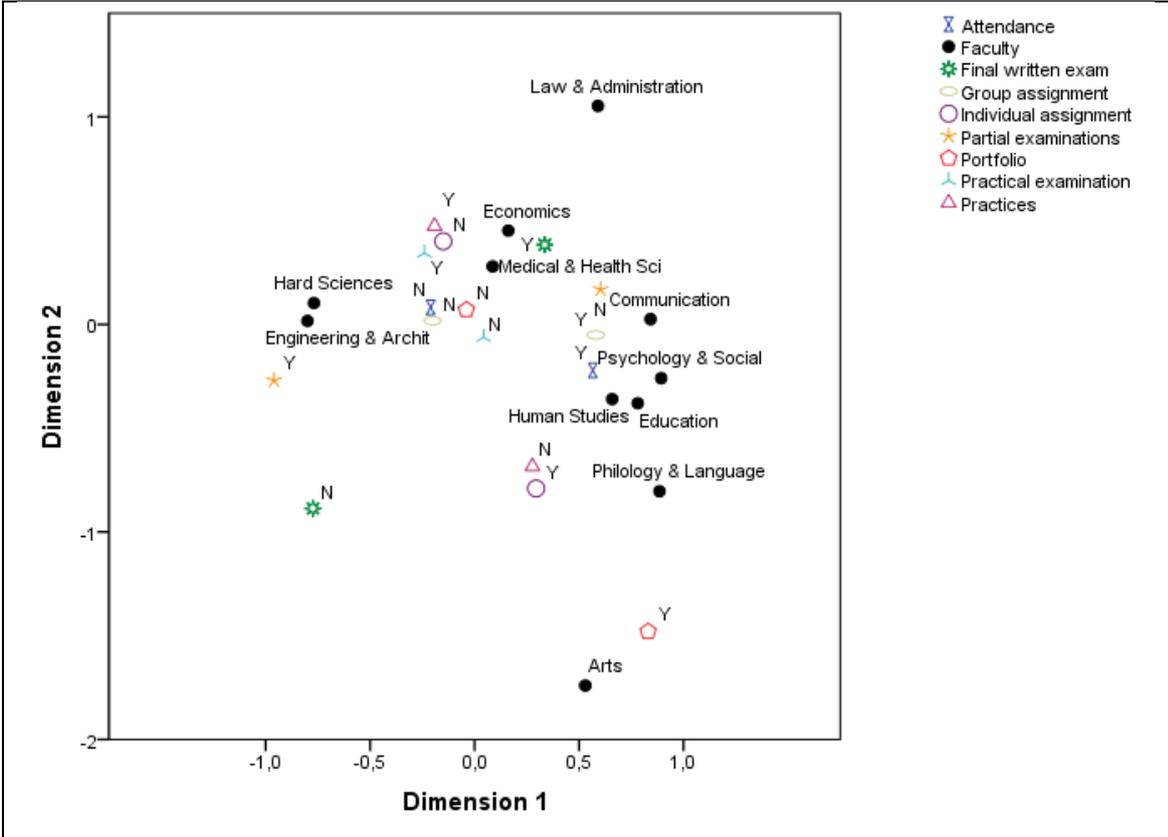


Figure 3. Assessment instrument utilized comparing faculties.

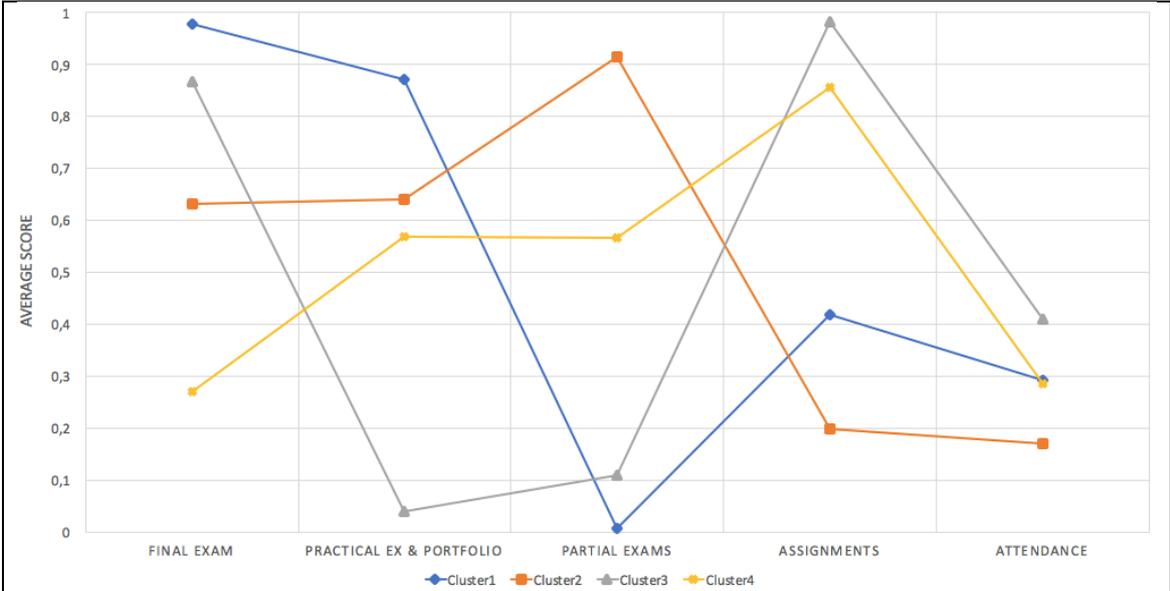


Figure 4. Assessment instruments' utilization profiles: 4- latent class model.

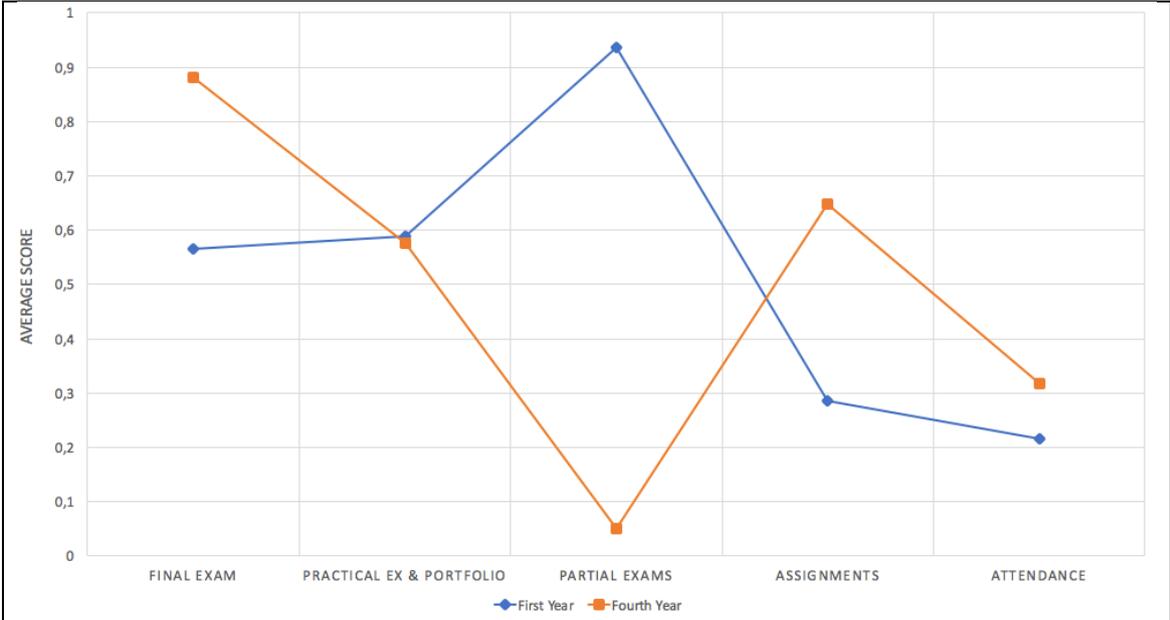


Figure 5. Assessment instruments' utilization profiles: main clusters for first and fourth-year courses.

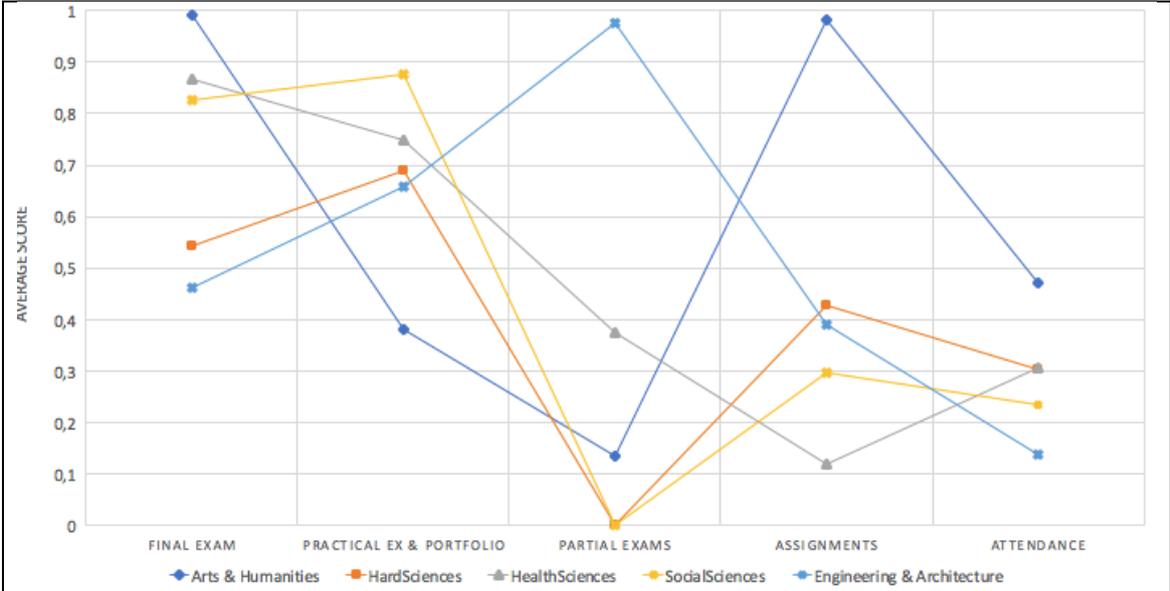


Figure 6. Assessment instruments' utilization profiles: main clusters for academic division.