How do Undergraduates Perceive the Use of Assessment? A Study in Higher Education

Diana Pereira[1], Irene Cadime[1], Gavin Brown[2], Maria Assunção Flores[1]

[1] *Universidade do Minho, Braga, Portugal*

[2] *The University of Auckland, Auckland, New Zealand*

Abstract

Drawing upon a wider piece of research, this paper focuses on the validation of a "use of assessment" scale five Portuguese public universities with 5549 students. The study aims to investigate the psychometric properties of the scale; to describe how students look at assessment uses; to analyse their utility perceptions of assessment; and to understand their perspectives regarding participation in the assessment process. The scale demonstrated adequate validity and reliability based on factoral analysis of internal structure. Preference for alternative methods of assessment was negatively correlated with preference for traditional methods of assessment. However, preference for alternative methods was correlated with higher scores for assessment effectiveness, fairness and level of participation and engagement with assessment. On the contrary, preference for traditional methods was negatively correlated with perceived fairness and with engagement with assessment.

Correspondence should be addressed to Dr. Diana Alexandra Ribeiro Pereira, Institute of Education, Universidade do Minho, Braga, Portugal or by email to dianapereira@ie.uminho.pt.

Introduction

Assessment in higher education has been the focus of attention, particularly after the implementation of the Bologna Process. The literature in this field suggests a paradigm shift that points, amongst other aspects, to a more student-and-learning-centered approach (Black and Wiliam 1998; Webber 2012; Myers and Myers 2015), emphasizing the student's active role in the learning and assessment process (Struyven, Dochy, and Janssens 2003). A recent review of assessment suggests a greater emphasis on alternative assessment methods as well as a greater diversity of assessment practices (Pereira et al. 2016). However, the assumptions inherent to an assessment approach that promotes students' autonomy, different learning approaches, motivation to learn and critical thinking do not fit well into the prevalence of the so-called traditional assessment methods such as written tests or exams (Goubeaud and Yan 2004; Wen and Tsai 2006; Duncan and Buskirk-Cohen 2011; Price et al. 2011).

Although research points to an approach based on the principles of formative assessment, because this improves the teaching and learning process (Biggs 2003; Boud and Falchikov 2007; Brown and Knight 1994; Brown, Bull, and Pendlebury 1997; Gulikers, Kester, Kirschner, and Bastiaens 2008), assessment in higher education entails essentially a summative dimension with an emphasis on the certification and grading of the students' learning. Extant literature identifies several processes and strategies used by educational institutions to improve the learning process through changes in the assessment process (such as the use of diversified instruments and techniques that promote the transparency and authenticity of the assessment process, the use of feedback, etc.) and innovative assessment practices (such as portfolios, self-assessment, peer-assessment, simulations, etc.) (Struyven and Devesa 2016; Struyven, Dochy, and Janssens 2005). However, more needs to be done to understand the higher education assessment processes from the student perspective. The goal of this paper is to examine

students' perceptions about the use of different assessment methods as well as their participation in the assessment process.

Use of Assessment in Higher Education

In recent years, higher education assessment studies have focused on students' participation, innovative practices and impact of different assessment methods, functions and moments as well as the use of feedback (Pereira et al. 2016; Jessop, El Hakim, and Gibbs 2014; Yan and Brown 2017; Huisman, Saab, van Driel, and den Broek 2018; Harland, Mclean, Wass, Miller, and Sim 2015; Wilson, Diao, and Huang 2015; Wilson, Ho, and Brookes 2018). Earlier empirical work suggests that students found assessment more effective and fairer when learner-centered assessment methods or mixed methods are used (Flores et al. 2015; Pereira et al. 2017a; Pereira et al. 2017b) because they may correspond to a formative assessment approach emphasizing, amongst other features, critical thinking and feedback (Light and Cox 2003; Falchikov 2005; Webber 2012).

Self- and peer-assessment are described in the literature as promoters of improved learning and academic performance (Brew, Riley, and Walta 2009; Lew, Alwis, and Schmidt 2010) and self-regulation of learning (Mok, Lung, Cheng, Cheung, and Ng 2006; Yan and Brown 2017). Amongst other features, these modes of assessment involve students in the assessment process, enhancing their autonomy, engagement, and self-confidence (Dochy et al. 1999; Fitzpatrick 2006). Involvement in assessment also enhances critical thinking enabling students to recognize their successes and failures and those of their colleagues (Pereira et al. 2016). For instance, the study by Sadeghi and Khonbi (2015) concluded that students show positive attitudes towards their experiences of self-assessment and peer-assessment.

Self-assessment has been recommended "as an appropriate approach to student involvement in formative assessment" (Panadero, Brown, and Strijbos 2016, p. 804) as it engages learners in the learning process, promotes feedback and develops students' ability to learn and to self-assess (Orsmond and Merry 2013). Yan and Brown (2017, p. 1284) defined self-assessment as "a process during which students collect information about their own performance, evaluate and reflect on the quality of their learning process and outcomes according to selected criteria to identify their own strengths and weaknesses". Recent studies highlight the importance of self-assessment because when students self-assess, they are able to recognize the strengths and drawbacks of their learning (Micán and Medina 2017). This practice allows students to use strategies that enable them to develop greater awareness of desirable and effective future actions. However, a review of self-assessment showed that definitions lack consistency and clarity based on different understandings and leading to different practices (Panadero, Brown, and Strijbos 2016).

Along with self-assessment students' perceptions of peer assessment are overall positive (Ashenafi 2017). A recent literature review of peer-assessment describes it as: "a scenario in which two or more students are involved in completing tasks that require fairly equivalent levels of participation for the entire process to be effective" (Ashenafi 2017, p. 245). Peer-assessment is appreciated by students for many reasons; it is useful and effective because it allows interaction among students (van den Berg et al. 2006; Merry and Orsmond 2018), it enables student engagement producing formative feedback (Patton 2012; Vickerman 2009), and promotes deep thinking and reflection (Segers and Dochy 2001), as well as the development of transferrable skills (McGarr and Clifford 2013). However, despite these benefits, studies suggested that university students have expressed frustration regarding peer assessment activities (Wilson, Diao, and Huang, 2015) and may

perceive peer-assessment as an unfair process (Carvalho 2013; McConlogue 2012). Some of the difficulty in peer assessment arises from the complex inter- and intra-personal demands it places on students (Panadero, 2016). However, just because students consider peer assessment as unfair, it does not mean that it is an unfair method of assessment. It is likely that students' perceptions of collaborative assessment as unfair is exacerbated where peer assessment activity is not adequately explained or moderated by tutors or course coordinators.

Indeed, Wilson, Ho and Brookes (2018) suggested that, although students understand the importance of developing working skills in teams for their professional future, most students did not feel well-enough prepared to develop these skills in courses; instead they may have valued more teamwork during laboratory sessions, team sports, and informal study groups than in formal assessment of teamwork. Matos, Cirino, and Brown (2009) also indicated that two dimensions (i.e. the formality of the assessment method and the locus of control) summarize the variation in student thinking. Assessment was seen as a set of formal practices similar to tests, under the control of teachers, not students; even self- and peer-assessment practices are controlled by teachers and not by students themselves. Their survey with Brazilian university students showed that formal methods of assessment were seen as *not* irrelevant but were associated with negative personal emotions, and that teacher-controlled formal assessments had only a small positive association with improved learning (Matos, Cirino, Brown, and Leite 2013).

Another empirical study carried out in New Zealand also indicates that, although students and teachers value certain aspects of the learning process, they may be difficult to achieve, namely research projects and essays designed to integrate knowledge, skills and understanding through autonomous learning (Harland, Mclean, Wass, Miller, and Sim 2015). However, Cooper (2017) concluded that collaborative summative assessment,

as a participatory process that engages the key stakeholders, such as academia and professional practice, provides students with opportunities to experience "performance of understanding" in a community of practice that reflects the "real" working life.

Recent studies also showed the predominance of summative assessment through the use of traditional methods, particularly tests and exams, and the grading of students without feedback from teachers (Barreira, Bidarra, Monteiro, Vaz Rebelo, and Alferes 2017). In a similar vein, Panadero, Fraile, Fernández, Castilla-Estévez, and Ruiz (2018) found that despite the large variation of assessment practices used in different universities and fields of knowledge in Spanish universities, there was frequent use of traditional assessment practices. Likewise, while formal examinations were the predominant basis of grading in American universities, there were disciplinary differences (i.e., education and English had fewer exams than mathematics, science, or psychology) (Lipnevich et al. 2020)

The challenge in this literature is that assessment methods are not inherently formative or summative; hence, understanding how students perceive the various uses, rather than the kinds, of assessment is needed. This paper presents data from a wider research project on assessment and reports on the validation of a scale drawing upon data from a large sample of higher education students.

The Study

This study is part of a larger research project entitled "Assessment in higher education: the potential of alternative methods" conducted in five Portuguese public universities (Flores et al. 2020). This paper aims to: (1) investigate the psychometric properties of the scale “use of assessment” by expanding the previous work of Pereira (2011, 2016) developed for the Portuguese context; (2) analyze the relationships between the uses of different types of assessment and the perceptions of effectiveness, fairness, and

participation in the assessment process; and (3) explore if these perceptions (i.e., use, utility, and participation) vary as a function of academic and students' personal variables.

Participants

In total, 5549 students from five Portuguese public universities participated (Table 1). The population (i.e. the number of students attending these universities) was approximately 117150 students. Therefore, a sample of 5549 has a margin of error rate as low as 1.28%. The vast majority (*n*=5407) completed the questionnaire in their classroom and 142 completed the questionnaire online using a link provided via e-mail. Approximately two-thirds of the participants were female students (n=3498, 64.4%). Just over half of the participants were between 20 and 25 years old (55.7%). Participants were enrolled in different academic programs, including Medical and Health Sciences, Exact Sciences, Engineering and Technology, Social Sciences, and Humanities. About four-fifths of the participants were in the first three years of an undergraduate or integrated masters' degree.

Table 1. Demographic characteristics of the participants

Demographic characteristics	***n***	**%**
University		
A	2039	36,8
B	620	11,2
C	1355	24,4
D	1123	20,2
E	412	7,4
Gender		
Male	1931	35,6
Female	3498	64,4
No information	120	
Age		
Less than 20	2079	37,6
20-25	3081	55,7
More than 25	375	6,8
No information	14	
Field of knowledge		

Demographic characteristics	***n***	**%**
Medical and Health Sciences	1362	24,4
Exact Sciences	211	3,8
Engineering and Technology	1519	27,2
Social Sciences	2015	36,3
Humanities	442	8,0
Cycle of study		
Undergraduate	2824	50,9
Master degree	162	2,9
Integrated master degree	2563	46,2
Year of study		
First year	1780	32,1
Second year	1531	27,6
Third year	1432	25,8
Fourth year	470	8,5
Fifth year	240	4,3
Sixth year	93	1,7
No information	3	0

Procedures for data collection

In accordance with the procedures approved by the Ethics Committee of the University of xxx the Presidents of Faculties/Institutes/Schools were asked for permission to carry out the study. Then the Program Directors/Coordinators were asked to provide the contacts of the faculty to schedule the time and place in which their students could be invited to complete the questionnaires. The research protocol was sent to all program Directors/Coordinators. An informed consent form was also designed. All of the participants were asked to sign the informed consent after a brief explanation of the context and goals of the research project. Data were collected in person and via email from February to July 2017.

The instrument

The construction of the scale was based upon previous work (Pereira, 2011, 2016) focusing on the use of assessment. The scale “use of assessment” included 33 items related to (1) use of alternative methods (e.g. Assessment based on portfolios, projects or reflections enables the development of new learning); (2) use of traditional methods (e.g.

I spend more hours in my study when assessment is performed through tests or exams); (3) effectiveness of assessment (e.g. Assessment is more effective when it encourages me to apply knowledge in real contexts/situations); (4) fairness of participation in assessment (e.g. Assessment is fairer when there is either self-assessment or peer assessment.); (5) current level of participation in assessment (e.g. In the first semester of 2016/2017 I was asked to do self-assessment); (6) engagement with assessment (e.g. When I prepare myself for an exam I only start to study shortly before the test and not during the semester); and (7) current level of continuous assessment (e.g. In the first semester of 2016/2017, assessment was made throughout the semester). The different items were derived from studies by Struyven, Dochy, and Janssens (2005), Falchikov (2005), Webber (2012), and Flores et al (2015). A five-point Likert scale was used ranging from 1 = totally disagree to 5 = totally agree in order to identify students' perceptions about the use of assessment. The reason to not use a scale already in the literature was because the existing instruments did not cover all the dimensions under investigation. Thus, the elaboration of this instrument allowed investigation of issues related to the nature of assessment methods, student engagement and participation in the assessment process, as well as perceptions of effectiveness of assessment.

Data analysis

The psychometric properties of the measure were studied, by analyzing its internal structure and the reliability of the scores. First, missing values were analyzed. Participants with more than 10% of missing values (i.e. more than three missing values in the items) were deleted (n=120; 2.16%). From the remaining 5429 cases, 773 (14.2%) had at least one missing value, but the missing values represented only 0.55% of the data set. The chi-square ratio to the degrees of freedom of the Little's MCAR test was not statistically significant ($\chi^2_{(5823)=}$6916.782; χ^2/df=1.19, p>.05), suggesting that these missing data were

random. No variable had more than 10% of missing data. Research suggests that maximum likelihood estimates should be preferred over multiple imputation when there are violations to the distribution of the observed variables (Yuan, Yang-Wallentin, and Bentler 2012). Therefore, 5429 cases were considered in the subsequent analyses with full information maximum likelihood (FIML) estimation.

Univariate and multivariate normality of the items were inspected using MVN package for R (Korkmaz, Goksuluk, and Zararsiz 2014). For univariate normality, the skewness and kurtosis values for each item ranged between -3 and +3 (i.e., skewness range -1.095 to 1.211; kurtosis range -1.121 to 2.292), thus suggesting no violation to univariate normality (Kim 2013). Multivariate normality was assessed by computing Mardia's multivariate kurtosis (MK) statistics. A significant test was obtained (MK=1388; z-statistic=165.4, p<.001). However, as the sample size is large, significance testing is not very informative. Therefore, the expected multivariate kurtosis was calculated and compared to the observed value. The expected MK is $p(p + 2)$, where p is the number of observed variables (Cain, Zhang, and Yuan 2017). Therefore, for this analysis, the expected MK was 1155 is slightly lower than the observed value of 1388. Therefore, these results suggest a moderate deviation from multivariate normality. To deal with this deviation, confirmatory factor analysis (CFA) was performed using the maximum likelihood estimation with robust standard errors (MLR). MLR estimator provides sandwich-type standard errors and rescaled test statistics (Muthén and Muthén 2012), which provides consistent estimation, even when there are moderate violations to normality (Cain, Zhang, and Yuan 2017; Yang and Liang 2013). This estimator also has the advantage of accounting for missing data.

Analyses were conducted with Mplus version 7 (Muthén and Muthén 2012). To evaluate model fit, the following criteria were used: a Comparative Fit Index (CFI) above

.90 (Byrne 2011; Hu and Bentler 1999), a Gamma Hat higher than .90 (Marsh, Hau, and Wen 2004), a Root Mean Square Error of Approximation (RMSEA) and a Standardized Root Mean Square Residual (SRMR) below .08 (Browne and Cudeck 1993; Hu and Bentler 1999; Schermelleh-Engel, Moosbrugger, and Müller 2003). Composite reliability was tested by computing the McDonald's omega (ω). Values higher than .70 are conventionally considered adequate (Hair, Black, Babin, and Anderson 2009). However, lower values are commonplace when few items are present within a factor or when items are heterogenous in content (Cattell and Tsujioka 1964)

In a second step, multi-group CFA was performed to test the invariance of the factor structure across gender, age groups, and cycle of study. Configural, metric and scalar invariance were tested in successive models. To examine the differences in model fit, the difference in CFI (ΔCFI) and the difference in RMSEA (ΔRMSEA) were calculated. Values of ΔCFI $\leq$.01 and values of ΔRMSEA $\leq$.015 indicate that the hypothesis of invariance should not be rejected (Chen 2007; Cheung and Rensvold 2002). The Bayesian Information Criterion (BIC) was also considered: lower BIC values indicate better fit. When full invariance was not achieved, partial invariance was established by freely estimating the parameters identified after examining the Lagrange Multiplier tests.

In a third step, multivariate analysis of variance (MANOVA) was used in order to test the effects of student's personal and academic variables on the factor scores. Partial eta squared (η_p^2) was used as measure of effect size, indicating the percentage of variance explained by each independent variable. The following guidelines were used for interpretation of effect size: $\eta_p^2>.5$ indicates a large effect; $.3<\eta_p^2<.5$, a medium effect, $.1<\eta_p^2<.3$, a small effect; values <.1, a negligible effect (Cohen 1988). Due to the large

number of comparisons, Bonferroni post-hoc pairwise comparisons were performed in case of significant multivariate and univariate effects.

Results

Model fit

The fit of the initial model was not adequate ($\chi^2_{(474)}$=10074.949, p<.001, RMSEA=.061 [90% CI .060, .062], CFI=.802, Gamma Hat=.903, SRMR=.063). In order to improve model fit, two procedures were applied: (a) items with factor loadings <.30 were removed; (b) items that were highly correlated and shared content with some other (i.e., were redundant) were removed. In the latter case, the item with the higher contribution to the factor was preferred and kept in the model. This led to the elimination of six items (items 13, 21 and 33 from factor 1, items 4 and 15 from factor 3 and item 16 from factor 5). The final model presented a good fit ($\chi^2_{(303)}$=3714.195, p<.001, RMSEA=.046 [90% CI .044, .047], CFI=.909, Gamma Hat=.955, SRMR=.050). Table 2 presents the item estimates in this final model which were higher than .30.

Table 2. Factor loadings

Items	Std. loading
Factor 1 – Preference for alternative methods	
12. Assessment based on portfolios, projects or reflections enables the development of new learning.	.664
24. Portfolios, projects or reflections enable a more effective assessment of learning.	.808
17. Portfolios, projects or reflections enable a fairer assessment of learning.	.772
10. I dedicate more hours to study when assessment is done through portfolios, projects or reflections.	.621
26. When I do a project or portfolio I study throughout the semester.	.564
22. I feel more confident when I am assessed through assessment methods in which I actively participate in the tasks.	.426
Factor 2 – Preference for traditional methods	
11. Written tests or exams enable a more effective assessment of learning.	.819
14. Written tests or exams enable a fairer assessment of learning.	.819
6. Assessment is fairer when it includes written tests or exams.	.753

Items	Std. loading
20. I feel more confident when I am assessed through tests or exams.	.617
19. I spend more hours in my study when assessment is performed through tests or exams.	.364
Factor 3 – Effectiveness of assessment	
2. Assessment is more effective when it enables me to improve my technical or scientific skills (related to my area of expertise).	.893
3. Assessment is more effective when it enables me to simultaneously improve my technical and soft skills (search and selection of information. teamwork. etc.).	.754
1. Assessment is more effective when it encourages me to apply knowledge in real contexts/situations.	.755
5. Assessment is more effective when it contributes to deep learning.	.571
Factor 4 – Fairness of participation in assessment	
9. Assessment is fairer when there is either self-assessment or peer assessment.	.935
8. Assessment is fairer when it includes peer assessment.	.795
7. Assessment is fairer when I also do self-assessment.	.574
Factor 5 – Current level of participation in assessment	
27. In the first semester of 2016/2017 I was asked to do self-assessment.	.735
28. In the first semester of 2016/2017, I participated in the assessment of my colleagues.	.808
32. Overall, the assessment methodology of the curricular units of the first semester of 2016/2017 was discussed and negotiated with students.	.475
Factor 6 – Engagement with assessment	
18. I usually forget most of the topics I studied after doing the written test.	.615
23. Usually, I only study the topics that will be the focus of the written tests.	.533
25. When I prepare myself for an exam I only start to study shortly before the test and not during the semester.	.513
Factor 7 – Current level of continuous assessment	
29. In the first semester of 2016/2017, assessment was made throughout the semester.	.950
30. In the first semester of 2016/2017, assessment only took place at the end of the semester. (Reverse coded)	.501
31. In the first semester of 2016/2017, assessment occurred every time I completed a task or activity.	.357

Table 3 presents the correlations between factors and the composite reliability values. The size of the factor inter-correlations was relatively small with no value >|.35| (absolute mean *r*=.13), indicating that these factors were substantially independent from each other. As expected, the use of alternative methods of assessment was negatively correlated with the use of traditional methods of assessment. The use of alternative methods was otherwise positively but weakly correlated with all other factors except

continuous. On the contrary, the use of traditional methods was weakly negatively correlated with the perceived fairness of participation and with engagement with assessment. While the absolute values of these correlations is consistently small, the difference between values does point to some interesting patterns. Because the Pearson correlation is a standardized value, the difference in values between the Alternative and Traditional methods can be used to estimate the size of effect in how these two different assessment methods relate to the other factors. Values between .10 and .30 are small so this applies to effectiveness, participation, and continuous. The difference in the correlation to fairness (.35) and to engagement (.44) is moderate in favor of the alternative, suggesting that the alternative assessments relate more strongly to those two features.

Regarding composite reliability, adequate values were obtained for factors 1 to 5, but was relatively low for factors 6 and 7. However, the content of the items in these two factors is quite heterogeneous and thus composite reliability is not a good measure of fit. Therefore, we maintained these factors in the analyses.

Table 3. Correlations between factors and composite reliability

Factor	1	2	3	4	5	6	7
1. Alternative Methods	(.813)	-.349***	.192***	.270***	.109***	.221***	-.015
2. Traditional Methods		(.815)	.034*	-.084***	.060**	-.222***	.179***
3. Effectiveness			(.836)	.102***	-.140***	.063**	.034*
4. Fairness				(.820)	.195***	.040	.006
5. Participation					(.720)	-.131***	.166***
6. Engagement						(.571)	-.174***
7. Continuous							(.655)

Note: Values on diagonal in brackets = omega estimate of scale reliability; *p<.05; **p<.01; ***p<.001

Invariance analysis

Table 4 presents the results of invariance analysis of the seven-factor structure. Strong invariance was achieved for age groups. However, the full scalar models for gender, the

cycle of studies and the field of knowledge failed to demonstrate equivalence. For gender, partial scalar invariance was achieved by freely estimating the intercept of item 25 in Factor 6; whereas for cycle of studies, partial scalar invariance was achieved by freely estimating the intercepts of item 23 in Factor 6; and items 27 and 32 in Factor 5. Regarding the field of knowledge, partial scalar invariance was achieved by freely estimating the intercepts of items 10, 12, 24 and 26 in Factor 1, item 19 in Factor 2, items 1 and 2 in Factor 3, item 7 in Factor 4, items 27 and 28 in Factor 5, items 23 and 25 in Factor 6 and item 31 in Factor 7.

Table 4. Results of invariance analysis

Model	χ^2_{SB}	*df*	*p*	CFI	ΔCFI	RMSEA [90% CI]	SRMR	BIC
Gender								
Configural	3916.388	606	<.001	.910	-	.045 [.044, .047]	.051	355089.121
Metric	3985.454	633	<.001	.909	.001	.045 [.043, .046]	.054	354952.077
Scalar	4386.945	660	<.001	.899	.010	.046 [.045, .047]	.057	355159.237
Partial scalar	4267.019	659	<.001	.902	.007	.045 [.044, .047]	.057	355029.793
Age group								
Configural	4417.085	909	<.001	.907	-	.046 [.045, .048]	.053	363317.927
Metric	4553.549	963	<.001	.904	.003	.045 [.044, .047]	.059	363037.524
Scalar	4929.703	1017	<.001	.896	.008	.046 [.045, .047]	.062	362968.801
Cycle of study								
Configural	4600.417	909	<.001	.905	-	.047 [.046, .049]	.052	363824.045
Metric	4667.684	963	<.001	.904	.001	.046 [.045, .047]	.059	363477.581
Scalar	5297.830	1017	<.001	.889	.015	.048 [.047, .050]	.069	363694.082
Partial scalar	5085.348	1011	<.001	.895	.009	.047 [.046, .048]	.067	363509.729
Field of knowledge								
Configural	5055.635	1515	<.001	.907	-	.046 [.045, .048]	.054	364061.660
Metric	5383.291	1623	<.001	.901	.006	.046 [.045, .048]	.069	363562.487
Scalar	6517.220	1731	<.001	.875	.026	.050 [.049, .052]	.081	363881.247
Partial scalar	5794.289	1679	<.001	.892	.009	.048 [.046, .049]	.075	363526.532

Note: SB=Satorra-Bentler; CFI = Comparative Fit Index; RMSEA = Root Mean Square Error of Approximation; SRMR = Standardized Root Mean Square Residual; BIC = Bayesian Information Criterion.

Differences as a function of gender, cycle of studies, age group, field of knowledge and year of study

Given that strong invariance was not achieved for gender, cycle of studies and field of knowledge, MANOVA was conducted using only age and year of study as independent variables and the factor scores of the seven factors as dependent variables. Given that students younger than 20 years attending year of study 4 or higher were almost non-existent (n<10), the interaction effect age group×year of study was not included in the analysis. The multivariate results of MANOVA indicated a statistically significant main effect of both age group and year of study (Table 5). However, the effect sizes were negligible for both variables, indicating statistical significance arose because of large sample size only.

Table 5. MANOVA multivariate results

Effect	Wilks' Lambda	*F*	*df*	*p*	η^2_p
Age group	.978	8.539	14	<.001	.011
Year of study	.956	6.914	35	<.001	.009

Discussion and conclusions

This paper had three major goals. First, it aimed to investigate the psychometric properties of the expanded, multi-factorial inventory "use of assessment" within the Portuguese context. Second, it analyzed the relationships between the uses of different types of assessment and the perceptions students had of the effectiveness, fairness, and their participation in each assessment process. Thirdly, it explored if those perceptions varied as a function of academic and students' personal variables.

In respect of the first goal of this study, the results suggest that the seven-factor model had adequate psychometric properties and therefore can be considered as a useful tool to research the use of assessment in higher education. The seven factors, as per the original design, were (1) preference for alternative methods; (2) preference for traditional methods; (3) effectiveness of assessment; (4) fairness of participation in assessment; (5)

current level of participation in assessment; (6) engagement with assessment; and (7) current level of continuous assessment. These factors had relatively small inter-correlations and logically inverse relations were observed between preference for alternative and traditional methods. Within scale estimates of reliability were sufficient, when combined with the CFA fit indices to treat the 'use of assessment' inventory as a valid and reliable measure. Together, these factors provide students a mechanism to evaluate the dominant forms of assessment in higher education and provide a sense of their utility and propriety perspectives of such assessment methods. It is hoped that this will serve as a basis for future adaptations to other contexts (i.e., systems, languages) in order to further understand students' use of assessment.

Regarding the second goal, it was noteworthy that preference for alternative methods of assessment was negatively correlated with preference for traditional methods of assessment. These two classes of assessment tend to have quite different relationships to learning. Alternative methods include portfolios, project-based assessments, and other "activities [which] include multiple drafts of written work in which faculty provide constructive and progressive feedback, oral presentations by students, student evaluations of each other's work" (Webber 2012, p. 203). Thus, alternative assessments take place during course instruction and offer opportunities for feedback and revision. In contrast, traditional methods are used summatively, with no opportunity for resit or even feedback beyond a score (Brown, Bull, and Pendelburry 1997).

However, traditional methods are not suitable for all purposes of assessment and can encourage reproduction and memorization (Perrenoud 1999; Biggs 2003). Thus, there is a need to use a variety of assessment methods based on their suitability for teaching and learning objectives as well as the nature of courses and subjects. Non-traditional methods of assessment have emerged to address skills and knowledge that are less valid

within the context of a traditional assessment. Further, because of the potential for richer feedback during a course of instruction, alternative assessments provide greater alignment with the objectives of formative assessment (MacLellan 2004). However, preference by students for alternative methods needs to be tempered with their experience of how such assessments are graded. Because alternative methods of assessment involve more teacher professional judgment, there is the possibility of quite unreliable scoring, as is seen already in the inconsistent awarding of grades or scores for written essays (Brown, 2009). If students experience alternative methods as being scored in an arbitrary or ad hoc manner, it is likely they will have greater preference for traditional supposedly 'objective' testing.

In relation to the students' perceptions of effectiveness of assessment this study found that student perceptions on the effectiveness of assessment are positively correlated with both the use of alternative and traditional methods of assessment. It seems that students see the opportunity to develop both technical and soft skills through a combination of both traditional and alternative methods. However, preference for alternative methods was correlated with higher scores for perceptions of effectiveness, fairness, and levels of participation and engagement with assessment. This latter association is only logical because alternative assessments inherently require much greater student participation and engagement. Consistent with their inverse relations, preference for traditional methods was negatively correlated with the perceived fairness and levels of participation and engagement with assessment. Again, this is somewhat definitional because teachers, rather than students, create, administer, and score traditional assessments. Because examinations are normally final and carry significant weight for grades they tend to engender considerable anxiety. The presence of anxiety around traditional assessments, however, is not necessarily a bad thing because it can

activate learning (Vogl & Pekrun, 2016) and has been empirically shown to associate with greater outcomes (Brown, Peterson, and Irving, 2009). It is somewhat surprising that the students perceived traditional assessments as not fair, in contrast to alternative methods of assessment, which had a positive correlation with fairness. This is somewhat surprising because it is a well-established principle of high-stakes assessment that standardization of tasks and administration leads to greater equity and fairness. Elsewhere, Hong Kong university students indicated traditional assessments were unfair because they were inaccurate and mis-measured students' real capabilities (Brown and Wang, 2013; Wang and Brown, 2014). It may be that students are less concerned about the fairness of scoring and administration than assessors. Perhaps students consider fairness as being about being able to show their own abilities (Pereira et al. 2017a). Alternative methods enhance the active and autonomous role of the student and his/her involvement in the assessment process (Sambell and McDowell 1998; Sluijsmans, Dochy, and Moerkerke 1999). Thus, when students perceive that the evaluative process allows personalization and involvement, they see assessment as fair as opposed to being subjected to the same tasks as everyone else. Additionally, the opportunity to apply knowledge and competencies in real contexts as well as the development of both technical and soft skills, which is inherent in alternative assessments, contribute to positive student perceptions of assessment.

In terms of academic context and student demographic characteristics, very small differences were found. In part, this might be attributable to the naturalistic data collection processes which meant that cell sizes for comparison purposes were frequently too small for safe analysis. However, the lack of invariance in the factorial model does limit our ability to comment meaningfully about differences arising from disciplines. Coherent, with expectations there were statistically significant effects for age group and year of

study, but these were negligible in size. Hence, it may be that student perceptions of assessment are shaped much more by the commonalities of the competitive university entrance system that relies heavily on traditional assessments and the predominance of traditional assessment methods across all institutions of higher learning. Nonetheless, this study suggests researchers do not need to worry themselves with the background variables evaluated in this study in future research; they simply do not matter to student perceptions of the uses of assessment.

Earlier studies suggest that skills' development and deep approaches to learning are promoted through the use of alternative assessment methods (Sambell and McDowell 1998; Dochy et al. 1999; Segers and Dochy 2001; Brew, Riley, and Walta 2009). It is, therefore, possible to identify some questions or hypotheses that may be answered in future research, particularly to understand how the utility of alternative methods combined with traditional ones may enhance the teaching and learning processes. This would be useful for student skills development and for future professional contexts. A recent study comparing different methods of assessment showed that students present better results when they are assessed by open-ended questions rather than by multiple choice questions and that open-ended questions promoted conceptual understanding and deeper learning (Melovitz, DeFouw, Holland, and Vasan 2018).

This study suggests that involving students in alternative methods of assessment in higher education will lead to greater perception of it as being engaging and fair. This is a positive basis for moving towards greater diversity in educational assessment methods in that such diversity will be seen as having greater validity and potential for learning. In addition to contributing to knowledge about assessment in higher education, the small or trivial differences found regarding gender, year and study cycle suggest that institutions do not need to greatly concern themselves with these factors in terms of their assessment

policies and practices. What matters most to these students is moving beyond traditional forms of assessment across all levels and types of study.

References

Ashenafi, M. 2017. “Peer-assessment in higher education – twenty-first century practices, challenges and the way forward.” *Assessment & Evaluation in Higher Education* 42: 226-251.

Barreira, C., G. Bidarra, F. Monteiro, P. Vaz-Rebelo, and V. Alferes. 2017. “Avaliação das aprendizagens no ensino superior. Perceções de professores e estudantes nas universidades portuguesas.” *Revista Iberoamericana de Educación Superior* 8:24-36.

Biggs, J. 2003. *Teaching for Quality Learning at University*. SHRE and Open University Press.

Black, P., and D. Wiliam. 1998. “Assessment and Classroom Learning.” *Assessment in Education: Principles, Policy & Practice* 5: 7-74.

Boud, D., and N. Falchikov. 2007. *Rethinking assessment in higher education: learning for the long term*. Routledge.

Brew, C., P. Riley, and C. Walta. 2009. “Education Students and Their Teachers: Comparing Views on Participative Assessment Practices.” *Assessment & Evaluation in Higher Education* 34: 641–657.

Brown, G. T. L. 2009. “The reliability of essay scores: The necessity of rubrics and moderation.” In *Tertiary assessment and higher education student outcomes: Policy, practice and research*, edited by L. H. Meyer, S. Davidson, H. Anderson, R. Fletcher, P. M. Johnston, and M. Rees, 40-48. Wellington, NZ: Ako Aotearoa.

Brown, G. T. L., and Z. Wang. 2013. “Illustrating assessment: How Hong Kong university students conceive of the purposes of assessment.” *Studies in Higher Education* 38 (7): 1037-1057. doi:10.1080/03075079.2011.616955

Brown, G. T. L., E. R. Peterson, and S. E. Irving. 2009. “Beliefs that make a difference: Adaptive and maladaptive self-regulation in students’ conceptions of assessment.” In *Student perspectives on assessment: What students can tell us about assessment for learning*, edited by D. M. McInerney, G. T. L. Brown, and G. A. D. Liem, 159-186. Charlotte, NC US: Information Age Publishing.

Brown, G., J. Bull, and M. Pendlebury. 1997. *Assessing student learning in higher education*. Routledge.

Brown, S., and P. Knight. 1994. *Assessing Learners in Higher Education*. Kogan Page.

Browne, M. W., and R. Cudeck. 1993. “Alternative ways of assessing model fit.” In *Testing Structural Equation Models*, edited by A. Bollen and J. S. Long, 136–162. Sage.

Byrne, B. M. 2011. *Structural equation modeling with Mplus: Basic concepts, applications and programming*. Routledge Academic.

Cain, M. K., Z. Zhang, and K. Yuan. 2017. “Univariate and multivariate skewness and kurtosis for measuring nonnormality: Prevalence, influence and estimation.” *Behavior Research Methods* 49: 1716–1735.

Carvalho, A. 2013. “Students' perceptions of fairness in peer assessment: evidence from a problem-based learning course.” *Teaching in Higher Education* 18: 491-505.

Cattell, R. B., and B. Tsujioka. 1964. “The importance of factor-trueness and validity, versus homogeneity and orthogonality, in test scales.” *Educational and Psychological Measurement* 24: 3-30.

Chen, F. F. 2007. “Sensitivity of goodness of fit indexes to lack of measurement invariance.” *Structural Equation Modeling: A Multidisciplinary Journal* 14: 464–504.

Cheung, G. W., and R. Rensvold. 2002. “Evaluating goodness-of-fit indexes for testing measurement invariance.” *Structural Equation Modeling: A Multidisciplinary Journal* 9: 233–255.

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Cooper, S. 2017. “A collaborative assessment of students’ placement learning.” *Assessment & Evaluation in Higher Education* 42: 61-76.

Dochy, F., M. Segers, and D. Sluijsmans. 1999. “The use of self-, peer and co-assessment in higher education: a review.” *Studies in Higher Education* 24: 31-50.

Duncan, T., and A. Buskirk-Cohen. 2011. “Exploring Learner-Centered Assessment: A Cross Disciplinary Approach.” *International Journal of Teaching and Learning in Higher Education* 23: 246-259.

education: experience in using a metacognitive approach in five case studies.” *Assessment & Evaluation in Higher Education* 31: 415- 433.

Falchikov, N. 2005. *Improving assessment through student involvement: practical solutions for aiding learning in higher and further education*. Routledge Falmer.

Fitzpatrick, J. 2006. “An Evaluative Case Study of the Dilemmas Experienced in Designing a Self-Assessment Strategy for Community Nursing Students.” *Assessment & Evaluation in Higher Education* 31: 37–53.

Flores, M.A, M. Veiga Simão, A. Barros, and D. Pereira. 2015. “Perceptions of effectiveness, fairness and feedback of assessment methods: a study in higher education.” *Studies in Higher Education* 40: 1523–1534.

Flores, M.A., G. Brown, D. Pereira, C. Coutinho, P. Santos, and C. Pinheiro. 2020. “Portuguese university students’ conceptions of assessment: taking responsibility for achievement.” *Higher Education* 79: 377–394.

Goubeaud, K., and W. Yan. 2004. “Teacher educators' teaching methods, assessments, and grading: a comparison of higher education faculty's instructional practices.” *The Teacher Educator* 40: 1-16.

Gulikers, J., L. Kester, P. Kirschner, and T. Bastiaens. 2008. “The effect of practical experience on perceptions of assessment authenticity, study approach, and learning outcomes.” *Learning and Instruction* 18: 172-186.

Hair, J. F., W. Black, B. Babin, and R. Anderson. 2009. *Multivariate data analysis* (7th ed.). Prentice Hall.

Harland, T., A. McLean, R. Wass, E. Miller, and K. Sim. 2015. “An assessment arms race and its fallout: high-stakes grading and the case for slow scholarship.” *Assessment & Evaluation in Higher Education* 40: 528-541.

Hu, L., and M. Bentler. 1999. “Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives.” *Structural Equation Modeling: A Multidisciplinary Journal* 6: 1–55.

Huisman, B., N. Saab, J, van Driel, and P. van den Broek. 2018. “Peer feedback on academic writing: undergraduate students’ peer feedback role, peer feedback perceptions and essay performance.” *Assessment & Evaluation in Higher Education* 43: 955-968.

Jessop, T., Y. El Hakim, and G. Gibbs. 2014. “The whole is greater than the sum of its parts: a largescale study of students’ learning in response to different programme assessment patterns.” *Assessment & Evaluation in Higher Education* 39: 73-88.

Kim, H.Y. 2013. “Statistical notes for clinical researchers: Assessing normal distribution using skewness and kurtosis.” *Restorative Dentistry & Endodontics* 38: 52-54.

Korkmaz, S., D. Goksuluk, and G. Zararsiz. 2014. “MVN: An R Package for assessing multivariate normality.” *The R Journal* 6: 151-162.

Lew, M., W. Alwis, and H. Schmidt. 2010. “Accuracy of Students’ Self-assessment and their Beliefs about its Utility.” *Assessment & Evaluation in Higher Education* 35: 135–156.

Light, G., and R. Cox. 2003. *Learning & Teaching in Higher Education: The Reflective Professional*. Sage Publications.

Lipnevich, A. A., T. R. Guskey, D. Murano, and J. K. 2020. “What do grades mean? Variation in grading criteria in american college and university courses.” *Assessment in Education: Principles, Policy & Practice* 1-21. doi:10.1080/0969594X.2020.1799190

MacLellan, E. 2004. “How convincing is alternative assessment for use in higher education?” *Assessment & Evaluation in Higher Education* 29: 311-321.

Marsh, H. W., K. Hau, and Z. Wen. 2004. “In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler’s (1999) findings.” *Structural Equation Modeling* 11: 320-341.

Matos, D., S. Cirino, and G. Brown. 2009. “Students’ conceptions of assessment in higher education in Brazil.” In *Student perspectives of assessment. What students can tell us about assessment for learning*, edited by D.M. McInerney, G.T.L. Brown, and G.A.D. Liem, 235-253. Information Age.

Matos, D., S., Cirino, G. Brown, and W. Leite. 2013. “A avaliação no ensino superior: Concepções múltiplas de estudantes brasileiros [assessment in higher education: Multiple conceptions of brazilian students].” *Estudos em Avaliação Educacional* 24: 172-193.

McConlogue, T. 2012. “But is it fair? Developing students’ understanding of grading complex written work through peer assessment.” *Assessment & Evaluation in Higher Education* 37: 113-123.

McGarr, O., and A. Clifford. 2013. “Just enough to make you take it seriously: exploring students’attitudes towards peer assessment.” *Higher Education* 65: 677-693.

Melovitz, C., D. Defouw, B. Hooland, and N. Vasan. 2018. “Analysis of testing with multiple choice versus open-ended questions: Outcome-based observations in an anatomy course.” *Anatomical Sciences Education* 11: 254-261.

Merry, S., and P. Orsmond. 2018. “Peer assessment: the role of relational learning through communities of practice.” *Studies in Higher Education* DOI: 10.1080/03075079.2018.1544236

Micán, A., and L. Medina. 2017. “Boosting vocabulary learning through self-assessment in an English language teaching context.” *Assessment & Evaluation in Higher Education* 42: 398-414. https://doi.org/10.1080/02602938.2015.1118433

Mok, M., C. Lung, D. Cheng, R. Cheung, and M. Ng. 2006. “Self-assessment in higher

Muthén, L. K., and B. Muthén. 2012. *Mplus user’s guide* (7th ed.). Muthén & Muthén.

Myers, C., and S. Myers. 2015. “The use of learner-centered assessment practices in the United States: the influence of individual and institutional contexts.” *Studies in Higher Education* 40: 1904-1918.

Orsmond, P., and S. Merry. 2013. “The Importance of Self-assessment in Students’ Use of Tutors’ Feedback: A Qualitative Study of High and Non-high Achieving Biology Undergraduates.” *Assessment & Evaluation in Higher Education* 38: 737–753.

Panadero, E. 2016. “Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions.” In *Handbook of human and social conditions in assessment*, edited by G. T. L. Brown and L. R. Harris , 247-266. New York: Routledge.

Panadero, E., G. Brown, and J. Strijbos. 2016. “The Future of Student Self-Assessment: a Review of Known Unknowns and Potential Directions.” *Educational Psychology Revie*, 28*:* 803-830.

Panadero, E., J. Fraile, J. Fernández Ruiz, D. Castilla-Estévez, and M. Ruiz. 2018. “Spanish university assessment practices: examination tradition with diversity by faculty.” *Assessment & Evaluation in Higher Education* Online first. 10.1080/02602938.2018.1512553

Patton, C. 2012. “Some Kind of Weird, Evil Experiment’: Student Perceptions of Peer Assessment.” *Assessment & Evaluation in Higher Education* 37: 719–731.

Pereira, D. 2011. *A avaliação das aprendizagens no ensino superior na perspetiva dos estudantes: um estudo exploratório* (Master’s thesis). Retrieved from https://repositorium.sdum.uminho.pt/bitstream/1822/19122/1/Diana%20Alexandra%20Ribeiro%20Pereira.pdf

Pereira, D. 2016. *Assessment in higher education and quality of learning: perceptions, practices and implications* (Doctoral dissertation). Retrieved from https://repositorium.sdum.uminho.pt/bitstream/1822/43445/1/Diana%20Alexandra%20Ribeiro%20Pereira.pdf

Pereira, D., L. Niklasson and M. A. Flores 2017b. “Students’ perceptions of assessment: a comparative analysis between Portugal and Sweden.” *Higher Education* 73: 153–173.

Pereira, D., M. Flores, and A. Barros. 2017a. “Perceptions of Portuguese undergraduate students about assessment: a study in five public universities.” *Educational Studies* 43: 442–463.

Pereira, D., M. Flores, and L. Niklasson. 2016. “Assessment revisited: a review of research in assessment and evaluation in higher education.” *Assessment & Evaluation in Higher Education* 41: 1008–1032.

Perrenoud, P. 1999. *Avaliação. Da excelência à regulação das aprendizagens: entre duas lógicas*. Porto Alegre: ARTMED.

Price, M., K. Handley, J. Millar, and B. O'Donovan. 2010. “Feedback: all that effort, but what is the effect?.” *Assessment & Evaluation in Higher Education* 35: 277-289.

Sadeghi, K., and Z. Khonbi. 2015. “Iranian university students’ experiences of and attitudes towards alternatives in assessment*.” Assessment & Evaluation in Higher Education* 40: 641-66.

Sambell, K., and L. McDowell. 1998. “The construction of the hidden curriculum: messages and meanings in the assessment of student learning.” *Assessment and Evaluation in Higher Education* 23: 391-402.

Schermelleh-Engel, K., H. Moosbrugger, and H. Muller. 2003. “Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures.” *Methods of Psychological Research Online* 8: 23–74.

Segers, M., and F. Dochy. 2001. “New Assessment Forms in Problem-based Learning: The Value added of the Students’ Perspective.” *Studies in Higher Education* 26: 327–343.

Sluijsmans, D., F. Dochy, and G. Moerkerke. 1999. “Creating a learning environment by using self, peer- and co-assessment.” *Learning Environments Research: An International Journal* 1: 293-319.

Struyven, K., and J. Devesa. 2016. “Students' perceptions of novel forms of assessment.” In *Handbook of Human and Social Conditions in Assessment*, edited by G. T. L. Brown and L. R. Harris, 129-144. Routledge.

Struyven, K., F. Dochy, and S. Janssens. 2003. “Students’ perceptions about new modes of assessment in higher education: A review.” In *Optimising new modes of*

assessment: In search of qualities and standard, edited by M. Segers, F. Dochy, and E. Cascallar, 171-223. Kluwer Academic Publishers.

Struyven, K., F. Dochy, and S. Janssens. 2005. "Students' perceptions about evaluation and assessment in higher education: a review." *Assessment & Evaluation in Higher Education* 30: 331–347.

van den Berg, I., W. Admiraal, and A. Pilot. 2006. "Peer Assessment in University Teaching: Evaluating Seven Course Designs." *Assessment & Evaluation in Higher Education* 31: 19–36.

Vickerman, P. 2009. "Student Perspectives on Formative Peer Assessment: An Attempt to Deepen Learning?" *Assessment & Evaluation in Higher Education* 34: 221–230.

Vogl, E., and R. Pekrun. 2016. "Emotions that matter to achievement: Student feelings about assessment." In *Handbook of human and social conditions in assessment*, edited by G. T. L. Brown and L. R. Harris, 111-128. New York: Routledge.

Wang, Z., and G. T. L. Brown. 2014. "Hong Kong tertiary students' conceptions of assessment of academic ability." *Higher Education Research & Development* 33 (5): 1063-1077. doi:10.1080/07294360.2014.890565

Webber, K. 2012. "The Use of Learner-Centered Assessment in US Colleges and Universities." *Research in Higher Education* 53: 201-228.

Wen, M., and C. Tsai. 2006. "University Students' Perceptions of and Attitudes toward (online) Peer Assessment." *Higher Education* 51: 27–44.

Wilson, L., S. Ho, and R. Brookes. 2018. "Student perceptions of teamwork within assessment tasks in undergraduate science degrees." *Assessment & Evaluation in Higher Education* 43: 786-799.

Wilson, M. J., M. Diao, and L. Huang. 2015. "I'm not here to learn how to mark someone else's stuff': an investigation of an online peer-to-peer review workshop tool." *Assessment & Evaluation in Higher Education* 40: 15-3.

Yan, Z., and G. Brown. 2017. "A cyclical self-assessment process: towards a model of how students engage in self-assessment." *Assessment & Evaluation in Higher Education* 42: 1247-1262.

Yang, Y., and X. Liang. 2013. "Confirmatory factor analysis under violations of distributional and structural assumptions." *International Journal of Quantitative Research in Education* 1: 61-84.

Yuan, K. H., F. Yang-Wallentin, and P. Bentler. 2012. "ML Versus MI for missing data with violation of distribution conditions." *Sociological Methods and Research* 41: 598–629.