Abstract

This study examined differences in the assessment criteria used by the US and Spanish university instructors to assign course grades. The US sample included two hundred and fifty course syllabi (159 from universities and 91 from four-year colleges) developed by randomly selected instructors from five academic disciplines (Education, Math, Science, Psychology, and English). Spanish data set included 175 syllabi, chosen from the national data base from the same five domains. The results revealed that university instructors employed a number of criteria when assigning course grades, with the US instructors relying equally on process and product criteria, and Spanish instructors using a higher proportion of product indicators. We also found that self- and peer assessment were used scarcely between the two countries, and that no syllabi employed progress criteria. Theoretical, practical, and policy implications are discussed along with avenues for further research.

How do Instructors Mark? Assessment Criteria Across the US and Spanish Universities

Assessment of student learning is at the core of university education. After all, decades worth of scholarly writings and practical field observations provided us with compendiums of literature, all of which comes to a similar conclusion: Assessment matters. For example, research has consistently demonstrated that taking tests could be highly effective for promoting academic achievement of students and can help professors adjust their instructional approaches to ensure optimal learning (Bjork, Storm, & deWinstanley, 2010; Kornell, Hays, & Bjork, 2009; Richland, Kornell, & Kao, 2009). There is extensive evidence showing that assessments can help students learn and encode important concepts that are taught in greater depth in subsequent lessons (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013) and that the feedback given to students after taking an assessment can enhance retention of information and lead to subsequent improvement of skills and knowledge (Hattie, 2012; Richland et al., 2009).

Creating high quality assessments, however, is an onerous undertaking that requires time and a high degree of pedagogical acumen that some university instructors may not have developed. Across the world, not all university instructors receive formal pedagogical training and, hence, may not be well-versed in the variety of available assessment tools and their optimal implementation (Postareff, Lindblom-Ylänne, & Nevgi, 2007). Further, how to translate results of individual assessments into meaningful cumulative grades adds another layer of challenge for many professors. Recently, several studies attempted to describe higher education assessment profile of individual countries. In their investigation, Panadero, Fraile, Fernández Ruiz, Castilla-Estévez, and Ruiz (2019) examined 1,693 syllabi selected at random out of a database of all the Spanish public universities syllabi to paint a comprehensive picture of the Spanish higher education assessment practices. The researchers found that university instructors used a variety

of instruments across the four years of schooling, that examination scores counted the most toward the final grade, that conventional approaches were still the most prevalent ones, that most variation in assessment were explained by differences on faculty/academic divisions, and that instructors refrained from formally using peer- and self-assessment in their grade assignment. Panadero et al. (2019) called for further improvement of assessment practices particularly emphasizing the fact that no differences were discovered between approaches used with year one and year four students, despite obvious differences in their knowledge, skill development, and general adaptation to the university setting.

Similarly, Lipnevich, Guskey, Murano, and Smith (2020) examined assessment approaches across universities and colleges in the US as reflected in course syllabi that came from five academic disciplines. The researchers discussed main criteria that college and university instructors used to assign grades, as well as main framing systems the instructors use for calculating final course grades (i.e., the open point system, the 100-point system, and the percentage system; see Smith & Smith, 2009, for details). Lipnevich et al. (2020) found that there were significant differences across academic disciplines in the usage of various criteria, as well as framing systems, with education and English relying less on final exams and other product criteria, as compared to psychology, math, and science – a finding that is also consistent with that reported in Panadero et al. (2019). Finally, a third line of work exploring institutional use of assessment practices at a national level has been conducted by Jessop and colleagues (Jessop & Maleckar, 2016; Jessop & Tomas, 2017; Tomas & Jessop, 2019). These scholars, using the Transforming the Experience of Students through Assessment (TESTA) database, offered important insights into assessment in institutions of higher education in the UK. Among other conclusions, the researchers found that the assessment practices were not improving the

students' perceptions of learning (Jessop & Maleckar, 2016), that there was a wide range of assessment practices used in the UK universities, and that those practices were predominantly summative (Jessop & Tomas, 2017). They also reported that teaching-intensive universities presented a broader number of assessment practices when compared to research intensive institutions (Tomas & Jessop, 2019).

To our knowledge, cross-national trends in assessment in higher education remain largely unexamined, and such comparisons carry a wealth of information for researchers, practitioners, and policy makers alike. In the current globalized world, as PISA and TIMSS attest, examining cross-cultural trends is critical in order to be able to move our collective thinking forward (e.g., Hopfenbeck, Lenkeit, El Masri, Cantrell, Ryan, & Baird, 2017). The current study intends to do just that. We aimed at drawing comparisons between assessments that the US and Spanish university instructors utilize, delving into criteria they use to assign final grades. We examined specific assignments they grade and explored differences in assessment approaches among academic disciplines.

An additional important aspect of this study is that the Spanish university system is ruled under the Bologna Process, a European Union agreement intended to equalize standards and quality of higher education credentials across countries (Wächter, 2004; Westerheijden et al., 2008). One of the pillars of this agreement is the emphasis on continued improvement of evaluation of student attainment and decreasing reliance on more traditional assessment methods. Therefore, due to this agreement and intended alignment of assessment approaches in universities across the European Union (European Commission/EACEA/Eurydice, 2018; Karran, 2005), assessment practices in Spain could be viewed as generally representative of those in other European countries, making comparisons with the US particularly meaningful.

Characteristics of assessment in higher education

There is a wide range of instructional activities that can serve for assessment purposes in a typical university course (Angelo & Cross, 1993; Brookhart & Nitko, 2008). Depending on their nature and their design, the types of assessment used in the classrooms gauge different cognitive levels (Bloom's taxonomy; Bloom et al., 1956) and aspects of learning (Brookhart, 2014). The types of assessments chosen by the university instructors and how they are used throughout the semester offer information about whether the lecturer is tracking the development of the students or not. The classical traditional example would be a course in which the assessment is conducted only through a final examination offered at the end of the semester. With this course structure, the instructor would not be able to track students' progress and would rely on a single indicator of performance. A more formative approach would include the use of varied assessment types presented to students throughout the semester, thus providing the instructor with multiple assessment points and, potentially, monitoring progress.

Generally, universities provide a lot of freedom for instructors in terms of their course structure and assessment design, so there may be a wide spread of assignments that professors use across and within disciplines that culminate in the final, feared or coveted, grade. Studies show that instructors may make assignment choices that reflect on what they experienced as students, thus perpetuating potentially ineffective practices (Allen, 2005). Therefore, it is crucial to explore what assessment types are used in the US and Spanish university contexts to offer recommendations for improvement. It is also critical to examine the proportion of each assessment type's contribution to the final grade across disciplines and between countries. This may provide us with useful information that may potentially lead to a range of policy implications.

Assessment Criteria: Product, Process and Progress

In addition to the variety of assessment types that university instructors use, research also reveals that they vary greatly in the procedures that are used to combine this evidence in assigning course grades (Cizek, Fitzgerald, & Rachor, 1996; Lipnevich et al., 2020; Panadero et al., 2019). Two reasons have been suggested to account for this variation. The first one is in the lacking clarity or agreement about the purpose of grading (Brookhart, 2011). Decisions about what evidence to use in determining students' grades are difficult to make when the instructors are not sure whether final marks can serve functions beyond that of a summary of vague and uneven levels of mastery (Brookhart & Nitko, 2008).

The second reason relates to how most college and university course grades are reported. The vast majority of reporting forms record a single grade for each course. This requires professors and instructors to combine, in some cases, a multitude of sources of evidence they gather throughout the course into a single symbol (Brookhart, 1991, 2009; Cross & Frary, 1999). Even when instructors report the weighting and computing strategies they employ to derive the final score, the final, amalgamated, grade remains to be messy and barely interpretable (Guskey, 2006; 2019). An alternative to amalgamated grades is differentiated grading procedures, when different categories of skills get assessed separately. To this end, Guskey (2006; 2019) suggested three broad categories of assessment criteria that include *product*, *process*, and *progress*.

Product criteria relate to *what* students know and are able to do at a specific point in time. Grades based on product criteria communicate a summative evaluation of student achievement (Brookhart, 2009; O'Connor, 2009). Professors and instructors who use product criteria normally base grades on final examination scores, final products (reports or projects), overall assessments, and other significant demonstrations of learning. Exams appear to be the

most prevalent examples of product criteria in higher education. Although exams represent relatively objective measures of student skills, there are many serious considerations that should be taken into account before using exams as the main source of evidence to assess students' course performance. Test anxiety, for example, has been shown to relate to decreased performance on exams (e.g. Cassady & Johnson, 2002; Seipp, 1991). Alternative approaches to product grades do exist, and are often favored by students. In a study by Turner, Roberts, Heal, & Wright (2013) 82% of students who made an oral presentation as a summative assessment in a teacher education program reported that they preferred this type of summative assessment in comparison to a traditional written exam or essay.

Alternatively, *process criteria* are used by professors and instructors who believe that grades should reflect not only students' final achievement but also *how* they got there. Professors who consider students' effort or work habits when assigning grades are using process criteria. Examples of process-based assessments include classroom quizzes, punctuality in turning in assignments, online posts, class participation, or attendance. Guskey (2019) separates process criteria into three categories. The first one is called *learning enablers* and includes formative assessments, homework, class participation and refer to ongoing indicators of student engagement with the course. The second set comprises *social and emotional characteristics*, whereas the final category is *compliance*. The latter is assessed by students' turning in assignments on time, punctuality, and professional behavior in the classroom. By assessing process, instructors are able to capture aspects of student performance not directly included in product criteria. Furthermore, this information is important because *process* criteria are explicitly deemed as critical outcomes of any university education (Oswald et al., 2004) with ten out of eleven specified categories representing indicators of process criteria.

Finally, *progress criteria* are based on how much students gain from their learning experiences. Other names for progress criteria include "learning gain", "improvement scoring", "value-added learning", and "educational growth" (Wiggins, 1996). Examples of progress indicators include various indices of improvement, changes from pre- to post-assessments, and revisions of work. In this study we will examine the use of the trichotomy of grading criteria to explore how it is currently being applied in the US and Spanish instructional contexts.

Peer- and self-assessment and differences among disciplines

One of the main claims in formative assessment is that students' involvement in peer and self-assessment would be positive for their learning (Black & Wiliam, 1998). Numerous studies have demonstrated significant positive effects of peer- and self-assessment on performance and self-regulated learning (e.g. Brown & Harris, 2013; van Zundert, Sluijsmans & van Merriënboer, 2010). However, research comparing the implementation of peer and self-assessment have found that higher education lecturers used it less than primary and secondary education teachers (e.g. Panadero & Brown, 2017). This is an interesting and a somewhat disappointing finding as university students may be far better prepared for being involved in these kinds of assessment due to greater metacognitive sophistication. Therefore, it is vital to explore how and to what extent peer and self-assessment is implemented in the two countries.

Finally, different academic traditions (e.g. psychology, education) may have different approaches to assessment (Pellegrino, 2006). Previous studies have clearly demonstrated that this is the case (Jessop & Maleckar, 2016; Lipnevich et al., 2020; Panadero et al., 2019). These differences are informative as they might inform policy decisions at the institutional and country

level. Therefore, we wanted to examine how patterns of differences among disciplines compare across the two countries.

The Current Study

In this study we aimed to explore the assessment practices university instructors use and compare these between the US and Spain. To this end, we explored the variation in the use of product, process, and progress criteria in assigning course grades in introductory level courses. We defined assessment criteria as *any products, practices, and procedures that are counted toward students' final grade*. We were interested in the percentage of course grades that are determined by performance on specific types of assignments. For each source of variation, we were also interested in any significant differences among academic disciplines and between countries. Finally, we also investigated the extent to which students are expected to be formally involved in assessment, as indicated by assignments involving self- and/or peer assessment. We examined the differences between Spain and the US in the following:

RQ1: What is the frequency of use of different assessment types and product/process/progress criteria?

RQ2: How is each assessment type weighted in the final grade?

RQ3: What is the involvement of students in assessment (i.e., peer- and self-assessment)?

RQ4: Can the likelihood of assessment types and product/process/progress criteria be predicted by country and academic discipline?

Method

Syllabi Analysis

A syllabus is an official document in which course instructors present descriptions of the course content, along with expectations, responsibilities, assignments, and criteria for evaluation

(Stanny, González, and McGowan, 2015). Syllabi are viewed as contracts between an instructor and a student, and it is expected that all syllabi would define assessment approaches and include information about types of assessment (e.g. examinations, essays), along with weights of each assessment in the final course grade. Syllabi have been used in prior studies to explore a range of research questions varying from the alignment of syllabi with learning outcomes to the evaluation approaches in Spain and the US, among others (Bers, Davis, & Taylor 2000; Cashwell & Young 2004; Lipnevich et al., 2020; Panadero et al., 2019; Rathbun, Leatherman, & Jensen, 2017). In both the US and Spain, instructors have substantial freedom in designing their courses, but must adhere to university policies in preparing their course syllabi. In most universities, departing from the policies delineated in the syllabus is considered a contractual violation and instructors can be penalized for not adhering to the syllabi. Hence, syllabi are a valuable source for identifying what grading practices may look like within and across disciplines and between countries.

Materials and Procedure

USA. Course syllabi (n = 250) were gathered from randomly selected college and university websites to determine the differences in grading criteria used by professors and instructors in various academic disciplines. The colleges and universities were chosen from five US regions of South, North, West, Midwest, Northeast. From each institution's website, one introductory level course at the undergraduate level was randomly selected in each of five academic disciplines: English, mathematics, science, psychology, and education. If the syllabi of the selected course did not include a detailed description of the criteria by which course grades were assigned, another introductory course within that department and institution was selected. If universities did not have one syllabus per academic domain available, we contacted department

chairs and requested representative syllabi in a specific domain. Our final sample included 50 syllabi of introductory level courses from each of the five academic disciplines. One hundred and fifty-nine of the selected syllabi came from large, comprehensive universities and 91 were from smaller, four-year colleges. These proportions approximate the numbers of students attending such institutions overall. All institutions were in the United States. Institutions awarding Ph.D. degrees were considered "universities", whereas all others were considered "colleges[1]". Syllabi were independently coded by two raters to establish high inter-rater reliability of <0.85.

Spain. We used part of the database created for the study of Panadero et al. (2019) that included 1,693 syllabi from all public universities in Spain with undergraduate studies (N = 48) and ensuring that a minimum of 30 cases from each university were selected. Those selected cases were coded by three coders into eight categories that corresponded to types of assessment. The Krippendorff's alpha coefficient among the three coders ranged between .68 (for group assignments) and .86 (practical examinations), and was above the recommended minimum of .60 (De Swert, 2012). For this study, we used data from 149 syllabi that came from the first year courses, and we randomly selected and coded 26 more, to reach a minimum of 20 per each academic discipline. The final Spanish sample comprised 173 syllabi.

Data merging. The US and Spanish were then merged. Every syllabus contained multiple assessments that counted toward the final grade. The US data set included 25 categories, because we had recorded *all* assessments that were counted toward the final grade. These were also coded into the three categories of product, process, and progress criteria. The Spanish data set included aggregate categories. Hence, we used the Spanish coding of assessment types to recode the US data, and we used the US codes for product, process, and progress criteria and academic

[1] The data upon which the findings of this study are based are available on request from the corresponding author.

disciplines to recode the Spanish data set. Every individual assessment was coded into only one category. Assessments that were graded on Pass/Fail basis were always included into a graded category. For example, an individual homework may have been grade as Pass/Fail, but all homework assignments accounted for 15% of the final grade. The following is a more detailed explanation of these components (also see Appendix A for explanation and examples).

Assessment types. Spanish syllabi were coded into eight categories: (1) final written exam, as the type of a comprehensive exam that students take at the end of the semester; (2) practical exam, which requires students to demonstrate their practical skills (e.g. sport skills), typically taken at the end of the semester; (3) partial exam/s, taken by students throughout the semester; (4) portfolio assessments, that represent compilations of exemplars of student work; (5) individual assignments, that include essays and reports; (6) group assignments, involving group collaboration; (7) practices, that include individual small tasks throughout the semester (e.g. homework, quizzes); and (8) attendance and participation. The US syllabi were recoded into the eight Spanish categories of assessment types, which were used for this study (see Appendix A). Both the US and Spain coded self- and peer-assessments, as well as self- and peer-grading, as explicitly mentioned on the syllabi.

Product, process, and progress criteria. According to Guskey (2006), assessment practices can be grouped into the three categories of product, process, and progress criteria. The US data set contained categories of these criteria, and the Spanish data set was then recoded to allow for comparisons. Examples of product criteria included final written exams, practical exams, partial exams, and portfolios; process criteria contained individual assignments, practices, and attendance, and progress criteria included any indicators of pre- and post-performance

assessments. Appendix A contains all assessment types across both countries and shows that we did not find a single example of progress criteria in either country.

Academic disciplines. US syllabi came from five disciplines into which we also coded Spanish syllabi. These included science, math, psychology, education and English/Spanish language arts.

Analytic Plan

Statistical analyses were conducted to answer the research questions of the study by using IBM SPSS Version 24. The descriptive statistics of all study variables were checked for skewness and distribution of scores in order to best examine multivariate relations and meet the assumptions for regression analyses. For the descriptive research questions (RQs 1-3) descriptive statistics including central tendency, skewness, and kurtosis were generated. The final research question (RQ 4) required multi-step analyses. First, descriptive statistics were explored by academic discipline as well as by country on each assessment type. Then, bivariate associations using chi-squared statistics were run to distinguish whether the presence of each of the assessment types was associated with country (US, Spain) as well as academic discipline. Finally, multivariate relations were examined using ordinary least squared (OLS) regressions for the linear outcomes (i.e., percentage of product oriented assessments) and logistic regressions for the presence assessment types (e.g., 1 = assessment type was utilized in syllabus, 0 = assessment type was not utilized in syllabus). Interaction terms between country and academic discipline were added onto the regression models, where country was dummy coded and the US was the reference category (US = 0) and “education” served as the reference category for academic discipline. The process assessment taxonomy was not used as an outcome in multivariate

analyses because it had an inverse relation to product assessment types (r = -1.00), making the results redundant.

Results

RQ1: What is the frequency of use of different assessment types and product/process/progress criteria?

Descriptive analyses of assessment types for the total sample indicated that the most frequently utilized assessment type was partial exam accounting for 28.21% of the total course grade. This was followed by final written exams (26.68%) and practices (23.55%). These three components, collectively, contributed to over 78% of students' course grade across the disciplines (see Table 1). These frequencies demonstrate that there were more product-oriented assessment types (57.64%) than there were process-oriented assessment types, irrespective of country and department. Progress criteria were not mentioned in either sample.

Table 1

Descriptive Statistics of Assessment Types by Percentage for Overall Sample and by Country

	Total sample (N = 423)						United States (n = 250)						Spain (n = 173)					
	M	Mdn	SD	Min	Max	Skew	M	Mdn	SD	Min	Max	Skew	M	Mdn	SD	Min	Max	Skew
Assignment Type																		
Final written exam	26.68	20	27.02	0	100	0.71	14.10	15	14.56	0	60	0.53	44.86	50	30.35	0	100	-0.37
Practical exam	1.51	0	6.84	0	70	5.65	0.07	0	0.87	0	12.82	13.52	3.58	0	10.31	0	70	3.45
Partial exam/s	28.21	25	29.46	0	100	0.69	33.33	31.88	27.87	0	100	0.4	20.8	0	30.19	0	100	1.24
Portfolio	1.24	0	6.87	0	75	6.71	1.56	0	7.9	0	75	6.21	0.78	0	5.01	0	40	6.97
Individual assignments	12.06	0	21.68	0	100	2	17.21	0	24.97	0	100	1.43	4.61	0	12.48	0	100	3.97
Group assignments	2.05	0	8.7	0	100	6.36	0.90	0	4.86	0	40	6.2	3.71	0	12.12	0	100	4.89
Practices	23.55	20	20.83	0	100	0.92	27.07	25	21.92	0	100	0.78	18.46	17	18.02	0	100	1.08
Attendance	4.72	0	70.42	0	50	2.14	5.76	0	9.17	0	50	1.71	3.20	0	6.87	0	50	3.22
Taxonomy																		
Product	57.64	65	30.15	0	100	-0.61	49.06	51	31.90	0	100	-0.24	70.02	70	22.27	0	100	-1.00
Process	42.37	35	30.15	0	100	0.61	50.94	48	31.90	0	100	0.24	29.98	30	22.27	0	100	1.00

RQ2: How is each assessment type weighted in the final grade?

Descriptive analyses of assessment types by country, U.S. and Spain, revealed discrepancies in the eight types of assessment (Table 1). Irrespective of academic discipline, syllabi from Spain allocated more weight in the final course grade to (1) final written exams, (2) practical exams, and (3) group assignments. It is worth noting that group assignments were utilized quite infrequently and the average point allocation for group assignments was only 2% for the total sample. The largest country discrepancy in grading allocation as specified in college syllabi was observed in final written exams, where syllabi from Spain utilized this assessment type (44.86%), resulting in almost a 30 percentage-point gap between countries. Syllabi from Spain revealed greater usage of product grading criteria (70.02%) when compared to process oriented grading criteria (29.98%), whereas syllabi from the US indicated a closer distribution of process and product grading criteria (50.94% for process and 49.06% for product).

Assessment types by academic discipline are detailed in Table 2 and indicate that psychology (41.27%), science (29.6%), and mathematics (38.24%) syllabi allocated the highest proportions of their final grade to partial exams. For the second most frequent assessment type, final written exams, 33.3% were allocated by science syllabi and 33.32% by mathematics syllabi. Practice assessments were the third most frequently used assessment type, and was equally utilized by all academic disciplines, accounting for the final grade in a 20.34% - 22.8% range.

Table 2

Descriptive Statistics of Assessment Types by Academic Discipline

	Science (n = 132)						English/Spanish Language Arts (n = 83)						Mathematics (n = 70)					
	M	*Mdn*	*SD*	*Min*	*Max*	*Skew*	*M*	*Mdn*	*SD*	*Min*	*Max*	*Skew*	*M*	*Mdn*	*SD*	*Min*	*Max*	*Skew*
Final written exam	33.33	30	27.74	0	90	0.32	22.2	10	27.71	0	100	0.96	33.32	30	22.98	0	100	0.73
Practical exam	3.28	0	10.15	0	70	3.88	1.57	0	6.94	0	40	4.83	0.43	0	2.66	0	20	6.66
Partial exam/s	29.6	29.86	29.2	0	100	0.59	13.47	0	23.48	0	100	2.01	38.24	40	24.26	0	100	-0.15
Portfolio	0.53	0	4.34	0	40	8.28	3.32	0	11.57	0	75	4.35	1.29	0	7.6	0	50	5.9
Individual assignment	4.87	0	13.07	0	100	4.15	27.57	20	31.63	0	100	0.79	1.65	0	5.78	0	30	3.67
Group assignment	2.2	0	10.29	0	100	7.36	1.71	0	6.5	0	40	4.36	0.57	0	2.89	0	20	5.55
Practices	22.22	22.5	19.01	0	100	0.83	22.69	16.67	25.39	0	100	1.15	22.8	20	14.46	0	80	1.05
Attendance	3.97	0	8.96	0	50	3.03	7.47	5	8.85	0	35	1.05	1.71	0	5.23	0	30	3.93
Product	66.74	70.00	23.74	0.00	100	-0.88	40.56	35	34.28	0	100	0.20	73.27	75	17.18	0	100	-1.52
Process	33.25	30	23.74	0.00	100	0.88	59.43	65	34.28	0	100	-0.20	26.72	25	17.17	0	100	1.52

(Table 2 continued)

	Psychology (n = 68)						Education (n = 70)					
	M	*Mdn*	*SD*	*Min*	*Max*	*Skew*	*M*	*Mdn*	*SD*	*Min*	*Max*	*Skew*
Final written exam	22.56	19.09	26.93	0	100	1.10	16.80	0.64	20.34	1.26	20.52	3.65
Practical exam	0	0	0	0	0		0	0	11.85	0	12.5	0
Partial exam/s	41.27	40	35.31	0	100	0.16	24.23	3.99	25.20	5.33	22.34	10.88
Portfolio	0	0	0	0	0		0	0	0	0	0	0
Individual assignment/s	9.07	0	16.01	0	80	2.23	85	30	90	30	70	60
Group assignment/s	2.04	0	9.17	0	66.67	5.87	1.32	6.66	1.15	4.84	0.72	3.49
Practices	20.44	15	21.91	0	100	1.04	16.80	0.64	20.34	1.26	20.52	3.65
Attendance	4.62	0	8.1	0	30	1.68	0	0	11.85	0	12.5	0
Product	63.82	70	28.56	0	100	-0.80	39.05	40	29.34	0	100	0.22
Process	36.18	30	28.52	0	100	0.80	60.95	60	29.34	0	100	-0.22

RQ3: What is the involvement of students in assessment (i.e., peer- and self-assessment)?

In the US sample, we encountered only 9 mentions of peer- and/or self-grading and assessment across the sampled syllabi. In the Spain sample, we encountered 14 mentions of such grading criteria. Some of these mentions occurred multiple times within a syllabus, which resulted in less than 1% of syllabi allocating any type of course grading to peer- and self-assessment. Due to the low frequencies of grade allocations to peer- and self-assessments, this outcome was not further explored in our analyses. The relative absence of such forms of assessment is highlighted in the discussion as implications for assessment in higher education.

RQ4: Can the likelihood of assessment types and product/process/progress criteria be predicted by country and academic discipline?

Given the non-normal distributions of the proportion of each assessment type, we were constrained to examine the presence of each assessment type (1 = present/proportion of final grade allocated to assessment type, 0 = absent/proportion of final grade not allocated to assessment type) in bivariate and subsequent multivariate analyses. Associations of the presence of assessment type by country and academic discipline are presented in Table 3. Results indicate statistically significant differences by country in allocation to each of the assessment types, with the exception of portfolio assignments. Results also indicated statistically significant differences by academic discipline in the percentage allocation to each of the assessment types, with the exception of group assignments.

Table 3

Associations of the Presence of Assessment Type by Country and Academic Discipline

	Country			Academic Discipline					
	US %	Spain %	χ^2	Science %	English/Spanish %	Math %	Psychology %	Education %	χ^2
Final written exam	55.20	74.86	17.12***	73.48	50.60	85.71	55.71	44.29	39.418***
Practical exam	0.80	15.43	34.65***	13.64	6.02	2.86	2.86	2.86	14.916**
Partial exam/s	72.00	43.43	35.08***	61.36	39.76	81.43	70.00	51.43	32.779***
Portfolio	5.20	4.00	.331	1.52	10.84	2.86	2.86	7.14	11.963*
Individual assignment/s	44.80	18.29	32.30***	19.70	54.22	8.57	37.14	58.57	66.573***
Group assignment/s	4.40	15.43	15.38***	8.33	8.43	4.29	10.00	14.29	4.502
Practices	79.60	67.43	8.05**	73.48	63.86	88.57	64.29	85.71	20.842***
Attendance	38.40	28.57	4.41*	26.52	53.01	14.29	34.29	47.14	33.987***

Note. *$p < .05$; **$p < .01$; ***$p < .001$

Logistic regression results predicting the likelihood of assessment types by country and academic discipline are presented in Table 4. Main effects of country (Spain = 1; U.S. = 0 as a reference category) and academic discipline (Science, English/Spanish, Mathematics, Psychology; with Education = 0 as the reference category) are reported. Practical exam, portfolio, and group assignment assessment types were not examined as outcomes in multivariate analyses due to the low frequencies in these outcomes and due to separation effects in the predictive models (i.e., the low frequencies were predicted by either of the main effects alone). Syllabi from Spain were more likely to contain higher percentage allocations to final written exams, individual assignments, and practical exams, with those odds ratios being *OR* = 4.53, 17.94, and 8.44, respectively. These significant differences by country largely indicate that higher education coursework in Spain is reliant on partial examinations and individual-based assignments to evaluate course completion, irrespective of academic discipline. Sample syllabi from Spain are nearly 18 times more likely to evaluate students based on individual assignments, compared to the US. Main effects of academic discipline indicated that the majority of the differences in assessment types are between mathematics and education, where the OR = 36.42, $p < 0.001$ for exams. This magnitude indicates that instructors of mathematics courses, when compared to education courses, are more likely to allocate larger percentages to examinations.

Country by academic discipline interactions indicated that moderating effects were present when predicting final written exams, partial exams, individual assignments, and practices. No interaction effects were statistically significant for predicting the likelihood of course grading allocation to student attendance. Figure 1 depicts the plots of the interaction effects of country and academic discipline by assessment type. As shown both in Table 4 and in

Figure 1, results indicate that there are differences in the likelihood of allocating a portion of the course grade to final written exams in Spain, depending on the academic discipline, where education has the lowest likelihood and mathematics has the highest likelihood. The likelihood of final written exams for the US sample was relatively similar across academic disciplines. Similar patterns were found for individual assignments and partial exams such that for Spain, the likelihood of allocating course grading to individual assignments and partial grades depended on the discipline.

Table 4

Logistic Regressions Predicting to Likelihood of Each Assessment Type by Country and Discipline

	Final Written Exam				Partial Exam				Individual Assignment			
	b	*SE*	*OR*	*p*	*b*	*SE*	*OR*	*p*	*b*	*SE*	*OR*	*p*
Main Effects												
Country (1 = Spain)	1.51	0.57	4.53	0.008	0.08	0.53	1.08	0.880	2.89	0.71	17.94	0.000
Science (= 1)	1.82	0.45	6.15	0.000	-2.36	0.59	0.09	0.000	2.2	0.46	9.01	0.000
English/Spanish (= 2)	-1.16	0.42	1	1	0.74	0.41	2.1	0.071	0.11	0.46	1.11	0.817
Math (= 3)	2.86	0.56	17.47	0.000	-3.1	0.78	0.05	0.000	3.6	0.62	36.42	0.000
Psychology (= 4)	0.34	0.41	1.41	0.411	-1.74	0.5	0.18	0.000	1.56	0.44	4.75	0.000
Interaction												
Spain*Science	-1.72	0.71	0.18	0.015	2.66	0.78	14.26	0.001	-2.26	0.84	0.1	0.007
Spain*English	0.29	0.76	1.34	0.702	-0.68	0.7	0.51	0.330	-0.7	0.88	0.5	0.424
Spain*Mathematics	-2.61	0.9	0.07	0.004	3.3	1	27.08	0.001	-3.13	1.16	0.04	0.007
Spain*Psychology	1.01	0.98	2.74	0.304	2.58	0.83	13.23	0.002	-2.45	0.91	0.09	0.007

Note. Practical exam, portfolio, and group assignment assessment types were not examined in multivariate analyses due to separation effects in the models. U.S. was the reference category (US = 0) for the country variable such that the estimate represents the difference between syllabi of Spain from the United States. A positive estimate indicates a higher value for Spain.

(Table 4 continued)

	Practices				Attendance			
	b	*SE*	*OR*	*p*	*b*	*SE*	*OR*	*p*
Main Effects								
Country (1 = Spain)	2.13	0.76	8.44	0.005	1.01	0.57	2.74	0.074
Science (= 1)	1.09	0.71	2.98	0.123	1.01	0.42	2.74	0.016
English/Spanish (= 2)	2.09	0.67	8.07	0.002	-0.16	0.4	0.85	0.687
Math (= 3)	0.31	0.79	1.36	0.696	1.82	0.48	6.16	0.000
Psychology (= 4)	2.26	0.66	9.60	0.001	0.82	0.41	2.28	0.046
Interaction								
Spain*Science	-1.19	0.88	0.31	0.179	-0.72	0.69	0.49	0.296
Spain*English/Spanish	-1.9	0.89	0.15	0.033	-0.5	0.72	0.61	0.487
Spain*Mathematics	-1.08	1.08	0.34	0.317	-0.47	1.01	0.63	0.643
Spain*Psychology	-2.49	0.95	0.08	0.009	-1.05	0.79	0.35	0.184

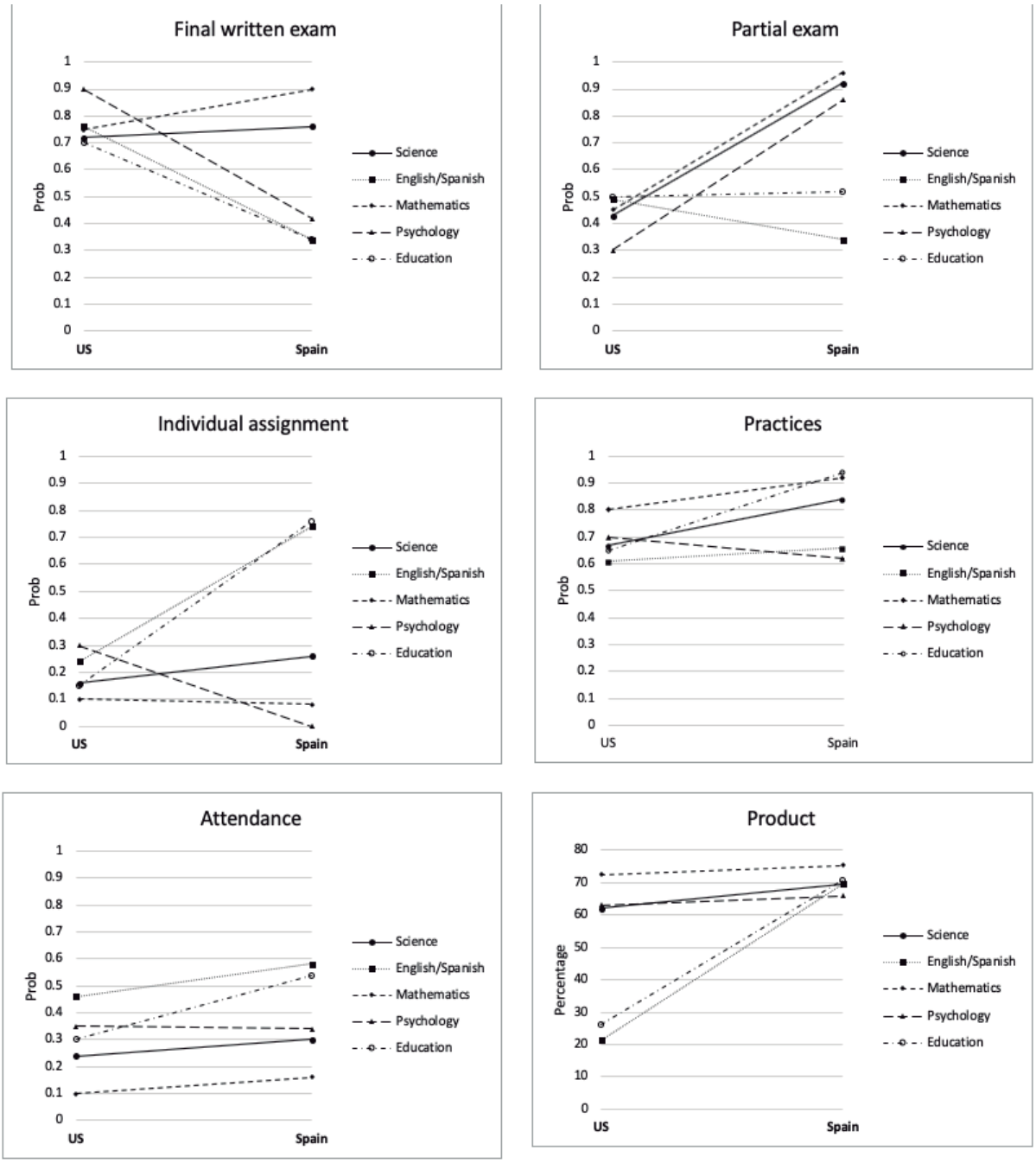

Figure 1. Plots of interaction effects of country and academic discipline by assessment type.

To summarize these results and to map these findings on a taxonomy of assessment in higher education, linear regressions predicting to assessment taxonomies by country and academic discipline were also examined (see Table 5). Results indicated that compared to the US, courses in Spain overemphasized product assessment types, as opposed to process assessment types; $b = 2.79$, 95% CI (0.194, 0.363), $p < 0.001$. Controlling for country, science, mathematics, and psychology courses showed higher grading allocations for product assessment types when compared to education courses– indicating, that these disciplines were more process oriented. Interaction effects were profound, where the interaction of Spain and the science, mathematics, and psychology disciplines was statistically significant. In sum, results indicated that there were significant country and discipline differences in product assessment in higher education and that product assessment grading practices differed by both country and academic discipline.

Table 5

Linear Regressions Predicting to Product Assessment Taxonomy by Country and Academic Discipline

	B	*S.E.*	95% Wald Confidence Interval		Exp(B)	95% Wald Confidence Interval for Exp(B)		*p*
			Lower	Upper		Lower	Upper	
Main Effects								
Country (1 = Spain)	0.279	0.0433	0.194	0.363	1.321	1.214	1.438	0.000
Science (= 1)	0.237	0.0344	0.170	0.305	1.268	1.185	1.356	0.000
English/Spanish (= 2)	-0.066	0.0388	-0.142	0.010	0.936	0.867	1.010	0.087
Math (= 3)	0.307	0.0343	0.240	0.374	1.359	1.271	1.454	0.000
Psychology (= 4)	0.275	0.0350	0.207	0.344	1.317	1.230	1.411	0.000
Interaction					1			
Spain*Science	-0.219	0.0516	-0.320	-0.118	0.803	0.726	0.889	0.000
Spain*English/Spanish	0.095	0.0587	-0.020	0.210	1.099	0.980	1.233	0.107
Spain*Mathematics	-0.272	0.0594	-0.389	-0.156	0.762	0.678	0.856	0.000
Spain*Psychology	-0.302	0.0609	-0.422	-0.183	0.739	0.656	0.833	0.000

Note. U.S. was the reference category (US = 0) for the country variable such that the estimate represents the difference between syllabi of Spain from the United States. A positive estimate indicates a higher value for Spain. Education was set as the reference category for academic discipline. The process assessment taxonomy was not used as an outcome in multivariate analyses because it has an inverse relation to product assessment types, making the results redundant.

Discussion

In this study we attempted to investigate assessment criteria that university instructors in Spain and the US use in their courses. To this end, we examined syllabi that came from five academic disciplines in both countries. We intended to gain better understanding of elements that instructors consider when assessing and determining students' grades. We also strived to uncover the deeper meaning that lies beneath the seemingly shallow nature of single grading symbols that condense a range of assessment tasks into a final grade. Our results revealed a slew of intriguing findings.

Assessment types across countries

First and foremost, across the entire sample, instructors relied on final written exams, partial exams (e.g., midterms), and individual assignments (e.g., reports and essays) the most. Hence, Spanish and US instructors utilize traditional approaches to assessments far more than alternative ones, the trend consistent with that from the UK that showed a predominant use of traditional summative assessment practices (Jessop & Tomas, 2017; Tomas & Jessop, 2019). However, a closer examination of differences between countries revealed some discrepancies. The largest country discrepancy in grading allocation as specified in college syllabi was observed in final written exams, where syllabi from Spain utilized this assessment type 44.86%, resulting in almost a 30 percentage-point gap between countries. This finding is consistent with that of Zabalza (2003) and Panadero et al. (2019) who demonstrated that university instructors in Spain primarily used examinations to assess student performance. It is of note, however, that although written final exams were the leading category of assessment types in Spain, the

remaining 55.14% was allocated to other approaches, with partial exams and practices being among the most common among them.

When it comes to the instructors in the US, they used partial exams with greater frequency compared to their Spanish counterparts, with the country difference of 13% points. Both partial and final exams can be categorized as traditional approaches, wherein more weight is assigned to “objective instruments” (e.g. exams). At the same time, on average, the US instructors were more likely to rely on alternative assignments using a variation of instruments (e.g. exercises + group work + exam) thus increasing the chance of measuring a wider range of behaviors and levels of cognitive processing (Angelo et al., 1993).

In terms of product and process criteria some differences were also revealed. Although for the entire sample a larger proportion of assessments fell into the product category, instructors in the US were more likely to employ process indicators (50.94% in the US, compared to 29.98% in Spain). The purpose of product criteria is to index what students know, as opposed to trying to capture the development of such knowledge, which is the purpose of process criteria. The latter includes *learning enablers*, which are ongoing indicators of student engagement with the course, *social and emotional characteristics*, as well as *compliance* (Guskey, 2006).

It is difficult to conclude whether one approach is better than the other. On the one hand, university education is about gaining knowledge and expertise in a certain domain of study, captured by product criteria. On the other hand, acquiring socio-emotional skills, which are reflected in process criteria (see Guskey, 2006), is no less critical. To investigate which characteristics should be cultivated and subsequently gauged by the university faculty, Oswald et al. (2004) examined educational objectives from 35 colleges and universities in the United States in search for common themes. The obtained statements were parsed into 174 meaningful

fragments and subsequently clustered into 12 criteria of college success. The identified dimensions were: (1) Knowledge, learning, mastery of general principles; (2) continuous learning, intellectual interest, and curiosity; (3) artistic and cultural appreciation and curiosity; (4) multicultural tolerance and appreciation; (5) leadership; (6) interpersonal skills; (7) social responsibility, citizenship, and involvement; (8) physical and psychological health; (9) career orientation; (10) adaptability and life skills; (11) perseverance; and (12) ethics and integrity. It is obvious that the product criteria, as defined by Guskey (2006), are represented by a single category, whereas the remaining eleven categories are indicators of process criteria. Similarly, Casner-Lotto, Barrington, and Wright, (2006) interviewed 400 employers about skills that college graduates should possess. Basic knowledge and skills were deemed as critical for success, however, employers emphasized that non-academic skills were as important in defining success. Hence, it is encouraging that instructors in Spain and the US include both process (e.g., attendance) and product (e.g., exams) indicators. We must note that no syllabi across the two countries utilized progress indicators. Progress criteria are typically covered by individual individualized educational plans (IEPs) that students with learning difficulties receive. Music and sports instructors also use a variety of progress criteria, indicating students' improvement relative to their prior performance (Parkes, 2018), which is very different from the remainder of academic domains that do not use these indicators. Encouraging instructors to employ value-added indicators could benefit student self-perception and improve their learning (Guskey, 2006). Having said that, we do acknowledge that progress criteria would be easier to incorporate at the secondary or primary school level, where teachers stay with students for the duration of the entire year, if not longer. University instructors may only have students for one semester, and this time may not be enough to adequately develop progress indicators.

In general, combining process and product criteria can be an acceptable solution. For example, if we look at the Spanish assessment types that fall into the product category (Table 1) three out of four are exams. The nature of exams is such that the time to complete them is limited, there is typically no opportunity for revisions after receiving the score, and the score usually accounts for a significant portion of the final grade. These characteristics of exams limit the possibility of capturing process criteria, which are usually more hands-on, develop throughout the curriculum, and students are at times given the opportunity to revise after receiving feedback. Process-based assessments, however, are usually more time consuming for the teachers to correct and may be quite subjective or even binary (e.g., attendance). Therefore, a combination of both is probably an adequate instructional decision under the current conditions. For example, in many Spanish public universities students have the right to take the final exam even if they have not attended the course. This was established as a right for those students that cannot attend classes during the regular lecture hours. Another constraint is the limited human resources: instructors' time is limited and a combination of process assessment types and product by exam might be a reasonable solution to a rather complex equation. A final potential influence could be the high modularization of higher education programs that leads to more intensive summative assessment practices as hypothesized in the UK (Jessop & Tomas, 2017; Tomas & Jessop, 2019).

How is the final grade calculated

Despite the rather positive finding indicating that instructors utilize a mixture of criteria, what we also found is that regardless of differentiating process and product criteria into various assignments, 100% of examined courses in the US and Spain used amalgamated grading. In other words, although accounting for knowledge and psychosocial skills indicators, virtually

nobody employed differentiated grading procedures wherein students receive grades for separate skills (e.g., product, process, or progress indicators). It is not particularly surprising. Historically, because of concerns about student motivation, self-esteem, and the social consequences of grading, and also wishing to ensure compliance with classroom rules, most professors combined multiple criteria to assign final grades (Brookhart et al., 2016; Guskey, 1996). An additional complication represents variations from student to student, thus further confounding the meaning of the grade (Tippin, Lafrenier, & Page, 2012). A grade of *A*, for example, may mean that a student accumulated superior skills and knowledge (product), did not learn as well as expected but displayed exceptional effort (process), or made a substantial improvement (progress). Hence, without understanding the underlying criteria and assessment types, looking at a single letter or a number representing a score is not particularly informative. It appears that the first step has been successfully made, with instructors realizing the value of process indicators.

A positive next step could be to explicitly differentiate among criteria and report them separately on the syllabi and to the students. Unlike colleges and universities, the landscape of K-12 education is undergoing rapid changes as the ESSA (Every Student Succeeds Act) standards are implemented. These standards require student assessment to be broad and inclusive of additional indicators of "success", such as student personal growth and engagement, in addition to traditional achievement scores on tests (US Department of Education, 2016). Certainly, the notion of using differentiated grades to reflect different aspects of student performance seems feasibly integrated into this new initiative. Whether colleges and universities follow suit, however, is to be determined. Considering the vast majority of institutions that already incorporate process and product criteria into an amalgamated grade, switching to a differentiated system of grading, where students could receive separate grades to reflect different

aspects of achievement, does not at all seem to be an implausible next step. ESSA could very well be the catalyst that shifts grading standards from amalgamated to differentiated, in which product, process, and progress criteria would all be recognized as equally important, yet independent, grades in and of themselves. Although this is not a common practice in the United States or Spain, many institutions in Canada have already adopted this system (O'Connor, 2010). Ultimately, differentiated grades that assign independent marks to separate measures of student achievement, as opposed to grades that reflect combined multiple aspects of student performance, can give a more meaningful and accurate account of student performance in various areas (Guskey, 2011; Rojstaczer, & Healy, 2012; Royal & Guskey, 2015).

Peer and self-assessment implementation

Regarding involvement of students in assessment via peer and self-assessment, the results are considerably discouraging. The presence of peer and self-assessment could be considered residual with only a handful of syllabi including them. There were nine instances of peer and self-assessment reflected in the US syllabi, and fourteen in Spain. This is negative for at least three main reasons. First, peer and self-assessment contribute to students' performance and self-regulated learning (e.g. Brown & Harris, 2013; van Zundert et al., 2010). Second, it seems crucial that students develop these skills in higher education before they start their active professional careers where instructional guidance is less structured than at university (Brown & Harris, 2018). And third, some authors have claimed that peer and self-assessment could decrease teachers load (Lipnevich et al., 2014) though this can probably only be achieved after investing significant instructional time into training students how to carry it out properly (e.g. Panadero & Brown, 2017).

Importantly, if we compared our results to those previously reported (Panadero et al., 2019), the results are quite similar. This leads to the conclusion that university instructors are either not willing or prepared to fully incorporate peer and self-assessment in their instructional repertoire, at least as reported in their syllabi. This is, in our opinion, an opportunity lost for students' learning and skill development. A major effort should be invested into transmitting the message of the importance of peer and self-assessment and preparing university instructors on how to effectively carry it out.

Country and academic discipline as predictors

In terms of country and discipline predictions of probabilities of utilizing various assessment types, we found differences in the likelihood of allocating a portion of the course grade to final written exams in Spain, depending on the academic discipline. Education had the lowest likelihood and mathematics has the highest likelihood of its grade allocated to final written exams. The likelihood of final written exams for the US sample was relatively similar across academic disciplines. Similar patterns were also found for individual assignments and partial exams such that for Spain, the likelihood of allocating course grading to individual assignments and partial grades depended on the discipline. That is, English/Spanish and mathematics had the highest percentages allocated to individual assignments, with math, science, and psychology having the highest likelihood of allocating the largest proportion of grades to partial exams. Interestingly, for final written exams, US instructors utilized this assessment type more frequently when assessing student performance in psychology, English/Spanish, and science, compared to Spain. For both countries, written exams accounted for the lowest percentage of the final grade in education, and, overall, education was the leader in the usage of the process criteria. This is a somewhat expected finding, with education faculty being the most

prepared to design and implement a variety of assessments. Curricular and assessment design is an ingrained part of educational programs, so, inevitably, instructors in education may be better equipped in preparing greater diversified assessments and realizing the importance of varied criteria of assessments (Guskey, 2006). Finally, attendance weights were similar between the two countries, with language (English/Spanish) allocating the highest percentage compared to math, which was the lowest. Due to the lack of previous research exploring these specific comparisons it is not possible to establish connections to other literature.

Limitations and Future Directions

This study is not without limitations. Only first year college syllabi were analyzed between the two countries and across the five academic disciplines. Future studies may examine syllabi across the four years and beyond, and may include additional academic disciplines. Panadero et al. (2019) showed minimal differences in assessment types between the first and the fourth years of study, but investigating whether these patterns persist across domains and countries could be of interest to the scientific community. Further, disciplines that are focused on individual achievement, such as music, drama, and sports, may more frequently include progress-based criteria and reveal different patterns of assessment. Future studies may explore assessment criteria in a greater range of domains and across academic levels.

Another limitation may include the scope of grading practices, which was not variable in our study. We had to reduce our analyses down to the binary absent or present, and we could not in all instances investigate the full scope of the amount/percentage of each of the assessment practices. Perhaps larger, more diverse syllabi samples would give us such information. Future studies should include a larger sample of international syllabi to boost specificity levels of grading criteria.

In our analyses we did not account for type of course (e.g., large lecture class, science laboratory class, online class). These course type differences may place opportunities or limitations on assessment practices. This could be a promising avenue for future research. Similarly, we did not account or control for class size or instructor course load, all of which can have effects on choosing assessment types. These could be useful areas for investigation for future researchers.

Conclusion

Our study shed light on various aspects of assessment implementation in higher education. In both the US and Spain, first year students get evaluated via different types of assessments, both product and process, which are then reported in one final score. On the one hand, this can be considered positive as shows that we have long moved from an assessment approach that was based only on a final exam. On the other hand, the most traditional assessment type, exam, is still the most commonly used and heavily weighted assessment type (partial in the US, final in Spain). Although this may not be ideal due to exams' limitations, in combination with other assessment types, it may be an acceptable solution to the complex problem of teachers' load and available resources. The picture of peer and self-assessment is without a doubt quite negative with hardly any syllabi explicitly mentioning them. It is our hope that university instructors will attempt to employ this powerful instructional tool with greater frequency. Finally, previously reported differences in assessment among academic disciplines were revealed in the current investigation, showing that differences in content and assessment traditions endemic to specific disciplines contribute to the choice of assessments instructors use.

References

Allen, J. D. (2005). Grades as valid measures of academic achievement of classroom learning. *The Clearing House, 78*(5), 218-223.

Angelo, T. A., Cross, K. P., & Rechnique, C. A. (1993). *A handbook for college teachers*. San-Francisco: Jose-Bass Publishers.

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.

Bjork, E. L., Storm, B. C., & DeWinstanley, P. A. (2010). Learning from the Consequences of Retrieval: Another Test Effect/AS Benjamin (Ed.) *Successful Remembering and Successful Forgetting: A Festschrift in Honor of Robert A. Bjork.* (pp. 347-354).

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7-74.

Bloom, B. S. (1956). *Taxonomy of educational objectives*. New York: David McKay Co. Inc.

Brookhart, S. M. (1991). Grading practices and validity. *Educational Measurement: Issues and Practice*, *10*(1), 35-36.

Brookhart, S. M. (2009). *Grading*. Upper Saddle River, NJ: Pearson Education.

Brookhart, S. M. (2011). Starting the conversation about grading. *Educational Leadership*, *69*(3), 10-14.

Brookhart, S. M. (2014). *How to design questions and tasks to assess student thinking*. Virginia, USA: ASCD.

Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T., & Welsh, M. E. (2016). A century of grading research: Meaning and value in the most common educational measure. *Review of Educational Research, 86*(4).

Brookhart, S. M., & Nitko, A. J. (2008). *Assessment and grading in classrooms*. Upper Saddle River, NJ: Pearson Education.

Brown, G. T., & Harris, L. R. (2014). The Future of Self-Assessment in Classroom Practice: Reframing Self-Assessment as a Core Competency. *Frontline Learning Research, 2(1)*, 22-30.

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.

Casner-Lotto, J., Barrington, L., & Wright, M. (2006). Are they really ready to work? Employers' perspectives on the basic knowledge and applied skills of new entrants to the twenty-first century U.S. workforce. Report Number BED-06-WF-KF. New York: The Conference Board, Corporate Voices for Working Families, the Partnership for Twenty-first Century Skills, and the Society for Human Resource Management.

Cassady, J. C., & Johnson, R. E. (2002). Cognitive test anxiety an academic performance. *Contemporary Educational Psychology, 27*(2), 270-295.

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' assessment practices: Preparation, isolation, and the kitchen sink. *Educational Assessment*, *3*(2), 159-179.

Cross, L. H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education*, *12*(1), 53-72.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14(1)*, 4-58.

European Commission/EACEA/Eurydice (2018). *The European Higher Education Area in 2018: Bologna Process Implementation Report*. Luxembourg: Publications Office of the European Union.

Guskey, T. R. (1996). Reporting on student learning: Lessons from the past – Prescriptions for the future. In T. R. Guskey (Ed.), *Communicating Student Learning. 1996 Yearbook of the Association for Supervision and Curriculum Development* (pp. 13-24). Alexandria, VA: Association for Supervision and Curriculum Development.

Guskey, T. R. (2006). Making high school grades meaningful. *Phi Delta Kappan*, *87*(9), 670-675.

Guskey, T. R. (2011). Five obstacles to grading reform. *Educational Leadership, 69*(3), 16-21.

Guskey, T. R. (2019). *Get set, go! Implementing successful reforms in grading and reporting*. Bloomington, IL: Solution Tree.

Hattie, J. (2012). *Visible learning for teachers: Maximizing impact on learning*. Routledge.

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J. A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the

programme for international student assessment. *Scandinavian Journal of Educational Research, 62(3)*, 333-353.

Tomas, C., & Jessop, T. (2019). Struggling and juggling: a comparison of student assessment loads across research and teaching-intensive universities. *Assessment & Evaluation in Higher Education, 44(1)*, 1-10.

Jessop, T., & Maleckar, B. (2016). The influence of disciplinary assessment patterns on student learning: A comparative study. *Studies in Higher Education, 41(4)*, 696-711.

Jessop, T., & Tomas, C. (2017). The implications of programme assessment patterns for student learning. *Assessment & Evaluation in Higher Education, 42(6),* 990-999.

Karran, T. (2005) Pan-European Grading Scales: Lessons from National Systems and the ECTS. *Higher Education in Europe, 30*(1), 5-22. doi: 10.1080/03797720500087949

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(4)*, 989.

Lipnevich, A. A., McCallen, L. N., Miles, K. P., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science,* 42(4), 539-559.

Lipnevich, A. A., Guskey, T., Murano, D., & Smith, J. K. (2020). Variation in Grading Criteria in College and University Courses. *Assessment and Evaluation: Principles, Policy, and Practice.*

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied, 15*(4), 319-333.

O'Connor, K. (2009). Grades: When, why, what impact, and how? *Education Canada, Spring 2010,* 38-41.

Oswald, F. L., Schmitt, N., Kim, B. H., Ramsay, L. J., & Gillespie, M. A. (2004). Developing a biodata measure and situational judgment inventory as predictors of college student performance. *Journal of Applied Psychology, 89(2),* 187-207.

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

Panadero, E., Fraile, J., Fernández Ruiz, J., Castilla-Estévez, D., & Ruiz, M. A. (2019). Spanish university assessment practices: examination tradition with diversity by faculty. *Assessment & Evaluation In Higher Education, 44*(3), 379-397. doi:10.1080/02602938.2018.1512553

Parkes, K. A. (2018). Instructional Feedback in Music. In A. A. Lipnevich and J. K. Smith (Eds.) *Cambridge Handbook of Instructional Feedback*. Cambridge Press.

Pellegrino, J. W. (2006). Rethinking and redesigning curriculum, instruction and assessment: What contemporary research and theory suggests. *Commission on the Skills of the American Workforce, Chicago,* 1-15.

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2007). The effect of pedagogical training on teaching in higher education. *Teaching and teacher education, 23(5)*, 557-571.

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The pretesting effect: Do unsuccessful retrieval attempts enhance learning? *Journal of Experimental Psychology: Applied, 15(3)*, 243.

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.

Rojstaczer, S., & Healy, C. (2012). Where A is ordinary: The evolution of American college and university grading, 1940-2009. *Teachers College Record, 114*(7), 1-23.

Royal, K. D., & Guskey, T. R. (2015). A case for differentiated grades. *Medical Science Education, 25,* 323-325.

Seipp, B. (1991). Anxiety and academic performance: A meta-analysis of findings. *Anxiety, Stress, and Coping, 4*(1), 27-41.

Smith, J. K., & Smith, L. F. (2009). The impact of framing effect on student preferences for university grading systems. *Studies in Educational Evaluation, 35,* 160-167.

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.

Tippin, G. K., Lafrenier, K. D., & Page, S. (2012). Student perception of academic grading: Personality, academic orientation, and effort. *Active Learning in Higher Education*, *13*(1), 51-61.

Turner, K., Roberts, L., Heal, C., & Wright, L. (2013). Oral presentation as a form of summative assessment in a master's level PGCE module: The student perspective. *Assessment & Evaluation in Higher Education, 38*(6), 662-673.

United States Department of Education (2016). *Transitioning to the Every Student Succeeds Act (ESSA): Frequently asked questions.*

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.

Wächter, B. (Ed.). (2004). *Higher education in a Changing Environment*. Bonn, Germany: Lemmens.

Webb, C., Endacott, R., Gray, M. A., Jasper, M. A., McMullan, M., & Scholes, J. (2003). Evaluating portfolio assessment systems: What are the appropriate criteria? *Nurse Education Today, 23,* 600-609.

Westerheijden, E. et al. (2008). The first decade of working on the European Higher Education Area [Report by the Center for Higher Education Policy Studies].

Wiggins, G. (1996). Honesty and fairness: Toward better grading and reporting. In T. R. Guskey (Ed.), *Communicating student learning: 1996 Yearbook of the Association for Supervision and Curriculum Development* (pp. 141-176). Alexandria, VA: Association for Supervision and Curriculum Development.

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