The reliability and sources of error of using rubrics-based assessment for student projects

José-Luis Menéndez-Varela[a]

menendez@ub.edu

Eva Gregori-Giralt[a]

gregori@ub.edu

[a]University of Barcelona. Faculty of Fine Arts. Adolf Florensa, s/n. 08028 Barcelona (Spain)

Tel.: +34 934039991. Fax: +34 934035976

Corresponding author

Eva Gregori-Giralt. University of Barcelona. Faculty of Fine Arts. Adolf Florensa, s/n. 08028 Barcelona (Spain). gregori@ub.edu Tel.: +34 934039991. Fax: +34 934035976

Abstract

Rubrics are widely used in higher education to assess performance in project-based learning environments. To date, the sources of error that may affect their reliability have not been studied in depth. Using the Generalizability theory as its starting-point, this article analyses the influence of the assessors and the criteria of the rubrics on the assessment of two service-learning projects. A sample of 365 novice students studying for three different undergraduate degrees was evaluated by eight student assessors and two teachers at three stages of assessment. Depending on the type of project and the stage of assessment, between 19.27–39.55% of the total variance

was attributed to the quality of the projects, 0–7.49% to the main effect of the raters, and 3.44–17.3% to the main effect of the criteria. The results demonstrated that acceptable levels of reliability (≥ .70) were obtained with three raters and eight criteria or four raters and nine criteria in contexts of relative or absolute decisions, respectively.

Keywords

Rubric, performance assessment, Generalizability theory, reliability, service-learning.

Introduction

In education, reliability is an evaluation of the accuracy of inferences extracted from a set of scores (Shavelson and Webb 1991). Consensus among assessors regarding assessment criteria and procedures, and regarding their perceptions of the phenomena assessed, are key factors in reliability. However, reliability also depends on: a) the suitability of the assessment tasks and items for evaluating the performance under examination, b), the suitability of the number of assessment tests and their distribution over time for evaluating the intended learning outcomes, and c) the differences in the reactions of students to the assessment tasks, items and occasions. The interaction of these four factors will decide whether the scores represent true differences in student performances regardless of when the tests are performed and in what circumstances, and regardless of who scores the student's responses, when, and in what circumstances.

Some authors have claimed that attempts to increase the reliability of a rubrics-based assessment system may be to the detriment of its validity (Montgomery 2002), and have stressed that validity should be prioritised (Jonsson and Svingby 2007). According to Messick (1989), validity in the educational context refers to the adequacy, appropriateness and usefulness of the

inferences derived from an assessment system to the purposes of learning; thus, the accuracy of the assessment system is a condition for ensuring its validity. Therefore, the conflict between validity and reliability occurs if the assessment system only focuses on the aspects of the construct that are easy to measure. By avoiding key aspects which are difficult to measure, the accuracy of the assessment system is improved but its fit to the object and to the aim of the assessment is reduced. In a well-designed assessment system, the relationship between reliability and validity means that an error in the former weakens the latter (Sadler 2009) because the causes of the inconsistency introduce construct-irrelevant variance (Iramaneerat et al. 2008).

The use of rubrics has a positive impact on the reliability of performance assessment because it improves the transparency and consistency of the assessment process (e.g., Schlitz et al. 2009; Wolf and Goodwin 2007) and lowers the risk of inaccurate scoring and bias (Oakleaf 2008). The key strength of rubrics is the fact that the qualitative descriptions of the criteria reduce the interference of assessors' personal preferences (Reynolds et al. 2009) and of differences in the interpretation of the criteria (Venning and Buisman-Pijlman 2013), especially when complex performances are being assessed. The correct use of rubrics differs from an assessment based on an overall impression of performance since it is grounded in a prior agreement on the aspects to be evaluated, on the criteria to use, and on the levels of performance that should be taken as a reference (Bird and Yucel 2013; Taub et al. 2011).

The use of rubrics does not in itself guarantee that a performance assessment system will present acceptable reliability (Gugiu, Gugiu, and Baldus 2012; Jonsson and Svingby 2007; Reddy and Andrade 2010). However, given that the design of rubrics is a reflection on professional knowledge and experience and their use requires consensus between practitioners to achieve better understanding and learning, a performance assessment without rubrics tends to

present poorer reliability scores (Wolf and Stevens 2007). Although the design of rubrics undoubtedly presents specific technical difficulties, the most serious problems arise from the complexity of the objects of evaluation required in competencies-based higher education and from the fact that rubrics are not yet habitually included in assessment systems.

Because of the complexity of performance assessment, analytic and topic-specific rubrics are recommended (e.g., Crotwell Timmerman et al. 2011; Jonsson 2014; Jonsson and Svingby 2007). Prior to their use, training programmes of varying lengths are offered which use complementary resources such as feedback codes, annotated exemplars and sample comments (e.g., Bird and Yucel 2013) and establish specific features and procedures such as the number of participating assessors (e.g., Gugiu et al. 2012; Tucker 2013) and the augmentation of the rating scale (e.g., Crotwell Timmerman et al. 2011). There is no doubt that training programmes are beneficial, but their use does not allay all the doubts regarding their ability to significantly improve reliability (Stuhlmann et al. 1999). The fact is that these programmes do not necessarily manage to integrate rubrics in the professional culture of the assessors, especially if they have been designed by experts from other fields rather than by the people who are intended to use them.

This second problem is the reason for the establishment of authentic learning communities to encourage learning and professional development (O'Malley 2010; Ward and Selvester 2012). The idea is that it is the assessors who should create the assessment systems and establish the consensus necessary to guarantee a shared understanding of standards and procedures; in this regard, the pre- and post-marking moderation meetings aimed at resolving the

problems of the assessment practice in its context are the ones that obtain the best results (Crotwell Timmerman et al. 2011; Vardi 2013).

Sources of error in rubrics-based performance assessment

Any assessment practice runs the risk of measurement error. There is always the possibility that extraneous information will be introduced (Sadler 2009); assessment may be affected by differences between disciplines (Knight 2006), or by differences in the experience of the assessors (Howell 2014), in the interpretations of the assessment criteria (Tan and Prosser 2004), or in the application of the assessment procedures (Jonsson and Svingby 2007). The time dedicated to the assessment may also play a role (Pinot de Moira et al. 2002).

Aspects of the design of the assessment instrument also affect the reliability. In the specific case of rubrics, the quality definitions present the added difficulty of recording both the explicit and the tacit dimensions of knowledge that underpin the assessment process (Bird and Yucel 2013). Other aspects, apparently less important, also influence the reliability. An excessive number of criteria, for instance, reduces the reliability because it is impossible for the assessor to manage them efficiently (Baryla, Shelley, and Trainor 2012). Much the same occurs with the proliferation in performance levels (Wolf and Stevens 2007), which makes it more difficult for assessors to distinguish clearly between the degrees of performance, and represents a significant overload.

With this diverse range of systematic errors, studies of the most frequently used types of performance assessment are needed that are able to identify the main sources of error, their relative weight in the overall measurement error, and the interactions that occur between them.

The contribution of Generalizability theory

The estimates of reliability provided by classical test theory have come in for criticism from psychometricians. For example, the limitations of correlation coefficients have been stressed, since they conceal the discrepancies that emerge when the rank-orders of the groups comprising the scores are similar (Gugiu et al. 2012). The suitability of the Cronbach's alpha coefficient for interpreting rater reliability has also been questioned, since, as a single-occasion reliability estimate, it does not reflect the inconsistencies across occasions (Vispoel and Tao 2013). However, the main limitation of classical test theory is that it treats the measurement error as an undifferentiated random error term (Feldt and Brennan 1989).

Generalizability theory (G theory henceforth) manages to identify and quantify multiple sources of measurement error and their relative importance, and improves their control by modifying of the measurement procedure (Cronbach et al. 1972). G theory works with facets, that is, uncorrelated components of systematic error variance, and so it allows the establishment of global strategies of control over each one of the facets.

In the first stage, referred to as the G (generalizability) study, researchers define the conditions of the study (known as the universe of admissible observations) and quantify each source of variance in a single analysis. In the second stage, known as the D (decision) study, researchers define the number and breadth of facets on which they wish to generalize (the universe of generalization), specify what constitutes measurement error in accordance with the type of decisions to be made – relative or absolute – and use the results of the G study to propose a more efficient application of the assessment system which meets the established requirements of reliability (Shavelson and Webb 1991).

In the broader ambit of performance assessment, in the last 15 years empirical studies based on G theory have been carried out mainly in the area of the health sciences and, to a lesser extent, in the education sciences. Few studies are available on the reliability of rubrics-based performance assessment. Two studies in the field of education sciences have been published, but it is surprising to find that there are none from the health sciences, and that the other three come from ambits that are underrepresented in higher education research (Crotwell Timmerman et al. 2011; Gugiu et al. 2012; Sudweeks, Reeve, and Bradshaw 2005), as indeed does our study. We know of no publications on rubrics-based performance and G theory in university arts studies. These four studies respond to the call made by Reddy and Andrade (2010) for more rigorous methods and validity and reliability analyses of the use of rubrics from expanded geographical and cultural perspectives.

In these studies, certain limitations have been detected which we have tried to avoid in the present work. First, only Gugiu et al. (2012) and Sudweeks et al. (2005) provided information on all essential aspects of the research; the other articles lacked information on the study design, universe score, main sources of error variance, and test statistics (G or φ). Second, the studies recorded the participation of different types of raters, but none of them combined expert and student raters. Third, none of the studies included a continuous assessment nor was studied the effect of the item facet, in spite of its direct relation with the rubrics' criteria.

Research questions and hypotheses

This empirical study focused on two research questions and associated hypotheses:

1. What amount of variance should be attributed to the sources of measurement error rather than to the quality of the students' projects? The difficulties of a performance assessment involving complex cognitive abilities call for a revision of the interpretations of the sources of error, especially those linked to certain interactions which are identified as error variance in G theory. We expect to show that, when particular circumstances obtain, the person-by-item interaction should be understood as a different response from students to a complex assessment task, in so far as it activates certain cognitive skills which are not homogeneously distributed among them. Thus, what is habitually considered a source of error is in fact true variance.

2. What changes should be introduced to improve reliability? Previous studies of G theory that have included generalizability and dependability coefficients (G or $E\rho^2$, and φ) showed wide variations in the data and, in general, presented low coefficients. In our case, the existence of a three-stage assessment procedure and the participation of student and teacher assessors was expected to increase the heterogeneity of the scores even more; however, the inclusion of ten assessors should allow greater room for manoeuvre depending on the results of the D studies. We expected to obtain good test statistics (G and φ coefficients) with a number of assessors that reflected more closely the reality of the teaching context, and that this reduction might have to be compensated by an increase in the number of items.

Method

Participants

As part of the teaching of three six-month compulsory subjects, two teachers created a project-based learning environment in which 365 first-year students designed service-learning projects in groups of five members (total: 73 work groups). These students were studying for the History of

Art, Fine Arts and Conservation-Restoration undergraduate degrees at a large public university. The mean age of the students was 24.95 years (SD = 11.23; mode = 19 and median = 20); 79.34% were female, and 62.5% were not in employment.

The projects were evaluated by two teachers and eight third-year undergraduate students (five women and three men). The two teachers, one woman and one man, had 14 and 20 years of teaching experience respectively. The eight student assessors had previously studied the subject with one of the teachers and were therefore familiar with the teaching environment. This was the main criterion for the recruitment of the student assessors; their participation was totally voluntary and they did not receive any compensation.

Instructional settings and resources

Depending on the subject, students performed a service-learning project for a group at risk of social exclusion (henceforth, SLSE), or for a cultural heritage element at risk of extinction (henceforth, SLCH). Inside these contexts, they chose the aim of their projects. Each project was presented orally in the classroom at three different stages of the semester. In each stage, the students revised their previous work and added new elements, and thus completed their end of year project. The weight of each stage in the final grade rose progressively from 10% to 30% and then to 60%

The projects were assessed using an analytical rubric designed by the teachers of the subjects during the previous year with the aid of three postgraduate students in order to adapt the language used in the rubric more closely to the student profile (Jonsson 2014; Reddy and Andrade 2010). The rubric helped students to understand the learning goals and provided them with useful information for the development of their projects, helped to focus the work sessions

and give students in-task guidance, and made it possible to carry out a continuous assessment in which undergraduates and teachers scored the projects and their oral presentation individually.

The rubric comprised the following dimensions: the *Project rationale*, that is, the title, objectives, analysis of the social group or heritage element at risk, and of the context of the application of the project (present in the three stages); *Documentation*, the sources and documents consulted (in the three stages); *Previous work*, a selection and study of similar projects (second and third stages); and finally, *Activities*, a detailed description and discussion of the activities in the project (the third stage).

No training programme was carried out, since the student assessors were already familiar with the two types of service-learning projects, the rubric and the details of the assessment system. In each semester, two assessment teams, each comprising one teacher and two undergraduates, were formed to evaluate all students' projects.

The student assessors and the teachers attended the oral presentations with the assessment sheets designed for each stage which were stored in the subject's course management system. The assessment protocol required each assessor to take notes on the aspects they considered relevant in relation to each criterion in the rubric, and also to take notes on any incident that disturbed their use of the rubric. The assessment procedure included moderation discussions attended by all the assessors, which were held immediately after each assessment session. In these discussions, participants described their impressions of the projects presented and backed up their views with the help of the notes they had taken for each criterion in the course of the oral presentations. The scores were not discussed at these sessions; later, each rater scored each student individually.

Data analysis

The study data included the 9855 scores (5022 from SLSE projects and 4833 from SLCH projects) given by the raters to all 365 students on all the assessment criteria which respectively constituted the rubric of the three stages of assessment.

Inside the framework of Generalizability theory, a two-facet fully-crossed random-effect design (*p* x *r* x *i*) was chosen. So, there were seven variance components: on the one hand, the main effect of the projects (*p*), raters (*r*) and items of rubrics (*i*), and on the other, the two-way interaction between projects and raters (*pr*), projects and items (*pi*) and raters and items (*ri*). Finally, the residual error variance in which the interaction between all the facets and other unidentified sources of variability (*pri,e*) was included. The two types of service-learning projects were analysed separately.

G theory distinguishes between relative or absolute decision-making contexts. In education this means, for instance, that scores serve to establish students' relative standing or, in contrast, to determine whether students have reached a specific level of performance. This distinction is important because it affects the definition of measurement error and, as a result, the appropriate type of coefficient. The generalizability coefficient (G or $E\rho^2$) only counts the variance that is caused by the interactions between the object of measurement (the projects in our case) and other variables ($\sigma^2 pr$, $\sigma^2 pi$, and $\sigma^2 pri,e$ in this study) as a measurement error; it is only useful for relative decisions. In contrast, the dependability coefficient (φ) considers all variance not attributable solely to the object of measurement ($\sigma^2 r$, $\sigma^2 i$, $\sigma^2 pr$, $\sigma^2 pi$, $\sigma^2 ri$ and $\sigma^2 pri,e$) as measurement error, and should be used for absolute decisions (see: Brennan 1996; Cronbach et al. 1972; Shavelson and Webb 1991). In so far as both coefficients are equivalent to classical reliability coefficients (e.g., Gebril 2009; Sudweeks et al. 2005; Zhang, Johnston, and Kilic

2008), the accepted lower limit of reliability is 0.7. The analyses were carried out using the general linear model of the SPSS 23.0 software.

In addition, an ANOVA was performed to study the existence of statistically significant differences between the scores of the items in each project and stage of assessment. SPSS 23.0 was also used for this analysis.

Results

Generalizability studies across raters and items as sources of error

Table 1 shows the data from the six G studies performed following the *p* x *r* x *i* design: one for each type of service-learning project at each of the three stages of assessment. The residual variance (*pri,e*) was acceptable in most cases (below 8%), with the sole exceptions of the second stages of both projects (with a peak value of 22.16% in the SLSE project), suggesting that the results were relatively unaffected by other unexplained sources of variance. The variance attributable to the object of measurement (*p*) – therefore, the true score variance – was practically always above 22% and even reached 39.55% in one case.

With regard to the sources of measurement error, we should stress first that the variances attributable to the main effects of the raters (*r*) remained low (surpassing 5% on only one occasion). Therefore, there were no marked differences in stringency or leniency between the assessors that might have distorted the assessment system. The two-way interaction between projects and raters (*pr*) strengthened this inference: percentages remain low, with the sole exceptions of the second and third stages of the SLSE project, and so the rank-order of the projects did not vary substantially depending on the assessors. In fact, the poorest result

attributable to the intervention of the raters accounted for only 20.04% of the variance, taking all the rater effects together (*r*, *pr* and *ri*), and this percentage was actually a notable exception.

Second, the variances attributable to the main effects of the items (*i*) were in the moderate range; they were not generally above 8%, with the exception of the first and third stages of the SLCH project, in which rates of 17.30 and 14.17% respectively were obtained. This means that the criteria of the rubrics did not generally vary in terms of difficulty, which underlines another component of the accuracy of the assessment system, but this variation affected the two projects in different ways. The small variance associated with the two-way interaction between raters and items (*ri*) shows that the student and professor assessors interpreted and applied the various criteria in a similar way.

Third, the two-way interaction between projects and items (*pi*) was by a long way the main source of measurement error with values that, with two exceptions, accounted for more than 41% of the variance. These results indicate that the projects were ordered differently depending on the item used; in other words, the students showed different levels of performance in their projects depending on the criteria of rubrics used.

Table 1 Estimates of variance components, and G and φ coefficients for the main effect of the projects, raters, items of rubrics and their interactions (*p* x *r* x *i* design)

Effect	Service-learning projects (social exclusion)						Service-learning projects (cultural heritage)					
	Stage 1 (n=1116)		Stage 2 (n=1674)		Stage 3 (n=2232)		Stage 1 (n=1074)		Stage 2 (n=1611)		Stage 3 (n=2148)	
	VC	%	VC	%	VC	%	VC	%	VC	%	VC	%
$\sigma^2(p)$	3.52	26.59	4.20	39.55	1.80	32.37	2.79	27.55	2.03	19.27	1.96	22.08

$\sigma^2(r)$	0.07	0.54	0.49	4.63	0.26	4.76	0.35	3.46	0.00[a]	0.00	0.66	7.49
$\sigma^2(i)$	1.00	7.60	0.36	3.44	0.26	4.72	1.75	17.30	0.75	7.12	1.26	14.17
$\sigma^2(pr)$	0.00[a]	0.00	0.96	9.07	0.79	14.29	0.00	0.00	0.19	1.82	0.24	2.78
$\sigma^2(pi)$	5.44	41.02	1.77	16.71	1.94	34.97	4.669	46.07	6.02	57.05	4.10	46.09
$\sigma^2(ri)$	2.41	18.18	0.47	4.44	0.05	0.99	0.06	0.63	0.36	3.42	0.29	3.33
$\sigma^2(pri,e)$	0.83	6.33	2.35	22.16	0.44	7.92	0.50	5.00	1.24	11.78	0.36	4.05
$E\rho^2$	.36		.45		.36		.35		.21		.29	
φ	.27		.40		.32		.28		.19		.22	

Note. p = students' projects; r = raters; i = items; VC = variance component; % = relative variance component; $E\rho^2$ = Generalisability coefficient; φ = Dependability coefficient; n = scores.

[a] Small negative variance components were set to zero.

Decision studies

Table 2 shows the D studies for different numbers of raters and items at each stage of assessment. To summarize the data as succinctly as possible, we include from eight to ten items in the case of one, three, four, five and six raters; for one and six raters we also include from two to four items which were the dimensions of the rubric used in the three stages. In the real case of six raters per semester and two to four items, the generalizability and dependability coefficients (G or $E\rho^2$, and φ) reached the following values: the former between .39 and .86, and the latter between .36 and .82. These results challenge the reliability of the assessment system in contexts of relative decisions, and even more so in contexts of absolute decisions.

We see that to maintain an acceptable G coefficient (≥ .70) for relative decisions it is enough in general to have three raters and eight criteria, but a good G coefficient (≥ .80) would

require six raters and nine criteria. In contexts of absolute decisions, four raters and nine criteria would only produce an acceptable dependability coefficient (φ).

Table 2 Estimates of G and φ coefficients for D studies

D studies		Service-learning projects (social exclusion)			Service-learning projects (cultural heritage)		
No. of raters	No. of items	Stage 1	Stage 2	Stage 3	Stage 1	Stage 2	Stage 3
		$E\rho^2/\varphi$	$E\rho^2/\varphi$	$E\rho^2/\varphi$	$E\rho^2/\varphi$	$E\rho^2/\varphi$	$E\rho^2/\varphi$
1	1	.36/.27	.45/.40	.36/.32	.35/.28	.21/.19	.29/.22
	2	.53/.42	.58/.52	.48/.43	.52/.42	.35/.32	.44/.33
	3	.63/.52	.64/.57	.53/.48	.62/.51	.44/.41	.53/.40
	4	.70/.59	.68/.61	.56/.51	.68/.57	.50/.48	.59/.45
3	8	.83/.79	.87/.83	.77/.73	.82/.75	.70/.68	.76/.66
	9	.85/.81	.87/.84	.78/.74	.84/.77	.72/.70	.78/.68
	10	.86/.82	.88/.84	.79/.75	.85/.78	.74/.72	.80/.69
4	8	.83/.79	.89/.85	.80/.76	.82/.75	.71/.69	.77/.68
	9	.85/.81	.89/.86	.81/.77	.84/.77	.73/.71	.79/.70
	10	.86/.83	.90/.87	.82/.78	.85/.79	.75/.73	.80/.72
5	8	.84/.80	.90/.87	.81/.78	.82/.76	.71/.69	.77/.69
	9	.85/.82	.90/.88	.82/.79	.84/.78	.73/.71	.79/.71
	10	.86/.83	.91/.88	.83/.80	.85/.79	.75/.73	.81/.73
6	2	.56/.50	.77/.73	.61/.58	.54/.46	.39/.36	.48/.40
	3	.66/.60	.83/.79	.69/.66	.64/.56	.49/.46	.58/.50
	4	.72/.67	.86/.82	.74/.71	.70/.62	.56/.53	.64/.56
	8	.84/.80	.91/.88	.82/.80	.82/.76	.72/.69	.78/.70
	9	.85/.82	.91/.89	.83/.81	.84/.78	.74/.72	.80/.72
	10	.86/.83	.92/.89	.84/.82	.85/.80	.76/.74	.81/.74

Note. $E\rho^2$ = Generalisability coefficient; φ = Dependability coefficient.

Study of the item effect

Table 3 shows statistically significant differences between the scores on the items or criteria of the rubric in each of the service-learning projects and at each stage of assessment. The lowest scores corresponded to the project rationale. This constituted the greatest cognitive challenge for the students, which in fact they only negotiated successfully in the third stage. The criteria of documentation and previous work did not present serious problems. The activities criterion was added only in the last stage of the projects, and so the students could not review or improve it; this is shown by the fact that it was the criterion with the lowest scores in the third stage.

Table 3 ANOVA of the items in each project and stage

Project	Stage	Group	Df.	F	Order of items (from lowest to highest scores)
Service-learning	1	Between	1	58.04[a]	PR and D
(social exclusion)		Within	1114		
(n = 5022 scores)	2	Between	2	23.36[a]	PR, D and PW
		Within	1671		
	3	Between	3	29.08[b]	A, D, PR and PW
		Within	2228		
Service-learning	1	Between	1	119.64[a]	PR and D
(cultural heritage)		Within	1072		
(n = 4833 scores)	2	Between	2	48.96[c]	PR, PW and D
		Within	1608		
	3	Between	3	89.42[b]	A, PW, PR and D
		Within	2144		

Note. PR = project rationale; D = documentation; PW = previous work; A = activities. Df = degrees of freedom. Maximum value permitted by chance ($p < .05$): [a] = 3.8; [b] = 2.6; [c] = 3.

Discussion

Sources of variance in the assessment of students' projects

In G theory, raters are considered to be among the principal factors that reduce reliability in performance assessments involving complex cognitive abilities. In simulated workplace-based assessment of clinical competencies (e.g., Cook et al. 2010; De Lima et al. 2013) or communication skills (e.g., Iramaneerat et al. 2008; Raymond, Harik, and Clauser 2011) using standardised instruments, assessor leniency/stringency emerged as one of the major causes of unreliability. These studies confirm Brennan (1996)'s claim that the rater reliability of the performance assessment is affected when the students choose the themes of their essays or projects, or produce unique products.

We analysed the results of 365 novice students on three different undergraduate degrees, in two different types of service-learning projects, with a three-stage assessment system involving 10 assessors (eight of them undergraduate students). In this situation, the variances attributable to the main effect of the raters (*r*: 0% to 7.49% of the whole variance) and to project-by-rater (*pr*: 0% to 14.29%) and rater-by-item (*ri*: 0.63% to 18.18%) interactions obtained mainly low values; even taking all the rater effects together, the variance reached a peak of 20%. This global effect was considerably more notable in the SLSE project than in the SLCH project.

The data show acceptable inter-rater reliability between expert and novice assessors. The maintenance of these low values between stages of assessment also indicates good intra-rater reliability. Our study corroborates others in which the rater facet did not appear as one of the

main sources of error. Although there are as yet very few studies of rubrics from the perspective of G theory, and although in the studies available the residual error variance generally appeared as the first (Gebril 2009; Sudweeks et al. 2005) or the second source of variability (Zhang et al. 2008), the use of well-designed rubrics increases the assessors' consistency. In our study, moreover, the moderation discussions played an important role in establishing a common assessment procedure among assessors; their impact was more evident in the first assessment sessions, before the calibration process was concluded.

The main effects of the items (*i*) rose to moderate levels with a variance ranging from 3.44% to 17.30%. These percentages behaved in different ways in the two types of projects and, in contrast to the rater effect, higher percentages were obtained in the SLCH project. This indicates an appreciable difference in the degree of difficulty of the items, and a different level of adaptation of the rubric's criteria to the types of project. Other studies in which the type of case or task effects obtained the highest amount of variance (Gebril 2009; Guiton et al. 2004; Leung, Wang, and Chen 2012; Swanson, Norman, and Linn 1995) have shown how difficult it is to design assessment tests homogeneous enough to evaluate complex performances. Our study corroborates this statement, because the results showed notable differences in the item facet (*i*) between different types of projects but not between different stages of assessment (see Table 1). However, two findings recommend caution: first, the variance attributable to the items was the third source of variance even in the SLCH project; second, the two-way interactions between raters and items (*ri*) and projects and items (*pi*) both showed similar trends in all the phases and in both types of project. Sudweeks et al. (2005) stressed that the main effects for tasks and other related interactions reveal differences in the nature of the tasks but also in the student's prior

knowledge or interest, to which one should add the influence of contextual factors such as the integration of this type of task or learning environment in the study plans.

If the person-by-task interaction shows the influence of content specificity (Iramaneerat et al. 2008), the person-by-item interaction may reflect different levels of achievement with respect to different learning outcomes in a context of complex performance assessment. Our study showed the highest variance in the interaction between projects and items (*pi*: 16.71–57.05%), which is generally considered a source of error. However, if this variance were interpreted as true variance, the universe score variance would rise from the current range of 19.27–39.55% to 56.26–76.62% with a very homogeneous distribution between projects and stages of assessment. This possibility was suggested by Praetorius, Lenske, and Helmke (2012) and the data from our study confirm their arguments. The low values obtained in the rater-by-item interaction (*ri*) reflect that the assessors were consistent in the use of all the criteria. These data, added to the percentages of the total variance attributable to all the rater effects together and main effects of the items, show that the variability is not due to the fact that the raters perceived the projects or applied the items in different ways, but because the projects showed different levels of achievement depending on the criteria. However, the existence of different numbers of items at each stage of assessment had a negative effect on the variance of the item facet, as was also reported in the study by Heijne-Penninga et al. (2008). Thus, the variable that aided the progress of students' learning was at the same time a cause of distortion in the analysis of the variance.

Improving the reliability of the assessment system

The D studies produced generalizability and dependability coefficients (G or $E\rho^2$, and φ) of .39–.86 and .36–.82 respectively in the real case of six raters and two, three or four rubric criteria, indicating that the levels of reliability fluctuated according to the type of project, stage and number of criteria. Our study thus confirms the difficulties of reaching adequate levels of generalizability in performance assessment contexts, already mentioned by Brennan (1996) and supported by other empirical studies on rubrics (see Crotwell Timmerman et al. 2011; Gebril 2009; Sudweeks et al. 2005; Zhang et al. 2008) or standardised instruments (Cook et al. 2010; Iramaneerat et al. 2008).

With a high main effect, the increase in the number of items or raters is the most effective way to improve reliability, even more than increasing the test administrations (Vispoel and Tao 2013) or tasks/cases (De Lima et al. 2013). However, low or moderate variance components mean that the increase in number does not entail clear improvements in the G and φ coefficients, as was the case in our study and in the study by Praetorius et al. (2012). In these circumstances, a balance must be struck between an acceptable reliability and the viability of the modifications, although this is not always possible (Kreiter et al. 2004). With three raters and eight items our study recorded G and φ coefficients of .70–.87 and .66–.83 respectively; that is, with increases considerably lower than in other studies (Cook et al. 2010; Dornan et al. 2012; Praetorius et al. 2012) we obtained similar coefficients. For absolute decision-making situations, four raters and nine criteria would produce an acceptable dependability coefficient ($\geq .70$).

These modifications are viable but their implementation requires time and effort. The number of four raters underlines the value of creating stable teacher teams who share areas of knowledge and teaching environments. Raising the number of criteria in the rubric from four to nine is perfectly feasible and even advisable in order to break down the ones that the students

find most difficult (see Table 3). For example, the item *project rationale* could be divided into its components (title, objectives, analysis of the social group or heritage element at risk, and of the context of the application of the project), and the *activities* item could be divided into a description of the activities performed, an analysis of problems and solutions, follow-up procedures and a review of the project. Thus, there would be a total of nine criteria, maintaining the items *documentation* and *previous work* in their original version.

The increase in the number of items does not necessarily entail greater difficulty in the assessment process since the new criteria will be less complex, but it does surpass the maximum of six criteria recommended in previous work (Wolf and Stevens 2007). However, other lines of study should be explored. First, the existence of a different number of items per stage of assessment has a negative influence on the distribution of the variance; the use of the same number might improve the sources of variance in which the items were involved. Second, adding a fourth stage of assessment – which is possible in a subject lasting six months – would improve the reliability, as the study by Van Moere (2006) showed.

Limitations and further research

First, the definition of assessment as a local practice (Knight 2006) means that we cannot generalize the conclusions of this investigation outside the context of university arts studies. The intervention was aimed at novice students and this feature had a strong effect on the design of the rubric, as did the fact that service-learning projects were a novelty in the educational context.

Second, the isolated effect of the novice and expert assessors on the rater bias was not investigated. This issue is of major importance since other studies have detected notable differences depending on the type of rater (e.g., Leung et al. 2012; Praetorius et al. 2012). We

need to establish whether the variance relative to the teachers was notably lower, because if so this would sanction the reduction in the number of the raters without negatively affecting the acceptable reliability coefficients.

Third, because of the clear differences between the two types of service-learning projects we were obliged to perform separate G studies. While this allows a more exact evaluation of other sources of variance, the intensity of the type of project effect is not taken into account. The tasks or cases were identified as main variance components (e.g., Guiton et al. 2004; Iramaneerat et al. 2008), and so their analysis is decisive for affirming the robustness of the rubrics-based performance system.

Conclusion

This study adds to the small body of studies of rubrics-based performance assessment approached from the perspective of the Generalizability theory. The results showed an acceptable true score variance and also showed that the influence of the items was slightly higher than that of the raters, without representing important sources of error in either case. We stress the need to convert rubrics into regulatory elements so that, in each educational context, students and teachers can construct professionally relevant consensuses about complex performances. The data indicated that undergraduate students perform reliable assessments when they have had sustained contact with the rubrics in the learning environments and when the assessment procedures are accompanied by moderation sessions in which the characteristics of the objects under evaluation are discussed.

Funding

This research was supported by the Spanish Ministry of Economy and Competitiveness and grants European Regional Development Fund [HAR2013-46608-R]; the Institute of Education Sciences [REDICE16-1420]; the Vice-rectorate for Teaching Policy and the Programme for Teaching Innovation at the University of Barcelona [GIDC-ODAS].

Notes on contributors

José-Luis Menéndez-Varela is a professor in the Department of Art History at University of Barcelona and director of the Observatory for Education in the Arts. His research interests include learning assessment, curriculum assessment, peer learning and informal learning environments.

Eva Gregori-Giralt is a professor in the Department of Art History at University of Barcelona. Her research interests include the development of learning resources and assessment specific to the arts, the use of information and communication technology, and informal learning environments.

ORCID

José-Luis Menéndez-Varela http://www.orcid.org/0000-0002-6733-9346

Eva Gregori-Giralt http://www.orcid.org/0000-0002-9774-1563

References

Baryla, E., G. Shelley, and W. Trainor. 2012. "Transforming rubrics using factor analysis." *Practical Assessment, Research and Evaluation* 17 (4): 1–7.

Bird, F., and R. Yucel. 2013. "Improving marking reliability of scientific writing with the Developing Understanding of Assessment for Learning programme." *Assessment and Evaluation in Higher Education* 38 (5): 536–553.

Brennan, R. 1996. "Generalizability of performance assessments." In *Technical issues in large-scale performance assessment*, edited by G. Phillips, 19–58. Washington: National Center for Education Statistics.

Cook, D., T. Beckman, J. Mandrekar, and V. Pankratz. 2010. "Internal structure of mini-CEX scores for internal medicine residents: factor analysis and generalizability." *Advances in Health Sciences Education* 15: 633–645.

Cronbach, L., G. Gleser, H. Nanda, and N. Rajaratnam. 1972. T*he dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Crotwell Timmerman, B., D. Strickland, R. Johnson, and J. Payne. 2011. "Development of a 'universal' rubric for assessing undergraduates' scientific reasoning skills using scientific writing." *Assessment and Evaluation in Higher Education* 36 (5): 509–547.

De Lima, A., D. Conde, J. Costabel, J. Corso, and C. Van der Vleuten. 2013. "A laboratory study on the reliability estimations of the mini-CEX." *Advances in Health Sciences Education* 18: 5–13.

Dornan, T., A. Muijtjens, J. Graham, A. Scherpbier, and H. Boshuizen. 2012. "Manchester Clinical Placement Index (MCPI). Conditions for medical students' learning in hospital and community placements." *Advances in Health Sciences Education* 17: 703–716.

Feldt, L., and R. Brennan. 1989. "Reliability." In *Educational measurement*, edited by R. Linn, 105–146. New York: Macmillan.

Gebril, A. 2009. "Score generalizability of academic writing tasks: Does one test method fit it all?" *Language Testing* 26 (4): 507–531.

Gugiu, M., P. Gugiu, and R. Baldus. 2012. "Utilizing Generalizability Theory to Investigate the Reliability of the Grades Assigned to Undergraduate Research Papers." *Journal of MultiDisciplinary Evaluation* 8 (19): 26–40.

Guiton, G., C. Hodgson, G. Delandshere, and L. Wilkerson. 2004. "Communication Skills in Standardized-Patient Assessment of Final-Year Medical Students: A Psychometric Study." *Advances in Health Sciences Education* 9: 179–187.

Heijne-Penninga, M., J. Kuks, J. Schönrock-Adema, T. Snijders, and J. Cohen-Schotanus. 2008. "Open-book tests to complement assessment programmes: analysis of open and closed-book tests." *Advances in Health Sciences Education* 13: 263–273.

Howell, R. 2014. "Grading rubrics: hoopla or help?" *Innovations in Education and Teaching International* 51 (4): 400–410.

Iramaneerat, C., R. Yudkowsky, C. Myford, and S. Downing. 2008. "Quality control of an OSCE using generalizability theory and many-faceted Rasch measurement." *Advances in Health Sciences Education* 13 (4): 479–493.

Jonsson, A. 2014. "Rubrics as a way of providing transparency in assessment." *Assessment and Evaluation in Higher Education* 39 (7): 840–852.

Jonsson, A., and G. Svingby. 2007. "The use of scoring rubrics: Reliability, validity and educational consequences." *Educational Research Review* 2: 130–144.

Knight, P. 2006. "The local practices of assessment." *Assessment and Evaluation in Higher Education* 31 (4): 435–452.

Kreiter, C., P. Yin, C. Solow, and R. Brennan. 2004. "Investigating the Reliability of the Medical School Admission Interview." *Advances in Health Sciences Education* 9: 147–159.

Leung, K.-K., W.-D. Wang, and Y.-Y. Chen. 2012. "Multi-source evaluation of interpersonal and communication skills of family medicine residents." *Advances in Health Sciences Education* 17: 717–726.

Messick, S. 1989. "Validity". In *Educational measurement*, edited by R. Linn, 13–103. New York: Macmillan.

Montgomery, K. 2002. "Authentic Tasks and Rubrics: Going beyond Traditional Assessments in College Teaching." *College Teaching* 50 (1): 34–39.

Oakleaf, M. 2008. "Dangers and Opportunities: A Conceptual Map of Information Literacy Assessment Approaches." *Portal: Libraries and the Academy* 8 (3): 233–253.

O'Malley, G. 2010. "Designing induction as professional learning community." T*he Educational Forum* 74 (4): 318–327.

Pinot de Moira, A., C. Massey, J. Baird, and M. Morrissy. 2002. "Marking consistency over time." *Research in Education* 67 (1): 79–87.

Praetorius, A.-K., G. Lenske, and A. Helmke. 2012. "Observer ratings of instructional quality: Do they fulfill what they promise?" *Learning and Instruction* 22: 387–400.

Raymond, M., P. Harik, and B. Clauser. 2011. "The Impact of Statistically Adjusting for Rater Effects on Conditional Standard Errors of Performance Ratings." *Applied Psychological Measurement* 35 (3): 235–246.

Reddy, Y., and H. Andrade. 2010. "A Review of Rubric Use in Higher Education." *Assessment and Evaluation in Higher Education* 35 (4): 435–448.

Reynolds, J., R. Smith, C. Moskovitz, and A. Sayle. 2009. "BioTAP: A Systematic Approach to Teaching Scientific Writing and Evaluating Undergraduate Theses." *BioScience* 59 (10): 896–903.

Sadler, D. 2009. "Indeterminacy in the use of preset criteria for assessment and grading." *Assessment and Evaluation in Higher Education* 34 (2): 159–179.

Shavelson, R., and N. Webb. 1991. *Generalizability Theory: A First.* Newbury Park: Sage.

Schlitz, S. et al. 2009. "Developing a Culture of Assessment through a Faculty Learning Community: A Case Study." *International Journal of Teaching and Learning in Higher Education* 21 (1): 133–147.

Stuhlmann, J., C. Daniel, A. Dellinger, R. Kenton, and T. Powers. 1999. "A generalizability study of the effects of training on teachers' abilities to rate children's writing using a rubric." *Journal of Reading Psychology* 20 (2): 107–127.

Sudweeks, R., S. Reeve, and W. Bradshaw. 2005. "A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing." *Assessing Writing* 9: 239–261.

Swanson, D., G. Norman, and R. Linn. 1995. "Performance-based assessment: Lessons from the health professions." *Educational Researcher* 24 (5): 5–11.

Tan, K., and M. Prosser. 2004. "Qualitatively different ways of differentiating student achievement: a phenomenographic study of academics' conceptions of grade descriptors." *Assessment and Evaluation in Higher Education* 29 (3): 267–282.

Taub, D., H. Servaty-Seib, C. Wachter Morris, S. Prieto-Welch, and D. Werden. 2011. "Developing Skills in Providing Outreach Programs: Construction and Use of the POSE

(Performance of Outreach Skills Evaluation) Rubric." *Counseling Outcome Research and Evaluation* 2 (1): 59–72.

Tucker, R. 2013. "The architecture of peer assessment: do academically successful students make good teammates in design assignments?" *Assessment and Evaluation in Higher Education* 38 (1): 74–84.

Van Moere, A. 2006. "Validity evidence in a university group oral test." *Language Testing* 23 (4): 411–440.

Vardi, I. 2013. "Effectively feeding forward from one written assessment task to the next." *Assessment and Evaluation in Higher Education* 38 (5): 599–610.

Venning, J., and F. Buisman-Pijlman. 2013. "Integrating assessment matrices in feedback loops to promote research skill development in postgraduate research projects." *Assessment and Evaluation in Higher Education* 38 (5): 567–579.

Vispoel, W., and S. Tao. 2013. "A Generalizability Analysis of Score Consistency for the Balanced Inventory of Desirable Responding." *Psychological Assessment* 25 (1): 94–104.

Ward, H., and P. Selvester. 2012. "Faculty learning communities: improving teaching in higher education." *Educational Studies* 38 (1): 111–121.

Wolf, K., and L. Goodwin. 2007. "Evaluating and Enhancing Outcomes Assessment Quality in Higher Education Programs." *Metropolitan Universities* 18 (2): 42–56.

Wolf, K., and E. Stevens. 2007. "The Role of Rubrics in Advancing and Assessing Student Learning." *The Journal of Effective Teaching* 7 (1): 3–14.

Zhang, B., L. Johnston, and G. Kilic. 2008. "Assessing the reliability of self- and peer rating in student group work." *Assessment and Evaluation in Higher Education* 33 (3): 329–340.