Rubrics for developing students' professional judgement: A study of sustainable assessment in arts education

José-Luis Menéndez-Varela and Eva Gregori-Giralt

Universitat de Barcelona

José-Luis Menéndez-Varela[a]

menendez@ub.edu

ORCID: 0000-0002-6733-9346

Eva Gregori-Giralt[a]

gregori@ub.edu

ORCID: 0000-0002-9774-1563

[a]Address: University of Barcelona. Faculty of Fine Arts, Department of History of Art, Adolf Florensa, s/n. 08028 Barcelona (Spain). Telephone number: (34) 934 037 589. Fax: (34) 934 035 976.

Correspondence concerning this article should be addressed to Eva Gregori-Giralt, Department of History of Art, Faculty of Fine Arts, University of Barcelona, Adolf Florensa s/n, 08028 Barcelona. Telephone number: (34) 934 037 589. Fax: (34) 934 035 976. E-mail: gregori@ub.edu

Abstract

The participation of students in assessment is known to generate higher-order learning outcomes. This study aims to determine the usefulness of rubrics in aiding the incorporation of undergraduate students into assessor teams for developing their professional judgement. A quasi-experimental study examined the effects of a brief training programme on the use of rubrics, and of the participation of students in rubric creation and moderation discussions. We calculated Cronbach's alphas, and intraclass correlation coefficients in order to examine the intra- and inter-rater reliability between all the members of the assessor teams. The results demonstrate that only participation in the rubric design and in the moderation discussions regulating their use helped undergraduate students to develop sound assessment skills. We infer that rubrics can help to promote professional judgement if they are conceived as instructional resources for defining and supporting the processes of negotiation and agreement that characterize an assessment culture.

Keywords: student evaluation; evaluation utilization; evaluation methods; rubric; professional judgement; reliability.

Introduction

With their criticisms of instrumental rationality and of the separation inside disciplines between theory and practice, the theories of experiential learning (see Kolb, 1984; Schön, 1983) and the theories of situated learning (see Lave, 1988; Lave & Wenger, 1991) drew attention to the educational prejudices that this division reflects and its incompatibility with the knowledge society. These authors also claimed that the formation of experts is not limited to epistemological issues, but involves a complex process of integration into a professional culture. Saying that experts "know" their discipline means that they are familiar with its

paradigms, methodologies and objects of study, with the social organization of work and the circuits and agents involved in the decision-making process, and with the everyday representations and practices they share with their colleagues.

This shift in the understanding of the discipline, from its conception as a mere epistemological construct to its acknowledgement as a professional culture, and the shift in the definition of the "expert", are directly linked to the recognition of tacit knowledge, originally defined by Polanyi (1958, 1962). It is disputed whether tacit knowledge is a type of independent knowledge, or rather a dimension of knowledge which can be progressively formalized until it is transformed into explicit knowledge (see Duguid, 2005; Klein, 2008; Tsoukas, 2002). The possibility of a relatively unarticulated and unconceptualized form of knowledge determines an idea of the expert's judgement as an act of connoisseurship which is founded both in metacognitive processes and in socially-situated interpretative practices.

Not even the approaches that explain metacognitive development as a progression from a tacit model to a formal model (e.g., Schraw and Moshman, 1995) deny the coexistence of mental processes characteristic of these models, or the predominance of one or the other depending on the circumstances. In contrast to Schraw and Moshman (1995), the models of Ecclestone (2001) and Suto and Greatorex (2008) regarding raters' cognitive operations in marking processes coincided in identifying a shift from mental processes characteristic of explicit knowledge to others more representative of tacit knowledge as the rater progressively gains experience. But these authors also acknowledged that both types of mental processes are active at all times and their use (by the expert as well) depends on their familiarity with the type of object being evaluated more than on its complexity.

Sustainable assessment for the development of professional judgement in students

The formation of experts requires the practice of judgement in real professional contexts or in teaching environments which simulate them. This training activity is centred on

metacognition in order to develop autonomous learning skills. As part of the assessment for learning approach, authentic performance assessment points towards a socially and culturally located practice of assessment (Watson & Robbins, 2008); for its part, sustainable assessment aims to develop students' professional judgement and their capacity to participate in the discipline in an increasingly pertinent manner. Since the sustainable assessment theory adds the requirement that the assessment should promote lifelong learning (Boud, 2000; Boud & Falchikov, 2005; Singh & Terry, 2008), it can thus be considered as a form of assessment *as* learning. For a mere idea or an opinion to be transformed into a professional judgement, students must participate in the assessment by becoming integrated in real contexts, in which they can share their responsibility and discuss the assessment criteria and procedures with other more experienced raters (Beck, Skinner, & Schwabrow, 2013).

In this way, we can establish a connection between the sustainable assessment approach and the conception of universities as learning communities (e.g. Carroll, 2005; O'Malley, 2010; Ward & Selvester, 2012). Brown, Collins, and Duguid (1989) affirmed that authentic learning is an apprenticeship to a community of practice, taking 'professional culture' and 'community of practice' to be synonyms. Nobody today disputes the fact that communities of practice enhance learning and professional development: the distinctive trait of these informal and heterogeneous work groups is that they are organized in accordance with the principles of trust, support and collegiality and thus promote a free dialogue and collaborative review of professional practices. This approach assigns considerable importance to the processes of the social construction of knowledge and of the construction of identity which operate, together with the cognitive dimension, in learning (e. g., Kobayashi et al., 2017; Kwon & Park, 2017). Nonetheless, the concept of *community of practice* in the context of formal education has been challenged, and some authors have stressed the need to

talk in terms of *communities of practice of learners* or – in vocational education – *quasi-communities of practice* (see Emad & Roth, 2016).

In the specific domain of assessment and evaluation, Bird and Yucel (2013), Brooks (2012), and Price (2005) noted the benefits of teachers' teams that work as communities of practice in the generation of shared assessment constructs and in the consistent application of the assessment procedures. The literature has focused fundamentally on faculty learning communities, but some studies also consider the participation of students. For example, Kearney (2013) tested two models of authentic assessment for sustainable learning grounded in legitimate peripheral participation in approximately 300 undergraduate education students. In this process, learners are inducted into communities of practice, take on more complex responsibilities, and occupy more important positions as their professional development progresses.

The concept of *communities of practice (of learners)* is relevant here because the sustainable assessment theory removes the idea of assessment as a learning activity and places the reflective practice of assessment at the centre of the teaching environment. Equipping the students with the skills to assess each situation, to identify their own learning needs and, in addition, to do this independently and responsibly, is not something that can be achieved by developing certain learning activities in isolation. Although the specialized literature has shown the positive influence of establishing self-and peer-assessment activities – focusing on professional situations, strengthening feedback and feedforward, and promoting student discussion and negotiation of the assessment system itself – it should not be forgotten that the sustainable assessment approach is a model of curriculum design.

Boud and Soler (2016) affirmed that assessment is always relational and there are no intrinsic qualities in the learning activity or assessment method that guarantee the attainment of the required learning outcomes. The need for a holistic perspective focused on assessment

makes it possible to talk about teaching strategies because the aim is to create an educational environment that favours metacognition, self-regulation and the social and professional evaluative skills on which the concept of sustainable assessment is based. If this theory places the development of informed judgment by students among the main educational goals (Boud & Falchikov, 2007), then pedagogy appears as a set of planned actions – strategies – whose coherence depends on the dynamism and flexibility that allow mutual recognition, the sharing of responsibilities, and collaboration among all the participating agents.

Rubrics as a resource for constructing professional judgement

The participation of students in marking constitutes an important way to learn a professional culture. Marking is consistent with sustainable assessment when it is used not as an aim in itself but a means for the development of higher-order evaluative processes. The question is whether the incorporation of students into assessors' teams should be promoted. The challenge is to make the expert's judgement comprehensible to students, and in this regard rubrics may be a particularly useful resource.

Researchers have insisted that assessment should include useful information which both helps students to understand the learning outcomes established and determines the extent to which they have achieved these outcomes, and put learning strategies into practice in order to enhance their performance (e.g., Lipnevich et al., 2014). There is sufficient evidence in the literature that prior knowledge of the criteria, performance levels and quality definitions for each of the assessment activities contributes to student learning. However, we should be careful not to see rubrics as an educational panacea. Three review articles on the subject coincide in calling for more methodologically sound research before a direct relationship between the use of rubrics and academic outcomes can be claimed (Jonsson & Svingby, 2007; Panadero & Jonsson, 2013; Reddy & Andrade, 2010). Asserting that rubrics have a positive effect on student learning is not the same as identifying their specific effect

compared with other teaching resources applied simultaneously, or assessing the quality of the learning outcomes that they provide for.

The first key point regarding rubrics resides in the difficulty of assessing complex phenomena. Sadler (2009) noted the existence of structural inadequacies in the methods of analytic grading – the problem of indeterminacy, in the author's own words – which invalidate them for the assessment of complex student works. Brooks (2012) stressed that the understanding of the criteria and the application of the assessment procedures require a complex process of familiarization which takes place inside the communities of assessors; this process generates specific adaptations of the assessment practices involved in the shift from a criterion-referenced assessment to a construct-referenced assessment.

Even if rubrics generate clear and transparent criteria, this clarity and transparency are not sufficient to guarantee a consistent application of the criteria, because these criteria are subject to social processes of the construction of meaning (Knight, 2006) and because rubrics do not include the implicit processes of judgement that are also included in assessment (Moskal and Leydens, 2001). This explains why students do not understand learning expectations or teachers' evaluations of their learning outcomes (O'Donovan, Price, & Rust, 2004) and why there are doubts about the effectiveness of students' use of rubrics (Jonsson & Svingby, 2007; Venning & Buisman-Pijlman, 2013), even when they are aware of the assessment criteria. Even studies of students' perceptions of rubrics present limitations: the fact that students value receiving information on assessment criteria sheds little light on how they interpret this information, or whether their interpretation benefits their learning achievements (El-Mowafy, 2014; Surgenor, 2013).

The second key point is the temptation of simplifying the phenomenon being assessed in order to to facilitate a shared and consistent use of rubrics by many assessors. Examples of bad practices are limiting rubrics to easily observable aspects of the object under assessment,

or limiting the assessment to a mechanical identification of the components specified in the rubrics. In the first case, the risk is that students may concentrate on trivial aspects (Bell, Mladenovic, & Price, 2013), which impedes their understanding of the object being assessed and of the assessment process, and distracts them from their attempts to achieve valid professional judgement. In the second case, the risk is that student autonomy and independent learning will be reduced (Boud & Falchikov, 2006; Nordrum, Evans, & Gustafsson, 2013; Norton, 2004), thus preventing the development of sustainable assessment skills.

Nonetheless, rubrics can help students to cope with the challenges facing them in their attempts to achieve expert judgement. When used as learning resources, rubrics highlight features of the assessment objects and procedures which merit discussion and function as reference points that can help guide the debate. In this way, they help students to identify the points of agreement inside a professional culture, to become aware of how these points of agreement are created and managed, and to understand how judgements operate. As Boud and Falchikov (2006) stressed, rubrics must be an instrument placed at the service of the students in order to develop their discernment.

In the literature, it is acknowledged that rubrics contribute to the creation of a shared reflective practice among professional assessors through processes of collaborative design, specific training programmes for their use, agreed practices for augmentation of ratings, and the implementation of post-marking moderation systems to reduce discrepancies. The emphasis placed on the importance of discussing rubrics with the students (Oakleaf, 2008; Sadler, 2009) on a regular basis (Beck, Skinner, & Schwabrow, 2013) demonstrates the recent tendency to consider them as instructional resources as well. There is a considerable body of research on the participation of students in rubric design (Fraile, Panadero, & Pardo, 2017; Hughes & Cappa, 2007; Wolf & Stevens, 2007), on training programmes directed at students (Crotwell Timmerman et al., 2011; Smith et al., 2013; Welch and James, 2007), on

the use of exemplars of the different performance levels (Bell, Mladenovic, & Price, 2013; Bird & Yucel, 2013; Cevik & Andre, 2014; Kyun, Kalyuga, & Sweller, 2013; Vardi, 2013), and on the participation of students in the moderation discussions (Bird & Yucel, 2013; Crotwell Timmerman et al., 2011; [deleted for peer review], 2016, 2017). So in educational research it is becoming clear that rubrics cannot be considered in isolation because in fact they are integral parts of the teaching and learning context (Howell, 2014).

Study aims

The aims of this study are to evaluate the utility of rubrics for incorporating undergraduate students in assessors' teams in order to develop their professional judgement, and to assess how far the use of rubrics needs to be accompanied by other instructional resources. The study started from the following premises: (a) the undergraduate students who acted as raters (henceforth, student assessors) would evaluate complex student works – in this case, the oral presentations of service-learning projects designed by first-year undergraduate students – and (b) that the student assessors had not performed service-learning projects during their undergraduate studies. These premises ensured that participants would have to activate higher-order evaluative processes, and controlled for the possible distortions caused by prior knowledge of the object under evaluation.

Two outcomes were studied: (a) how the student assessors used the rubric after attending a training programme, and (b) how they used the rubric after joining the assessors' team alongside the teachers; that is, after participating in the redesign of the rubric, and attending moderation discussions after each assessment test. To check whether the student assessors developed consistent evaluative skills, their assessments were compared with the expert judgement of the teachers via an analysis of inter-rater reliability. The hypothesis was that the training programme would have a positive effect, but that the best results would be

obtained with the incorporation of the students in a team of raters which would allow them to construct a deeper learning of the assessment criteria and procedures.

If this hypothesis is confirmed, this would mean that a study which started as an assessment of reliability also provides evidence of the validity of the instructional settings, in so far as the results would suggest that the ability of the assessor students to assess complex work has improved.

Method

Participants

The study sample comprised 600 students studying for the History of Art, Fine Arts, Conservation-Restoration and Design undergraduate degrees at a large public [deleted for peer review] university. Inside the framework of three six-month compulsory subjects, the students designed service-learning projects. One hundred and twenty work groups were created, comprising five students each, and remained unchanged throughout the semester. The mean age of the students was 21 years (SD = 8.53; mode = 19 and median = 19); 78.33% were female, and 74.52% were not in employment.

The projects were evaluated by two teachers and 24 undergraduate students. In this case, the mean age of the students was 22 years (SD = 2.14; mode = 21 and median = 19); 75% were female, and 71.33% were not in employment.

Study design

An only-post quasi-experimental study with several groups and simultaneous quasi-control was performed to examine the results of the undergraduate students' assessment competency. Before the start of the academic year, the two teachers designed a system comprising one rubric to assess the planning of service-learning projects and their oral presentation by the first-year undergraduate students. This rubric-based performance assessment was developed in three stages over the course of each semester.

The intervention took place over three semesters and involved a total of six experimental groups and six control groups, two of each type in each semester. The experimental groups and control groups comprised two teachers (always the same) and two students, who were different in each team of assessors. Each student assessor evaluated 20 projects along with his/her companion and the two teachers.

In the control groups, the student assessors assessed the projects, together with the two teachers, after attending a training programme on the use of rubrics. In the experimental groups, the two student assessors assessed the same projects, also together with the two teachers, after participating in the redesign of the rubric and in the moderation discussions which followed each of the three stages of the assessment.

Instructional settings

The undergraduate students designed and planned service-learning projects directed at a group at risk of social exclusion. The projects comprised three stages. Weekly meetings were scheduled between each work group and the teachers. This guarantee of follow-up and continued feedback over the course of the semester helped students to reflect on the results obtained at each stage and thus to improve their projects.

The two teachers had designed a rubric for the formative and summative assessment, and before the intervention, the rubric was redesigned with the aid of the 12 student assessors who participated in the experimental groups. The rubric was used to help students understand the learning goals and provided them with useful information for the development of their projects; how to focus the work sessions giving in-task guidance to the students; and to carry out a continuous assessment in which all the participants scored the projects and their oral presentation individually.

The analytic rubric was used in the three stages of assessment of the projects. The assessment criteria always included four performance levels, and the quality definitions were

descriptive in order to facilitate their use by the students. As for the structure of the rubric, the assessment criteria were organized in two broad blocks focusing on the project content and the oral presentation. The first block covered the project's content: the *Project rationale*, that is, the title, objectives, and study of the group at risk of social exclusion, and of the context of the application of the project (present in the three assessment 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). The second block included the principal aspects of oral presentation skills such as *Appropriateness to audience*, *Delivery*, *Grammar and language*, *Eye contact* and *Time management* (the last criterion appeared only in the third stage).

Procedure

The assessment of the projects was planned at the beginning of the two academic years. The work groups of undergraduate students were allowed 20 minutes for the first assessment stage, 30 for the second and 45 for the third.

The student assessors and the teachers attended the oral presentations with the assessment sheet designed for each stage which were stored in the subject's course management system. To avoid discrepancies due to environmental conditions, all the assessors sat together and in the same place during the three semesters. The assessment protocol required each assessor to take notes on the aspects he/she considered relevant in relation to each criterion included in the rubric, and also to take notes on any incident that disturbed its use of the rubric. Finally, each assessor gave the students' work a numerical score.

The assessors scored each criterion of the rubrics using a fixed scale of measurement from 0 to 10, the habitual scale used in the [deleted for peer review] educational system: 0-2

for the lowest performance level, 3-4 for the second, 5-7 for the third and 8-10 for the highest. The study data included the 105,600 scores given by the raters to all 600 students on all the assessment criteria which constituted the three stages of assessment.

The training programme was given by the teachers to the student assessors of the control groups before the start of the assessment process. The programme lasted six hours over two 3-hour sessions, and discussed: (a) what a service-learning project is, (b) what distinguishes it from other teaching environments that also work with projects, or from volunteer programmes and social work programmes, (c) the fundamental structure in which a project of this kind should be organized, (d) the commonest problems associated with its design and planning, (e) the main aspects of an oral presentation, (f) the management of the assessment of a presentation by different speakers, and (g) how all these aspects are included in the rubric and how it should be used.

The work of the student assessors of the experimental groups began with a redesign of the rubric. The work flow was as follows. In the first stage, the two teachers and the student assessors worked individually on improving the initial rubric; these students studied the rubric and recorded the doubts presented by its use. Their contribution was important because it helped the team to understand the problems that a student who had never used them might encounter. In the second stage, separate meetings were held with the student assessors, on the one hand, and the two teachers, on the other, in order to clarify the proposals for modification. These meetings were held separately in order to preserve the point of view of the student assessors. In the third stage, a meeting was held involving the entire team in order to discuss the different proposals and to agree on the modifications to be made to the rubric.

The work of the experimental groups also included moderation discussions involving all the assessors which were held immediately after each assessment session. In these discussions, participants described their impressions of the service-learning 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. Subsequently, the assessors introduced their scores in the application created for this purpose in the course management system.

Research questions

The research questions posed in this study were the following:

(1) Was the assessors's behaviour in this rubric-based performance assessment consistent?

(2) What aspects influenced the development of the assessment skills used by the student assessors?

Data analysis

The following statistical studies were performed in each subject. First, an analysis of variance (ANOVA) was conducted with the assessors' final scores at each stage of the assessment to determine whether there were significant differences. Second, Cronbach's Alphas were calculated to establish the consistency of the results after repeating the assessment with the same assessment system (the rubric) and in similar conditions (the stages of the projects). Third, the intraclass correlation coefficients were obtained to study the equivalence of the assessors' judgements in each assessment criterion.

Results

Table 1 presents the results of the analysis of variance (ANOVA) calculated from the set of the scores of all the raters who participated in each subject and semester.

Table 1

Joint ANOVA according to stage and subject

Semester	Subject	Stage	Group	Study	Sum of squares	Df.	Mean square	F[a]	Sig. (p<.05)
1	Concepts of Modern Art A	1	Experimental	Between	7.84	3.00	2.61	1.47	0.22
				Within	705.13	396.00	1.78		

Semester	Subject	Stage	Group	Study	Sum of squares	Df.	Mean square	F[a]	Sig. (p<.05)
				Total	712.97	399.00			
			Control	Between	39.83	3.00	13.27	7.66	0.00
				Within	685.95	396.00	1.73		
				Total	725.78	399.00			
		2	Experimental	Between	3.54	3.00	1.18	0.87	0.45
				Within	533.72	396.00	1.35		
				Total	537.26	399.00			
			Control	Between	9.06	3.00	3.02	2.67	0.05
				Within	448.54	396.00	1.13		
				Total	457.60	399.00			
		3	Experimental	Between	3.50	3.00	1.17	1.03	0.38
				Within	446.16	396.00	1.13		
				Total	449.66	399.00			
			Control	Between	9.12	3.00	3.04	2.86	0.04
				Within	420.07	396.00	1.06		
				Total	429.19	399.00			
	Concepts of Modern Art B	1	Experimental	Between	13.29	3.00	4.43	1.15	0.33
				Within	1518.36	396.00	3.83		
				Total	1531.65	399.00			
			Control	Between	12.69	3.00	4.23	2.71	0.04
				Within	618.53	396.00	1.56		
				Total	631.22	399.00			
		2	Experimental	Between	13.14	3.00	4.38	1.17	0.32
				Within	1486.66	396.00	3.75		
				Total	1499.80	399.00			
			Control	Between	17.46	3.00	5.82	4.73	0.00
				Within	487.54	396.00	1.23		
				Total	505.00	399.00			
		3	Experimental	Between	2.15	3.00	0.72	0.55	0.65
				Within	517.95	396.00	1.31		
				Total	520.10	399.00			
			Control	Between	9.32	3.00	3.11	2.70	0.05
				Within	455.92	396.00	1.15		
				Total	465.24	399.00			

Semester	Subject	Stage	Group	Study	Sum of squares	Df.	Mean square	F[a]	Sig. (p<.05)
2	Theory of Art	1	Experimental	Between	6.53	3.00	2.18	2.56	0.05
				Within	336.88	396.00	0.85		
				Total	343.41	399.00			
			Control	Between	16.55	3.00	5.52	6.57	0.00
				Within	333.37	396.00	0.84		
				Total	349.92	399.00			
		2	Experimental	Between	3.01	3.00	1.00	0.71	0.54
				Within	558.67	396.00	1.41		
				Total	561.68	399.00			
			Control	Between	7.59	3.00	2.53	2.97	0.03
				Within	336.88	396.00	0.85		
				Total	344.47	399.00			
		3	Experimental	Between	1.39	3.00	0.46	0.35	0.79
				Within	518.89	396.00	1.31		
				Total	520.28	399.00			
			Control	Between	8.43	3.00	2.81	2.75	0.04
				Within	403.97	396.00	1.02		
				Total	412.40	399.00			
	Foundations of the History of Aesthetics	1	Experimental	Between	6.75	3.00	2.25	1.29	0.30
				Within	691.11	396.00	1.75		
				Total	697.86	399.00			
			Control	Between	24.94	3.00	8.31	3.13	0.03
				Within	1050.98	396.00	2.65		
				Total	1075.93	399.00			
		2	Experimental	Between	3.03	3.00	1.01	0.63	0.60
				Within	635.42	396.00	1.60		
				Total	638.45	399.00			
			Control	Between	20.44	3.00	6.81	4.10	0.01
				Within	657.91	396.00	1.66		
				Total	678.35	399.00			
		3	Experimental	Between	2.33	3.00	0.78	0.68	0.57
				Within	454.34	396.00	1.15		
				Total	456.67	399.00			
			Control	Between	11.16	3.00	3.72	3.27	0.02

Semester	Subject	Stage	Group	Study	Sum of squares	Df.	Mean square	F[a]	Sig. (p<.05)
				Within	450.63	396.00	1.14		
				Total	461.79	399.00			
3	Concepts of Modern Art	1	Experimental	Between	1.88	3.00	0.63	0.33	0.80
				Within	744.53	396.00	1.88		
				Total	746.41	399.00			
			Control	Between	10.49	3.00	3.50	2.61	0.05
				Within	530.64	396.00	1.34		
				Total	541.13	399.00			
		2	Experimental	Between	1.86	3.00	0.62	0.39	0.76
				Within	624.55	396.00	1.58		
				Total	626.41	399.00			
			Control	Between	7.83	3.00	2.61	2.86	0.04
				Within	361.78	396.00	0.91		
				Total	369.61	399.00			
		3	Experimental	Between	1.82	3.00	0.61	0.41	0.75
				Within	590.70	396.00	1.49		
				Total	592.52	399.00			
			Control	Between	20.41	3.00	6.80	5.70	0.00
				Within	473.20	396.00	1.19		
				Total	493.61	399.00			
	Theory of Art	1	Experimental	Between	0.88	3.00	0.29	0.17	0.92
				Within	670.31	396.00	1.69		
				Total	671.19	399.00			
			Control	Between	8.77	3.00	2.92	2.68	0.04
				Within	430.42	396.00	1.09		
				Total	439.19	399.00			
		2	Experimental	Between	1.00	3.00	0.33	0.25	0.86
				Within	531.69	396.00	1.34		
				Total	532.69	399.00			
			Control	Between	7.33	3.00	2.44	2.77	0.04
				Within	348.16	396.00	0.88		
				Total	355.49	399.00			
		3	Experimental	Between	1.35	3.00	0.45	0.32	0.81

Semester	Subject	Stage	Group	Study	Sum of squares	Df.	Mean square	F[a]	Sig. (p<.05)
				Within	550.76	396.00	1.39		
				Total	552.11	399.00			
			Control	Between	35.85	3.00	11.95	12.58	0.00
				Within	377.33	396.00	0.95		
				Total	413.18	399.00			

[a]Maximum value permitted by chance = 2.6

In all the control groups, statistically significant differences were detected between the assessors in the three stages of assessment, demonstrating the existence of important discrepancies in the use of the rubric. In the experimental groups, however, these differences were no longer present in any of the stages; all the raters scored in a similar way, indicating a shared use of the assessment system.

To study the possible causes of the discrepancies, we assessed the detailed scores obtained by students in each stage and calculated the Cronbach's α and intraclass correlation coefficients (ICC) between raters and for each one of the assessment criteria (Table 2).

Table 2

Cronbach's α and intraclass correlations coefficients according to stage and subject

Semester	Subject	Stage	Group	Study	PR	DO	PW	A	AA	DE	GaL	EC	TM
1	Concepts of	1	Experimental	*a*	.93	.88			.85	.86	.85	.87	
	Modern Art			ICC	.77	.65			.59	.61	.59	.62	
	A		Control	*a*	.62	.87			.85	.80	.76	.85	
				ICC	.29	.63			.59	.50	.44	.58	
		2	Experimental	*a*	.93	.90	.89		.87	.87	.80	.87	
				ICC	.78	.69	.67		.62	.63	.51	.63	
			Control	*a*	.54	.74	.86		.70	.72	.76	.87	
				ICC	.23	.42	.60		.36	.40	.44	.62	
		3	Experimental	*a*	.93	.90	.87	.91	.90	.90	.90	.90	.91
				ICC	.78	.69	.64	.71	.69	.69	.70	.69	.72

Semester	Subject	Stage	Group	Study	PR	DO	PW	A	AA	DE	GaL	EC	TM
			Control	*a*	.67	.55	.90	.91	.77	.75	.76	.86	.90
				ICC	.34	.23	.70	.71	.46	.43	.44	.61	.70
	Concepts of	1	Experimental	*a*	.92	.85			.85	.85	.85	.85	
	Modern Art B			ICC	.73	.59			.58	.58	.59	.58	
			Control	*a*	.90	.87			.85	.78	.89	.86	
				ICC	.70	.63			.58	.47	.66	.61	
		2	Experimental	*a*	.92	.89	.91		.90	.90	.90	.91	
				ICC	.73	.66	.72		.70	.68	.70	.71	
			Control	*a*	.91	.56	.86		.78	.75	.74	.86	
				ICC	.72	.24	.60		.47	.42	.41	.61	
		3	Experimental	*a*	.92	.90	.90	.90	.91	.91	.90	.91	.90
				ICC	.73	.70	.69	.70	.72	.72	.70	.72	.70
			Control	*a*	.54	.91	.90	.90	.78	.75	.78	.90	.90
				ICC	.23	.71	.68	.68	.47	.43	.46	.69	.69
2	Theory of Art	1	Experimental	*a*	.86	.83			.89	.86	.87	.85	
				ICC	.60	.55			.67	.60	.62	.59	
			Control	*a*	.82	.62			.85	.86	.90	.84	
				ICC	.53	.29			.59	.61	.70	.57	
		2	Experimental	*a*	.79	.87	.90		.90	.91	.90	.89	
				ICC	.48	.63	.71		.70	.71	.70	.68	
			Control	*a*	.84	.86	.80		.76	.72	.75	.88	
				ICC	.57	.60	.50		.44	.39	.43	.65	
		3	Experimental	*a*	.90	.88	.90	.90	.90	.90	.90	.90	.90
				ICC	.69	.65	.70	.70	.70	.70	.70	.70	.70
			Control	*a*	.50	.61	.90	.90	.75	.69	.78	.90	.90
				ICC	.20	.28	.69	.70	.43	.36	.48	.68	.70
	Foundations	1	Experimental	*a*	.91	.87			.85	.85	.85	.85	
	of the History			ICC	.71	.63			.59	.59	.59	.59	
	of Aesthetics		Control	*a*	.76	.84			.68	.77	.80	.83	
				ICC	.44	.57			.35	.45	.51	.55	
		2	Experimental	*a*	.91	.89	.88		.86	.87	.85	.86	
				ICC	.71	.68	.64		.60	.62	.59	.61	
			Control	*a*	.90	.82	.73		.76	.77	.74	.83	
				ICC	.69	.54	.40		.44	.45	.42	.56	
		3	Experimental	*a*	.92	.90	.90	.90	.90	.90	.90	.90	.90

Semester	Subject	Stage	Group	Study	PR	DO	PW	A	AA	DE	GaL	EC	TM
				ICC	.73	.70	.69	.70	.70	.70	.69	.69	.69
			Control	*a*	.91	.90	.90	.83	.76	.79	.79	.89	.90
				ICC	.72	.70	.68	.55	.45	.48	.48	.67	.69
3	Concepts of	1	Experimental	*a*	.90	.85			.86	.86	.85	.85	
	Modern Art			ICC	.69	.58			.60	.60	.59	.59	
			Control	*a*	.62	.70			.85	.75	.86	.86	
				ICC	.29	.34			.55	.43	.60	.60	
		2	Experimental	*a*	.91	.91	.91		.90	.91	.92	.91	
				ICC	.71	.71	.71		.70	.71	.73	.72	
			Control	*a*	.60	.68	.91		.74	.72	.77	.91	
				ICC	.27	.34	.71		.42	.39	.45	.77	
		3	Experimental	*a*	.91	.90	.91	.91	.91	.91	.92	.91	.91
				ICC	.71	.69	.72	.72	.71	.72	.74	.72	.72
			Control	*a*	.90	.89	.91	.83	.51	.75	.75	.90	.91
				ICC	.69	.68	.71	.54	.21	.43	.43	.70	.71
	Theory of Art	1	Experimental	*a*	.97	.94			.85	.85	.85	.85	
				ICC	.88	.79			.59	.59	.59	.59	
			Control	*a*	.86	.84			.84	.71	.88	.90	
				ICC	.61	.57			.57	.38	.65	.69	
		2	Experimental	*a*	.90	.90	.91		.88	.88	.88	.88	
				ICC	.70	.68	.71		.65	.65	.65	.65	
			Control	*a*	.78	.87	.83		.72	.73	.75	.91	
				ICC	.47	.62	.55		.39	.41	.43	.71	
		3	Experimental	*a*	.90	.90	.90	.91	.91	.90	.91	.90	.91
				ICC	.69	.70	.70	.71	.71	.70	.72	.69	.71
			Control	*a*	.89	.89	.90	.78	.75	.74	.79	.90	.90
				ICC	.67	.67	.69	.47	.44	.42	.48	.69	.70

Note. PR = Project rationale; D = Documentation; PW = Previous work; A = Activities; AA = Appropriateness to audience; DE = Delivery; GaL = Grammar and language; EC = Eye contact; TM = Time management. α = Cronbach's alpha; ICC = Intraclass correlation coefficient.

In the control groups, 54.5% of the Cronbach's α were above .80, and mainly affected the criteria of the project's content. Just over a third of the coefficients (34.09%), distributed

unevenly over all the criteria, were situated between .70–.79. Only on 15 occasions (11.36%) were coefficients below .70, practically all in the criteria of project's content. In these groups, the α coefficients evolved erratically between the three stages, a finding that raised doubts about the student assessors' learning progress.

In the experimental groups, neither of these distortions appeared. With the sole exception of one coefficient, the Cronbach's α were above .80 in every assessment criterion and in the three stages. The high Cronbach's α obtained in the three stages means that it cannot be conclusively claimed that the student assessors notably improved their assessment skills. However, a significant improvement was observed in the criteria referring to the oral presentation, which rose from values around .85 in the first stage to minimum values of .90 in the final stage.

The ICC presented similar trends. According to the evaluation table proposed by Fleiss (1986), the coefficients in the control groups were for the most part fair to good (between .41 and .75), but were irregularly distributed between the stages. What is more, on 23 occasions (17.42%), generally criteria of the project's content, the coefficients obtained were poor (< = .40).

In the experimental groups, the coefficients were nearly always fair to good (between .41 and .75): in five cases they were excellent (> .75), four of them in the Project rationale criterion. Over the stages, the ICC moved towards the upper area of the Fleiss scale in all the subjects and semesters, and in the third stage were around .70. The study of the ICC allowed corroboration of what had been only a working hypothesis in the analysis of the Cronbach's α in the experimental groups: the student assessors notably improved their assessment skills as the stages progressed.

The analysis of the Cronbach's α and the ICC highlighted the difficulty of evaluating an event of short duration (the oral presentation), the differences in the communication skills

of the students, and the fact that each student may achieve different performance levels depending on the assessment criterion used.

Discussion

The contribution of rubrics to the consistent use of a performance assessment system

The fact that the Cronbach's α were generally above .80 in the three stages of the experimental and control groups shows that all the assessors used the assessment system to evaluate a complex phenomenon (the oral presentation of service-learning projects) in a reliable way. The Cronbach's α refers to the classical definition of reliability as internal consistency, replicability and stability of decisions represented by the scores (Gugiu, Gugiu, & Baldus, 2012; Hart & Hemker, 2013; Moskal & Leydens, 2001). This means that each assessor was consistent with respect to the meaning of the criteria and procedure assessment and of his/her vision of the objects evaluated; that is, intra-rater reliability was high.

The rubric was the only resource common to the experimental and control groups. In fact it may even have played a more important role because the student assessors had never previously come into contact with service-learning projects. The rubric helped the assessors to use the same criteria, to focus on the same aspects of the students' performances, and prevented the intrusion of subjective concerns and prejudices (Reynolds, 2009; Taub et al., 2011; Venning & Buisman-Pijlman, 2013; Wolf & Stevens, 2007). The study thus continues the line of previous work that acknowledges the contribution of well-designed rubrics to the reliability of the assessment practice (Bird & Yucel, 2013; Diller & Phelps, 2008; Jonsson & Svingby, 2007; Oakleaf, 2008; Reddy & Andrade, 2010; Schlitz et al., 2009).

Three main characteristics of the rubric helped assessors to construct a consistent understanding with respect to the assessment criteria and procedures. The first was the clarity of its criteria, levels of performance, and quality definitions. To avoid the difficulties with the interpretation and use of rubrics by students (Lapsley & Moody, 2007; O'Donovan, Price, &

Rust, 2004; Price & Rust, 1999), particular emphasis was placed on ensuring that the terminology and the expression of the quality definitions were suitably adapted to the undergraduate students.

The second characteristic of the rubric was its specificity. If its clarity was key to ensuring that it were understood by the assessors, its specificity meant that the assessors used it in accordance with the established purposes. It has been stressed that clear criteria are insufficient to ensure the quality of rubrics (Knight & Yorke, 2008; Sadler, 2009). The problems with rubrics do not appear only at the design stage; they may also emerge at the application stage and may principally affect how the criteria and levels of performance are used (Sadler, 2005; Surgenor, 2013). For this reason, the design of the rubric was specific to the discipline (Dunbar, Brooks, & Kubicka-Miller, 2006; Torrance, 2007), but also to the assignments and learning goals in which the students were engaged at various points in the semester. The rubric-based system was progressive because new criteria were added as the assessment stages advanced and because, inside each stage, new aspects were added as the levels of performance increased and their quality definitions became more complex.

The third characteristic of the rubric was its analytic character. We preferred to use an analytic rather than holistic rubric because it specify the information conveyed to the students (Moskal, 2001; Wolf & Goodwin, 2007) and facilitate students' assessment as it do not require an expert judgement based on an overall impression (Dunbar, Brooks, & Kubicka-Miller, 2006). However, we also had to consider the arguments against a highly detailed description of the rubrics (O'Donovan, Price, & Rust, 2001; Sadler, 2005) and avoid the risk that a superficial use would provide only a strategic approach to learning (Moskal & Leydens, 2001; Torrance, 2007) and undermine students' autonomous learning (Bell, Mladenovic, & Price, 2013; Boud & Falchikov, 2006). In our opinion, the problem lies not in the detail of the quality definitions but in whether they are aimed towards higher-order

learning outcomes or trivialities. As a consequence, our rubric listed the essential aspects of a service-learning project and its oral presentation, stressing their complexity, and in fact making this complexity the central focus of the discussion between students, student assessors and teachers.

However, the presence of intra-rater reliability is not sufficient grounds to claim that the undergraduate students reached higher-order assessment skills. In accordance with the criticisms raised by Gugiu, Gugiu, and Baldus (2012) regarding the use of stability or internal consistency estimators, the analysis of the inter-rater reliability was required, by calculating the intraclass correlation coefficients.

Instructional resources for developing students' higher-order assessment skills

While the ANOVA showed significant differences between the assessors included in control groups, the intra-rater reliability gave only a partial indication that all the assessors belonged to the same assessment culture. The use of inter-rater reliability estimators helped to explain the extent to which assessors shared representations regarding the assessment criteria and procedures and the perception of the phenomena evaluated.

The lower ICC values in the control groups and above all their erratic distribution over the semester were insufficient to demonstrate that the undergraduate students were fully incorporated in the assessor team and had achieved significant learning related to higher-order assessment skills. Only the ICC of the experimental groups showed a maintained learning progress over the semester; the predominance of the good coefficients in the criteria of the third stage indicated that the undergraduate students achieved the learning outcomes desired. The evolution of the ICC in these groups also reflected the difficulty involved in developing consistent assessment skills and demonstrated the need for instructional settings in which suitable assessment is carried out on a regular basis.

The results demonstrated that the training programme performed in the control groups was not sufficient for the undergraduate students to attain the learning goals, even though its duration was similar to that of other studies (Diller & Phelps, 2008; Dunbar, Brooks, & Kubicka-Miller, 2006; Reynolds et al., 2009). Linn (1994) detected a positive influence of training on assessment validity and reliability, a finding confirmed by other investigators with student assessors (Crotwell et al., 2011; Welch & James, 2007). Without conclusively rejecting this evidence, this study did not find that scorer training was the most important factor for achieving this validity and reliability, as Boulet et al. (2004) claimed. We believe that a brief training programme for students does not guarantee a significant improvement in their judgements, even if the programme includes assessment practices with exemplars of real student work (Smith et al., 2013).

In contrast, the experimental groups explored the influence on the undergraduate students of more profound forms of participation in the assessor team, that is, their involvement in the design of the rubrics and in the discussion of the projects evaluated. If the engagement of the students in discussions about the assessment system favours the development of independent problem-solving skills and professional judgements (Oakleaf, 2008; Sadler, 2009), greater benefits will be obtained if they take part in its design (Venning & Buisman-Pijlman, 2013; Wolf & Stevens, 2007). However, the usefulness of promoting the participation of students in the creation of the rubrics has been questioned (Beck, Skinner, & Schwabrow, 2013); in fact, few studies of the reliability of student scoring have been conducted to date, and their results are inconclusive (Crotwell et al., 2011; Hughes & Cappa, 2007).

We agree with Diller and Phelps (2008) that the collaboration of all the assessors in the creation of rubrics helps to avert the problems deriving from the interpretation of the criteria and procedures. The design stage is a recurrent process of construction of a consensus

regarding: (a) the selection of the attributes of the object under evaluation and their relative value according to the context and the educational aims, and (b) the best way of informing all the stakeholders, especially the students, of the decisions taken (Bird & Yucel, 2013; Smith et al., 2013). In our opinion, the introduction of the undergraduate students in the processes of negotiation increased levels of validity and reliability: they acquired a clearer understanding of the quality of the service-learning projects and how this quality is reflected in the criteria and performance levels, and their involvement helped to adapt the language used in the rubrics more closely to the student profile (Jonsson, 2014; Reddy & Andrade, 2010).

However, it is not enough just to understand the meaning of the assessment criteria and procedures, or to have an overall vision of how they should be used. What defines expert judgement is its specific application to each object assessed: in the specific case of this study, an ability to adapt the assessment system to the different way in which a service-learning project is presented. In other words, the key here is not to understand the assessment system, but to be able to interpret the object under evaluation in all its complexity. This was the main objective of the moderation discussions.

In general, the post-marking moderation strategies such as double-marking techniques start with the identification of discrepant scores. They aim to obtain shared scores and then to refine the assessment system via a consensual review of the criteria and the application procedures. The effect of moderation discussions of this kind on reliability and agreement is controversial (e.g. Bloxham, 2009; Vardi, 2013). In our view, its weakness lies in the fact that the search for consensus is made when the assessors have already made an evaluation, and so achieving shared scores is more complex: first, because each assessor has an established idea of the value of the object assessed and, second, because the negotiation involves spurious aspects for the assessment process, such as questions of professional ranking and prestige of the participants. For this reason, following the studies by Bird and Yucel (2013) and Hughes

and Cappa (2007), we prefer to implement a pre-marking moderation strategy. Our proposal differed in that the moderation discussions were scheduled immediately after each assessment session and preceded the final stage in which each assessor gave their scores individually. The object was not to attain a shared understanding regarding the construct definition and assessment criteria (which was the object of the design stage), but to compare and debate the different perceptions that the assessors had of each of the projects presented. In this way, undergraduate students and teachers collectively constructed the same vision of the objects assessed.

Limitations of the study and proposals for further research

The first limitation is that the study design did not allow a distinction to be made between the relative impact of the participation of the students in the design of the rubric and in the moderation discussions. Future research should use broader samples of undergraduate students in order to study the results of two interventions: in one, the participation of the student assessors should be limited to the design of the rubrics, and in the other, some of these students would also attend the moderation discussions. Second, the possible benefits of intensive training (Taub et al., 2011) were not examined because the teachers did not have sufficient time to prepare all the resources and teach a programme that was not included in their teaching activities. In this case as well, a more flexible conception of the curricula is required that would allow the incorporation of programmes of this type and provide greater institutional support for teaching innovation and educational research. Third, there are other aspects that should be studied in future research using either qualitative or mixed methodologies. The first is the study of students' representations of the assessment, and of their role in it, after the training programme or their incorporation into the assessors' team alongside the teachers. The second is the study of their perceptions of certain key aspects of the rubric: the meaning and utility that they conferred on the rubric-based evaluation system

compared with other systems present in their immediate educational environment; the procedures they used to apply the rubric in the assessment process; the difficulties they encountered, and how they resolved them; and the changes in their impressions regarding the meaning and use of the rubric during their participation.

Conclusion

This study supports the utility of rubrics in the development of professional judgement in undergraduate students, starting from the premise of sustainable assessment. Rubrics favour the development of independent evaluative skills of professional relevance, because they guide the student in assessment practice by establishing preferred themes and points of reference that can help to conduct the debates. Our study demonstrates that a brief training programme focusing on the nature of the objects assessed and on the meaning and the use of the rubrics was not sufficient to achieve the higher-order learning outcomes desired; to do so, the full incorporation of the undergraduate students in the assessor team was necessary. In these circumstances, rubrics are instructional resources that help the student to identify the points of consensus of a professional culture and to understand how they develop. Thus, the greatest benefits of the rubrics do not lie in the instrument of evaluation created, because their scope is limited to a specific educational context and their applicability is subject to continuous revision. Rather, their main contribution lies in the fact that they define and support the processes of negotiation and agreement which determine the assessment criteria and procedures and the perceptions of the objects assessed.

Conceived in this way, rubrics strengthen the assessor teams by promoting the emergence of authentic communities of practice open to the participation of the students. In this process, the assessor teams become work groups fully geared towards fostering the development of professional judgement in real contexts, providing specific pedagogical training for the students acting as instructors, and promoting learning environments in which

students develop a fuller commitment to their teaching and learning activities.

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 [GINDOC-UB/103].

References

Beck, R. J., Skinner, W.F., & Schwabrow, L.A. (2013). A study of sustainable assessment theory in higher education tutorials. *Assessment & Evaluation in Higher Education*, *38*(3), 326–348.

Bell, A., Mladenovic, R., & Price, M. (2013). Students' perceptions of the usefulness of marking guides, grade descriptors and annotated exemplars. *Assessment & Evaluation in Higher Education*, *38*(7), 769–788.

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

Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, *34*(2), 209–220.

Boud, D. (2000). Sustainable assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167.

Boud, D., & Falchikov, N. (2005). Redesigning assessment for learning beyond higher education. Paper presented at the 28th HERDSA Annual Conference. Higher Education in a Changing World. Sydney, July 3–6. Retrieved 23 June 2012 from http://www.herdsa.org.au/wp-content/uploads/conference/2005/papers/boud.pdf

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, *31*(4), 399–413.

Boud, D., & Falchikov, N. (2007). Developing assessment for informing judgement. In D. Boud & N. Falchikov (Eds.), *Rethinking assessment for higher education: learning for the longer term* (pp. 181–197). London: Routledge.

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413.

Boulet, J. R, Rebbecchi, T. A., Denton, E. C., Mckinley, D. W., & Whelan, G.P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education*, *9*(1), 47–60.

Brooks, V. (2012). Marking as judgment. *Research Papers in Education*, *27*(1), 63–80.

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, *18*(1), 32–42.

Carroll, T. (2005). Induction of teachers into 21st century learning communities: Creating the next generation of educational practice. T*he New Educator*, *1*(3), 199–204.

Cevik, Y. D., & Andre, T. (2014) Studying the impact of three different instructional methods on preservice teachers' decision-making. *Research Papers in Education*, *29*(1), 44–68.

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

Diller, K. R., & Phelps, S. F. (2008). Learning Outcomes, Portfolios, and Rubrics, Oh My! Authentic Assessment of an Information Literacy Program. *Portal: Libraries and the Academy*, *8*(1), 75–89.

Duguid, P. (2005). 'The art of knowing': Social and tacit dimensions of knowledge and the limits of the community of practice. T*he Information Society*, *21*(2), 109–118.

Dunbar, N. E., Brooks, C. F., & Kubicka-Miller, T. (2006). Oral Communication Skills in Higher Education: Using a Performance-Based Evaluation Rubric to Assess Communication Skills. *Innovative Higher Education*, *31*(2), 115–128.

Ecclestone, K. (2001). 'I know a 2:1 when I see it': Understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education*, *25*(3), 301–313.

El-Mowafy, A. (2014). Using peer assessment of fieldwork to enhance students' practical training. *Assessment & Evaluation in Higher Education*, *39*(2), 223–241.

Emad, G. R., & Roth, W.-M. (2016). Quasi-communities: rethinking learning in formal adult and vocational education. *Instructional Science*, *44*(6), 583–600.

Fleiss, J. L. (1986). T*he design and analysis of clinical experiments*. New York: Wiley.

Fraile, J., Panadero, E., & Pardo, R. (2017). Co-creating rubrics: The effects on self-regulated learning, self-efficacy and performance of establishing assessment criteria with students. *Studies in Educational Evaluation*, *53*, 69–76.

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

Hart, H., & Hemker, B. T. (2013). On the reliability of vocational workplace-based certifications. *Research Papers in Education*, *28*(1), 75–90.

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

Hughes, C., & Cappa, C. (2007). Developing generic criteria and standards for assessment in law: processes and (by)products. *Assessment & Evaluation in Higher Education*, *32*(4), 417–432.

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

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

Kearney, S. (2013). Improving engagement: the use of 'Authentic self-and peer-assessment for learning' to enhance the student learning experience. *Assessment & Evaluation in Higher Education*, *38*(7), 875–891.

Klein, J. H. (2008). Some directions for research in knowledge sharing. *Knowledge Management Research & Practice*, *6*, 41–6.

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

Knight, P., & Yorke, M. (2008). Assessment Close Up: The Limits of Exquisite Descriptions of Achievement. *International Journal of Educational Research*, *47*(3), 175–183.

Kobayashi, S., Berge, M., Grout, B. W. W., & Østerberg Rump, C. (2017). Experiencing variations: learning opportunities in doctoral supervision. *Instructional Science*, *45*(6), 805–826.

Kolb, D. A. (1984). *Experiential learning: experience as the source of learning and development*. New York: Prentice-Hall.

Kwon, K., & Park, S. J. (2017). Effects of discussion representation: comparisons between social and cognitive diagrams. *Instructional Science*, *45*(4), 469–491.

Kyun, S., Kalyuga, S., & Sweller, J. (2013). The Effect of Worked Examples When Learning to Write Essays in English Literature. T*he Journal of Experimental Education*, *81*(3), 385–408.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lapsley, R., & Moody, R. (2007). Teaching tip: Structuring a rubric for online course discussions to assess both traditional and non-traditional students. *Journal of American Aacademy of Business*, *12*(1), 167–172.

Lave, J. (1988). *Cognition in practice: Mind, Mathematics and Culture in Everyday Life*. Cambridge: Cambridge University Press.

Lave, J., & Wenger, E. (1991). *Situated learning: legitimate peripheral participation*. Cambridge: Cambridge University Press.

Linn, R. L. (1994). Performance assessment: Policy promises and techncel measurement standards. *Educational Researcher*, *23*(9), 4–14.

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

Luft, J. A., Kurdziel, J. P., Roehrig, G. H. & Turner, J. (2004). Growing a garden without water: Graduate teaching assistants in introductory science laboratories at a doctoral/research university. *Journal of Research in Science Teaching*, *41*(3), 211–233.

Menéndez-Varela, J.-L., & Gregori-Giralt, E. (2016). The contribution of rubrics to the validity of performance assessment: a study of the conservation?restoration and design undergraduate degrees. *Assessment & Evaluation in Higher Education*, *41*(2), 228–244.

Menéndez-Varela, J.-L., & Gregori-Giralt, E. (2018). The reliability and sources of error of using rubrics-based assessment for student projects. *Assessment & Evaluation in Higher Education*, *43*(3), 488–499.

Moskal, B. M. (2001). Scoring Rubrics: What, When and How? In L. M. Rudner, & W .D. Schafer (Eds.), *Practical Assessment, Research & Evaluation 2000-2001* (pp. 22–29). Retrieved 23 June 2012 from http://www.eric.ed.gov/PDFS/ED458254.pdf

Moskal, B. M., & Leydens, J. A. (2001). Scoring Rubric Development: Validity and Reliability. In L. M. Rudner, & W .D. Schafer (Eds.), *Practical Assessment, Research & Evaluation 2000-2001* (pp. 71–81). Retrieved 23 June 2012 from http://www.eric.ed.gov/PDFS/ED458254.pdf

Nordrum, L., Evans, K., & Gustafsson, M. (2013). Comparing student learning experiences of in-text commentary and rubric-articulated feedback: strategies for formative assessment. *Assessment & Evaluation in Higher Education*, *38*(8), 919–940.

Norton, L. (2004). Using assessment criteria as learning criteria: A case study in psychology. *Assessment & Evaluation in Higher Education*, *29*(6), 687–702.

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

O'Donovan, B., Price, M., & Rust, C. (2001). The student experience of criterion-referenced assessment (through the introduction of a common criteria assessment grid). *Innovations in Education and Teaching International*, *38*(1), 74–85.

O'Donovan, B., Price, M., & Rust, C. (2004). Know what I mean? Enhancing student understanding of assessment standards and criteria. T*eaching in Higher Education*, *9*(3), 325–335.

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

Panadero, E., & Jonsson, A. (2013). The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review. *Educational Research Review*, *9*, 129–144.

Polanyi, M. (1958). *Personal knowledge: Towards a post-critical philosophy*. London: Routledge & Kegan Paul.

Polanyi, M. (1962). Tacit knowing: Its bearing on some problems in philosophy. *Reviews of Modern Physics*, *34*(4), 601–616.

Price, M. (2005). Assessment standards: The role of communities of practice and the scholarship of assessment. A*ssessment & Evaluation in Higher Education*, *30*(3), 215–230.

Price, M., & Rust, C. (1999). The experience of introducing a common criteria assessment grid across an academic department. *Quality in Higher Education*, *5*(2), 133–144.

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

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

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175–194.

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

Schlitz, S. A., O'Connor, M., Pang, Y., Stryker, D., Markell, S., Krupp, E., Byers, C., Jones, S. D., & Redfern, A.K. (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.

Schön, D. A. (1983). T*he reflective practitioner: How professionals think in action*. London: Temple Smith.

Singh, K., & Terry, J. (2008). Fostering students' self-assessment skills for sustainable learning. Paper presented at the EDU-COM 2008 International Conference. Sustainability in Higher Education: Directions for Change, Perth, November 19–21. Retrieved 23 June 2012 from http://ro.ecu.edu.au/cgi/viewcontent.cgi?article=1038&context=ceducom

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2013). Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education*, 38(1), 44–60.

Surgenor, P. W. G. (2013). Measuring up: comparing first year students' and tutors' expectations of assessment. *Assessment & Evaluation in Higher Education*, *38*(3), 288–302.

Suto, W. M. I., & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insight into the GCSE marking process. *British Educational Research Journal*, *34*(2), 213–233.

Taub, D. J., Servaty-Seib, H. L., Wachter Morris, C. A., Prieto-Welch, S. L., & Werden, D. (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.

Torrance, H. (2007). Assessment as Learning? How the Use of Explicit Learning Objectives, Assessment Criteria and Feedback in Post-Secondary Education and Training can come to Dominate Learning. *Assessment in Education*, *14*(3), 281–294.

Tsoukas, H. (2002). Do we really understand tacit knowledge? *Paper presented at the Knowledge, Economy and Society Seminar*, LSE Department of Information Systems, London, June 14.

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

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

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

Watson, D., & Robbins, J. (2008). Closing the chasm: reconciling contemporary understandings of learning with the need to formally assess and accredit learners through the assessment of performance. *Research Papers in Education*, *23*(3), 315–331.

Welch, M., & James, R. C. (2007). An Investigation on the Impact of a Guided Reflection Technique in Service-Learning Courses to Prepare Special Educators. T*eacher Education and Special Education*, *30*(4), 276–285.

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

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