The Content Aspect of Validity in a Rubric-based Assessment System for Course Syllabuses

Eva Gregori-Giralt[a]

gregori@ub.edu

ORCID: 0000-0002-9774-1563

José-Luis Menéndez-Varela[a]

menendez@ub.edu

ORCID: 0000-0002-6733-9346

Affiliations: [a]Department of History of Art, University of Barcelona, Barcelona, Spain

Affiliation addresses: [a]University of Barcelona. Department of History of Art, Montalegre, 6, 08001 Barcelona (Spain) Tel.: +34 934037589. Fax: +34 934035976

Corresponding author: José-Luis Menéndez Varela. University of Barcelona. Department of History of Art, Montalegre, 6, 08001 Barcelona (Spain) Tel.: +34 934037589. Fax: +34 934035976

Abstract

The growing trend among universities to promote systems of programme and course evaluation entails more responsibility for faculties and departments. These systems require resources to ensure that they are not only valid and reliable but also effective and sustainable. The design of rubric-based assessment systems may provide a solution, but there is a gap in the research on curriculum evaluation concerning their use and validation. We examine the content aspect of validity in a rubric-based assessment system for course syllabuses using a

mixed method that combines an analysis of the agreement among 23 experts with a phenomenographic study. With data gathered through a questionnaire linked to the Delphi technique, content validity indexes were calculated and the experts' different perspectives were identified. The content validity indexes (greater than .80) met the standards set out in literature, and the qualitative study of the experts' feedback showed three different perspectives on the system's use. Beyond providing evidence of the system's content validity, the study highlights the extent to which it is important to give appropriate consideration to experts' – and by extension final users' – experience in order to ensure the successful implementation of rubric-based assessment systems.

Keywords: programme evaluation; evaluation methods; rubric; course syllabus.

Introduction

Recently, accreditation programmes have adopted an evolutionary and developmental approach that accords with the main purpose of educational quality assurance systems, which is to provide the means necessary for the continuous improvement of educational programmes (e.g., Boyle & Bowden, 1997). It is not enough to use the rates of academic success customarily cited by universities as evidence of the smooth running of their degree programmes. This is partly due to the disrepute arising from grade inflation (e.g., Bachan, 2017; Chowdhury, 2018; Finefter-Rosenbluh & Levinson, 2015) and partly because such rates of academic success say nothing of the impact that university programmes have on their students' learning processes.

This state of affairs has led universities to focus on programme and course evaluation, resulting in increased responsibility for faculties and departments. These must provide

evidence that their degree programmes foster high-level learning outcomes and that these outcomes meet the needs of society. This involves the implementation and upgrading of systems to support data capture and entry on student learning, the use of appropriate systems of analysis, and the provision of planning and rationales for policies to improve degree programmes in accordance with the interpretation of the findings (Goldstein, 2010). The focus of evaluation is no longer the design of an educational programme, but rather on its performance (Caffarella, 2002; Hixon, Barczyk, Buckenmeyer, & Feldman, 2011). This, in turn, has even led to studies on meta-assessment, that is, the evaluation of the suitability of programme assessment processes (Orem, 2012).

In addition to the conceptual difficulties inherent in the evaluation of educational quality, programme and course evaluation must also solve problems of implementation. The design and application of evaluation systems require plentiful resources (e.g., Uribe, 2013); therefore, they must be not only valid and reliable, but also effective and sustainable (e.g., Barrie, Hughes, Crisp, & Bennison, 2014). Steps must be taken to compensate for the lack of a suitable assessment culture among faculty, due partly to inadequate pedagogical training (Grainger, Adie, & Weir, 2016) and partly to the persistence of a university tradition that impedes a paradigm shift in education (see Boyle & Bowden, 1997; Brownell & Tanner, 2012). The latter factor is the more important one because it affects faculty commitment to the continuous assessment of programmes. For instance, Brancaccio-Taras et al. (2016) developed a programme assessment system in which one of the criteria was the presence of a suitable institutional atmosphere for the implementation of evidence-based teaching practices.

The implementation of programme and course evaluation involves giving a prominent role to faculty members (Gerretson & Golson, 2005), which in turn has led to arguments in favour of

shifting the focus from programme-level to course-level assessment (Reed, Levin, & Malandra, 2011). Consequently, there is a pressing need to bolster the creation of faculty learning communities (e.g., O'Malley, 2010; Ward & Selvester, 2012) that are open to teachers from different universities who teach in the same field or discipline (Sefcik, Bedford, Czech, Smith, & Yorke, 2018). In fact, these communities can serve as professional environments seeking to enhance learning and professional development governed by the principles of trust, support and collegiality (Author, 2018). With adequate institutional support, these professional safe environments could also foster the creation of teaching innovation groups to act as levers of change within faculties and departments. For example, teaching innovation groups could promote the design and application of programme and course assessment systems or their transfer from other contexts.

Analysing the teaching process as a unified construct is complex, and it is difficult for faculties and departments to gather direct measures in order to carry out an analysis that is valid, reliable and sustainable. This is why universities have promoted the collection of indirect evidence on quality assurance in addition to survey-based studies. Thus, recent research on programme and course evaluation has pushed forward in three directions. First, studies have examined whether a graduate will have attained the stated profile based on the distribution of competencies in the curriculum. This is exemplified by research on curriculum mapping (e.g., Perera, Babatunde, Zhou, Pearson, & Ekundayo, 2017; Veltri, Webb, Matveev, & Zapatero, 2011; Wijngaards-de Meij & Merx, 2018). Second, studies have checked whether students have attained the competencies set out in curricula; this focus has proved fertile ground for the use of rubrics. Prominent examples include analyses of core competencies, primarily the competencies of information literacy (e.g., Whitlock & Ebrahimi, 2016) and oral or written communication (e.g., García-Ros, 2011; Good, Osborne,

& Birchfield, 2012), but also including studies on specific competencies (e.g., Romkey, Chong, & El Gammal, 2015; Tractenberg, Umans, & McCarter, 2010). Third, studies have verified whether the learning environments set out in course syllabuses are consistent with the pedagogical principles of a competency-based higher education.

The course syllabus is a document that lays out the learning outcomes, content and teaching environment that define a course. The quality of a course syllabus requires that all of its components should be aligned, but it also calls for a clear and specific explanation of the components so that a student can make an informed decision on whether or not to enrol in a course or degree programme. An analysis of the content of syllabuses does not produce direct evidence of the actual learning process that takes place in the classroom. However, it does provide information on: a) how well the teaching design is aligned with the student learning outcomes; b) the teaching culture and practices of the faculty; and c) the extent of their commitment to a competency-based educational model. Teacher-specific syllabuses are typically accessible only to enrolled students, but they co-exist in some university systems with generic syllabuses that are the product of a consensus among the teachers of a single course and they establish the general framework of the course. Course syllabuses are used as teaching resources, course plans, and evidences for teacher evaluation and programme accreditation (Grunert O'Brien, Millis, & Cohen, 2008; Slattery & Carlson, 2005; Willingham-McLain, 2011). In these roles, they function as communication tools and educational contracts (Fink, 2012; Parkes & Harris, 2002, Singham, 2007). The fact that generic syllabuses are sometimes available to the public makes them indicators of the attention that a university pays to its educational mission. Being useful to compare course programmes (e.g., Álvarez-Pérez, González Morales, López-Aguilar, Peláez Alba, & Peña Vázquez, 2018), generic syllabuses are analysed by the national quality agencies in the

processes of programme accreditation (e.g., National Agency for Quality Assessment and Accreditation, 2013).

As shown in research spanning from the study by Bers, Davis, and Taylor (2000) to more recent work conducted by Goodwin, Chittle, Dixon, and Andrews (2018) and Mathers, Finney, and Hathcoat (2018), there remains a large scope for improvement in aligning course syllabuses with the teaching practices that are most highly valued in the literature. Consequently, this line of evaluation research needs to be incorporated into already existing quality assurance systems at universities. In addition, syllabus content analysis is the type of programme and course assessment that is least time-consuming (Stanny, Gonzalez, & McGowan, 2015; Willinghan-McLain, 2011), that best takes advantage of the faculty's expert knowledge and that can most directly and most immediately be useful in faculty development (Bers et al., 2000).

An examination of the latest publications reveals a variety of approaches, methodologies and evaluation tools, illustrating the interest in the topic and the exploratory stage of current research. There have been survey-based studies, such as the one by Bergsmann, Klug, Burger, Först, and Spiel (2018), who analysed the presence of competency-based teaching and real student competencies and did so using a screening model with an online questionnaire to which students and teachers alike responded. Iudica (2011) made use of two checklists to examine the alignment between course syllabuses in educational technology leadership and the state and national technology standards. Goodwin et al. (2018) put forward a qualitative analysis of the information on learning outcomes, reading requirements, learning activities, assessment types, and policy listing and adherence as set out in the course syllabuses of an undergraduate programme. Lastly, the way in which assessment systems are

described in course syllabuses has received special attention from a number of viewpoints: a) whether students receive appropriate communication about assessment aims (Thomas et al., 2018); b) whether the assessment system is aligned with learning outcomes (Sefcik et al., 2018); and c) which assessment modes and instruments are most commonly used in educational contexts (Tucker, 2012).

Rubrics are also valuable resources because they have given rise to the largest number of studies in the field in question. Brancaccio-Taras et al. (2016), Halim (2008) and Stanny et al. (2015) are examples of studies that analyse all aspects of the learning environments set out in course syllabuses and their alignment with the pedagogical principles and best teaching practices that are most widely recognised in the literature. Raybon (2012) and Legon (2015) have the distinctive characteristic that their rubrics feature specific criteria for online courses. However, it is more common to find rubrics used to assess whether the course syllabuses are consistent with the principles of learner-centred teaching (Blumberg, 2009; Blumberg & Pontiggia, 2011; Cullen & Harris, 2009), students' meaningful learning (Koh, 2013) or the assessment for learning (Alonzo, Mirriahi, & Davison, 2018; Wolf & Goodwin, 2007).

At a time when the rubric-based assessment systems for student performance had not yet reached a stage of maturity in higher education, Jonsson and Svingby (2007) and Reddy and Andrade (2010) confirmed the need for a greater number of studies on validity that have rigorous research methods and analyses. More recently, Dawson (2017) found very few publications that were suitable for replication studies because they contained insufficient information on their research method. Unsurprisingly, this is also the case now with the rubric-based assessment systems for course syllabuses. Of the studies cited earlier, only those

of Halim (2008), Koh (2013) and especially Alonzo et al. (2018) address the matter of validity.

The present study is a contribution to the as yet limited body of literature on the rubric-based assessment systems for course syllabuses and, more specifically, to the models used to assess all the components set out in course syllabuses. The proposed rubric has six dimensions that reflect the customary sections of course syllabuses – i.e., learning outcomes, course content, learning resources, learning activities, learning mode, and assessment – and it features four performance levels (see Appendix A). The aim of this paper is to analyse some of the aspects that underpin the validity of the inferences that are drawn from the application of the rubric. Accordingly, two research questions are posed: (1) What rating is given to the content aspect of validity for the rubric-based assessment system? and (2) What are the professional experiences and concerns shown by the experts who took part in its evaluation?

Study context

The European Higher Education Area (EHEA), which was launched in 2010, is a joint undertaking of 48 European countries that are cooperating in the construction of a framework of comparable and compatible higher education systems. As a result, national systems of higher education have begun to be regulated by common quality standards. Also, the European Association for Quality Assurance in Higher Education (ENQA) and national organisations for quality assurance have been set up to conduct the follow-up and evaluation of degrees and to issue recommendations on quality improvement to higher education institutions.

The need to comply with accreditation standards has led to greater awareness of universities' educational missions. In reality, however, the assessment of teaching–learning environments has not yet reaped all the benefits of quality agency programmes, and this is a worldwide spread phenomenon (Storey & Asadoorian III, 2014). The main purpose of an educational quality assurance system consists precisely in improving student learning opportunities (Ellis, Jarkey, Mahony, Peat, & Sheely, 2007). For this to be possible, educational quality assessment must take place at classroom, department and degree levels, in compliance with the general institutional framework and evaluation agency guidelines. Institutions, however, do not always dispose of appropriate assessment instruments to produce evidence of implemented educational improvements (Blumberg & Pontiggia, 2011). Moreover, these instruments are even less easily accessible to teachers and students.

In this general context, the faculty of Spanish public universities must annually update the generic syllabuses of every course in order to make them available to the public as part of the published information provided on degree programmes. Undoubtedly, this task gives concrete shape to the universities' commitment to accountability and transparency policies. However, the process of updating course syllabuses has shortcomings that jeopardise the ongoing quality improvement of teaching environments and lead faculty members to perform the task of updating as a matter of routine. On one hand, teachers have only general guidelines on the information that must be put in the various sections of a somewhat inflexible software application. The result is that course syllabuses are very homogeneous – with the exception of the subject matter – and not very specific.

On the other hand, teachers do not have: a) educational quality criteria specifically adapted to their discipline on which to base the review of their course syllabuses; b) best practices for

the preparation of course syllabuses in their own degree programmes to serve as examples; and c) quality assessment systems for course syllabuses to enable them to evaluate the current state of their courses and plan an ongoing quality improvement process. The lack of benchmarks and resources is disorienting and discouraging for faculty members. It promotes the belief that the activity is nothing more than a bureaucratic hurdle and it prevents any real ongoing improvement process for course syllabuses (see Carrión Martínez, 2006; Guerra García, 2013).

In light of these key factors, we have taken the view that addressing the last issue may lay the groundwork for successfully tackling the other issues. A rubric-based assessment system for course syllabuses in the hands of faculty should generate the data necessary to identify any shortcomings, determine their relative importance and plan a short-term and medium-term improvement programme, provided that the assessment system can demonstrate its validity and reliability. With widespread use of the proposed system, faculty could agree on the quality criteria of course syllabuses in their respective academic fields and, accordingly, develop a catalogue of best practices over time.

Method

The approach is based on Messick's unified construct validity theory, which states that "validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of interpretations and actions based on test scores or other modes of assessment" (Messick, 1990, p. 5). Messick's concept of validity has three corollaries: a) validity is a unified construct grounded on the integration of the analysis of empirical evidence from multiple sources; b) validity is a judgement subject

to constant review; c) validity not only refers to the plausibility of the rationales but also to the consequences of such rationales for their immediate surroundings.

Messick (1989, 1990, 1995) established six facets or aspects of the concept of construct validity: content, substantive, structural, generalizability, external and consequential. To analyse all six facets is all the more important because there are so few studies that cover every aspect of validity (Docktor, 2009). In this paper, the analysis focuses solely on the content aspect in the conviction that the remaining aspects must be tackled in subsequent studies. In accordance with Messick, the content aspect includes evidence on whether the rubric's criteria and quality indicators are relevant and representative according to expert assessors and the specialised literature.

Participants

To examine the content aspect of validity, the study brought together a panel of 23 experts, none of whom withdrew from the research. A purposive sampling was made on the basis of three criteria: a) that they came from different universities and disciplines; b) that they represented different categories of faculty, and c) that they were linked to universities of different sizes that provided education both in situ and online.

One of the criteria for identifying the experts was that they must have had a recent line of research on rubrics accredited by some academic publication in the past five years. Also, the experts had to share the language of the rubric in its original version, given that grammar was one of the aspects for them to evaluate. The selection process was complex because the acceptance of each expert made it necessary for the authors to go back and revise the initially selected list of experts in order to achieve a balance among the three criteria set out in the

previous paragraph. Consequently, the process took two months and 53 invitations were sent out. The response rate was 43.39%.

The experts carry out their academic activity in different disciplines in the Education Sciences and the Arts and Humanities and they are linked to eight universities in two countries. These are the two Spanish-speaking countries with the highest number of publications on rubrics. The researchers did not succeed in enlisting the participation of any experts from the hard sciences. The professional status of the participants was distributed as follows: three professors, one reader, four senior lecturers, eight lecturers (with between one and seven years of academic experience), four associate lecturers, one research associate and two predoctoral fellows. They received no remuneration or compensation of any kind for their participation.

Ethical considerations

Scrupulous attention was paid to the principles and codes of best practices widely recognised in the scientific community. First, the experts took part in the study voluntarily after accepting an invitation by email that included information on the purpose and the object of study, the methodology, the type of participation and the anticipated length of the research. The aim was to ensure that the experts gave their informed consent to take part. The invitation was also explicit about the following issues: a) the confidential treatment of the data; b) the right of participants to leave the study at any time without the need to provide a reason beyond their own express wish to do so; c) the right of participants to review their contributions prior to publication; and d) their right to receive information on the results of the research upon its conclusion. In the qualitative analysis, any element that might identify

the participants was removed and their comments were preserved verbatim with the exception of eliminating any typos that might hamper comprehension.

Second, we have always sought to ensure the highest level of objectivity in analyses and discussions throughout the research. We have presented the findings in ways that permit scrutiny by other researchers and we have cited any studies by other researchers that have been used in the present study in accordance with international standards.

Data collection techniques and instruments

The Delphi method was used as an anonymous process of consensus building among the 23 experts on the content relevance and representativeness of the rubric's criteria and quality indicators. In the first round, a questionnaire was sent to experts containing 23 closed items worded as statements that required a mandatory response (see Table 1 below and Appendix B). Also included was an open-ended text field for optional response so that the experts could add any qualitative feedback to their evaluation. The items were grouped into four components: grammar (two items), consistency (five items), structure (three items) and content (13 items). These are explained in greater detail below. A four-point Likert scale was used, with the anchors ranging from least to most agreement: totally disagree, disagree on the whole, agree on the whole, totally agree.

After receiving the experts' ratings, the content validity index was calculated to analyse the degree of consensus among the experts. The grammar component and the consistency component were found to have a rating slightly below .80, which made it necessary to do a second round. For the two items related to grammar, the experts concurred that the wording overall was sufficiently clear (item 1), but they warned that too much technical language in

the quality indicators could hamper the ability of non-specialist readers to understand the rubric (item 18). As a result, the researchers reviewed the rubric and identified expressions from the field of Education Sciences that could be reworded without losing any of their accuracy or precision. In one instance, it was imperative to keep the term in question and the experts introduced a paraphrase to help in understanding the text. Take for example the Learning outcomes dimension: given the widespread use of the concept of competence in the university context, it was deemed appropriate to keep the term "competences" in parentheses and explain it as "the main learning outcomes" (see Appendix A).

For the five items in the consistency component, no issues were found in relation to whether the rubric covered the main aspects of course syllabuses (item 2), whether the rubric enabled evaluation of their content (item 26), or whether the rubric contained widely recognised quality indicators for teaching environments (item 51). The experts were also in agreement that the rubric would need a training programme in order to be able to apply it (item 29). While this raised questions about the rubric's ease of use, it accorded with an issue that has been stressed repeatedly in the literature on rubrics (e.g., Bird & Yucel, 2013; Crotwell Timmerman, Strickland, Johnson, & Payne, 2011; Reddy & Andrade, 2010; Sue, 2020; Venning & Buisman-Pijlman, 2013). It was also consistent with the challenge of assessing course syllabuses in all their complexity.

The responses to item 11 raised another reason to revise the rubric concerning its usability in the experts' academic settings. Indeed, some of them gave this item the lowest rating on the scale. In this regard, we made two kinds of modifications to the rubric, which affected the dimensions of Learning activities, Learning mode and Assessment. First, it was necessary to remove the more specific statements in the quality indicators and replace them with more

generic wording. For the Learning activities dimension, one example was to remove direct references to problem-based or project-based learning environments and to put in their place the desirability of connecting learning activities with real or simulated professional contexts. Second, the level required to meet some of the quality indicators was lowered. This was the case with students' working in groups, their opportunities to select some aspects of the learning environment, and their participation in the course.

After modifying the statements in the rubric according to the experts' suggestions, we prepared a second questionnaire containing 52 closed items, also mandatory, plus the open-ended text field for optional response. In the second questionnaire, the items were grouped into the same four components. The component on grammar contained five items that analysed the ambiguity, length, complexity, specialisation and bias of the statements in the rubric. The component focusing on the consistency of the rubric contained 12 items that assessed the rubric's fitness for purpose, for the educational context and for the discipline, how demanding it was in relation to the phenomenon evaluated and how difficult it was for the evaluator to use. The third component contained 11 items that analysed the structure of the rubric: the number and extent of the dimensions and performance levels and the organisation and internal consistency of the qualitative descriptions. Lastly, the fourth component on content contained 24 items on learning activities, the evaluation system and the role played by the student in the learning environments set out in course syllabuses. Table 1 shows a summary of the distribution of the items by component and the relative weight of each component in the questionnaire as a whole.

Table 1

Distribution of items by component

Component	Item	% of total
Grammar	1*, 12, 18*, 27, 52	9.61
Consistency	2*, 10, 11*, 17, 19, 26*, 28, 29*, 35, 39, 50, 51*	23.07
Structure	3*, 13*, 16, 25, 30*, 38, 41, 43, 45, 47, 49	21.15
Content	4*, 5, 6*, 7, 8, 9*, 14, 15*, 20, 21*, 22*, 23, 24, 31*, 32*, 33, 34, 36*, 37*, 40*, 42, 44*, 46, 48*	46.15

Note. *Items also included in the first round questionnaire.

The second questionnaire had a larger number of items because the aim in the second round was threefold: a) determine if the problems identified by the experts in the first round have been fixed, b) clarify the agreement of the experts by adding more specific items, and c) analyse the consistency of the experts' responses. For this last reason, the items appeared randomly on the questionnaire in order to minimise the error of rating related items in the same way, and control items were added to identify any inconsistencies in the responses.

In the second round, one of the authors, acting as a facilitator, sent the experts the corresponding report with the results from the first round, the new version of the rubric and the new questionnaire. The results from the second round reached the established threshold of agreement as the Results section shows.

Data analysis

To determine the extent of agreement among the experts, the content validity index was used. This index measures whether a set of items – in this case, the quality indicators included in the rubric – provides a relevant representation of the construct in question, that is, the quality of course syllabuses. The distinctive feature of the present study is that the analysis of the rubric does not solely entail an assessment of the extent to which each quality indicator is pertinent. It also requires assessments of the relationship between each quality indicator and

other indicators that are part of the same qualitative description, of whether there is a clear distinction between contiguous performance levels and of whether there is a comparable level of demand across homonymous performance levels. Consequently, these aspects are examined by the component on structure, whereas the components on consistency and content address whether the rubric has an appropriate sample of quality indicators. For the same reason, the content validity index for the items was calculated only to obtain the content validity indexes for each of the component scales because the latter indexes measure the extent of agreement among the experts on each of the rubric's determining aspects.

The universal agreement calculation method was rejected for the scale-level content validity index. Instead, the averaging calculation method was used to offset any increase in disagreement resulting from the participation of a large number of experts (see Polit & Beck, 2006). The content validity index for the overall scale was also calculated. In accordance with Davis (1992), the minimum for the scale-level content validity index was set at .80, taking into account: a) that the rubric involved a type of use not widespread yet; b) that Lynn (1986) set the acceptable item-level content validity index at .78 when there were between six and ten experts, and c) that Polit and Beck (2006) characterised a scale-level content validity index of .90 as excellent content validity.

In addition, a qualitative analysis was carried out on the feedback that the experts input into the optional open-ended text field. Of the 23 experts, 18 made observations. Phenomenography was chosen as the methodology for the analysis of their feedback for two reasons. First, it emphasises the role of the interaction of the actors who come into contact with a phenomenon in order to understand it; second, phenomenography fits with the purpose

of the present analysis: to sort the perspectives of the rubric among experts from different universities, disciplines and professional categories. As Marton stated in his seminal article:

"[...] phenomenography is a research method for mapping the qualitatively different ways in which people experience, conceptualize, perceive, and understand various aspects of, and phenomena in, the world around them. [...] Phenomenography is concerned with the *relations* that exist between human beings and the world around them." (Marton, 1986, p. 31)

Marton's definition synthesises the aspects that distinguish this methodology from other phenomenological approaches. This is particularly relevant because phenomenography shares two features with the other approaches: first, the social, dynamic and empirical nature of the methodology; and, second, a focus on the influence that the experiences and sense-making processes of a human group have on the understanding of a given phenomenon. However, the approach is different, and that matters.

Trigwell (2000) noted that phenomenography focuses on the variation of participants' perceptions and this focus differentiates the approach from phenomenological analyses; on the other hand, its interest in identifying the internal relationships among the variations distinguishes it from thematic analyses. The distinctive feature of phenomenography lies in the fact that it identifies and describes the perspectives with which a human group experiences a phenomenon, and interprets the internal relationships among those perspectives in order to explain the phenomenon.

In the context of this study, it is one thing to identify the judgements on the rubric that are shared among the experts (the first-order perspective typical of phenomenological analysis),

and it is another thing to identify the different positions from which they make those judgments and then interpret how their positions relate to one another (the second-order perspective typical of phenomenography). The first case gives priority to the phenomenon of the “rubric”, while the second case gives priority to "the experts at work", that is, to their positions at the time they judge the rubric (see Larsson & Holmström, 2007; Jobin & Turale, 2019). This perspective on how an expert interacts with the rubric is not a minor issue, given that the object of analysis is content validity and, to paraphrase Marton (1996), the expert does not act independently of the professional experience that he or she has had with rubrics.

Given the descriptive nature of the study, inductive coding was carried out in two steps: 1) the two authors of the paper analysed and coded the information separately; 2) both authors argued together over the different categories and coded meaning units with the aim of keeping the diversity of experiences and sorting the feedback into the *outcome space*. One of the key aspects of phenomenography, the outcome space is the result of the sorting of the experts' different experiences of the rubric by the researchers (see Marton & Booth, 1997).

Results

The extent of agreement among the experts

Table 2 shows the content validity indexes for items and the scale-level content validity indexes for each component and for the overall scale (research data available in [dataset] Author, 2020).

Table 2

Content validity indexes for items (I-CVI) and for scales (S-CVI)

Item	I-CVI	Item	I-CVI	Item	I-CVI	Item	I-CVI	Component (scale)	S-CVI
1	.85	14	.88	27	.87	40	.80	Grammar	.83
2	.99	15	.88	28	.95	41	.96	Consistency	.83
3	.97	16	.88	29	.68	42	.71	Structure	.90
4	.87	17	.71	30	.91	43	.87	Content	.87
5	.93	18	.76	31	.91	44	.88	Total Rubric	.87
6	.96	19	.86	32	.84	45	.83		
7	.99	20	.92	33	.74	46	.85		
8	.83	21	.88	34	.87	47	.96		
9	.80	22	.96	35	.96	48	.88		
10	.77	23	.88	36	.83	49	.97		
11	.87	24	.92	37	.90	50	.88		
12	.78	25	.68	38	.95	51	.93		
13	.97	26	.97	39	.79	52	.88		

As Table 2 above shows, only seven of the 52 items had item-level content validity indexes lower than .78, which is the threshold set by Lynn (1986) when there are between six and ten experts. Given that only 13.5% of the items were affected and that the present study featured more than double the number of experts indicated by Lynn, this outcome lends credibility to the analysis. Of the seven items, one was in the grammar component (20%), three in the consistency component (25%), one in the structure component (9.1%) and two in the content component (8.3%). According to Table 2, the experts identified a few mismatches between the rubric and the context in which it might be applied, and a few shortcomings in the wording of the quality indicators.

However, in the review of the scale-level content validity indexes, all the components came out far higher than the threshold of .80 set by Davis (1992) and the structure component even reached the value of .90 set by Polit and Beck (2006) for excellent content validity. Similarly,

the content validity index for the overall scale stood at .87, which represents the rubric as a whole.

Qualitative analysis on the experts' feedback

The teaching perspective.

As for the qualitative analysis, Table 3 below shows the categories obtained after coding (research data available in [dataset] Menéndez-Varela & Gregori-Giralt, 2020).

Table 3

Categories sorted by frequency

Category	Description	N	%
Resources	Resources to improve higher education.	13	13.27
Teaching improvement	The rubric promotes the improvement of teaching environments.	12	12.24
Reflection	The rubric fosters reflection in education.	10	10.20
Complexity of assessment	Difficulties in the assessment of student performance, curricula and teaching.	9	9.18
Agreement	The rubric encourages agreement between teachers and students.	9	9.18
Professional development	The rubric has an impact on the professional development of teachers.	8	8.16
Quality assurance	The rubric is useful for quality assurance in higher education institutions.	7	7.14
Institutional support	Institutional policies to support rubric-based assessment.	6	6.12
Competencies	The rubric aligns with competency-based higher education.	6	6.12
Teaching innovation	The rubric is related to teaching innovation.	4	4.08
Research and teaching	The rubric and the model of university research and teaching.	4	4.08

Accountability	The rubric fosters accountability.	3	3.06
Social impact	The social mission of the university.	2	2.04
University	Current university models and trends.	2	2.04
Pilot experiences	The rubric needs pilot experiences of teaching assessment.	1	1.02
Student evaluation of teaching	The rubric complements the Student Evaluation of Teaching.	1	1.02
Partnership	University partnerships with other bodies.	1	1.02
	Total	98	100

The above findings show that the feedback was predominantly to do with teaching issues. The rubric was perceived to be a useful resource (category: Resources) in tackling the task of evaluating curricula through the establishment of a shared view of educational quality (categories: Reflection and Agreement). For example:

"The truth is that [...] it seems like a good instrument to me, it's very comprehensive and clear, and it can also be used for evaluation and for moulding teachers by providing better teaching guidelines."[1] (Participant 5)

The use of the rubric for the assessment of course syllabuses was repeatedly combined with its use in student performance assessment, very possibly because this is the use the experts were most familiar with; on several occasions they stressed that it was also beneficial for clarifying teaching environments and improving communication between teachers and students (category: Agreement).

[1] The experts' feedback has been translated into English. The original version in Spanish is available in [dataset] Author, 2020.

"I congratulate the team that designed the rubric. Any teacher knows how hard it is to give an appropriate assessment of students' achievements, and doing so is an important factor in ensuring that students feel safe. Rubrics motivate students when they see that nobody is hiding any cards up their sleeve. The commitment to transparency that goes together with the use of rubrics plays an informational role, but it also helps to build a positive emotional state that makes it possible for students and teachers to come together more honestly." (Participant 12)

In this general context of educational assessment, experts offered a good deal of feedback on the difficulty of conducting evaluations in a manner that was valid, reliable and educationally beneficial (category: Complexity of assessment). An expert said:

"[I]t seems to me that teachers today, rather naively, have learned about the notion of styles as if they were watertight compartments, so that we are either kinaesthetic or visual or verbal and so on. I believe there is a whole debate on the issue, but you [researchers] made an appropriate choice because it is necessary to promote all of those styles of learning. Another construct, learning approaches (whether the approach to learning is superficial, strategic or deep), could give rise to the question of whether complex learning is being achieved, at a high level, close to what [students] will face in the professional world [...]" (Participant 5)

The rubric was associated with improvement in teaching environments (category: Teaching improvement) and with the training, responsibility and recognition of faculty members (category: Professional development), given the interdependence that exists between these two aspects. For example:

"There should be greater recognition of the fact that faculty members work in teams to make the decisions needed to design rubrics or any other new assessment instruments. The same could be said about their impact on the professional growth of teachers." (Participant 13)

On one occasion, an expert presented the rubric as a tool to empower faculty precisely in their teaching activity, which should, in the expert's view, constitute the core of their professional practice (category: Accountability). This reassertion of teaching was part of a rationale that included the need to improve higher education, to relieve the growing pressure on faculty, and to counteract the deterioration of their employment conditions.

"Evaluation is one of the keys because of its resonance within the bodies that are responsible for ensuring (or surveilling, as Foucault would say) the quality of our degree programmes and our work both within and outside the university. Especially if you accept the proposition (as I do) that the job of university faculty is first and foremost education [...]. Rubrics are a good example of empowering the teaching professional in his or her own work environment [...]. Expert teachers that have their own ideas about educational improvement and innovation are in a position to demand that their voice be heard when other institutions propose regulatory measures that have an effect on their direct area of activity and even on their conditions of employment." (Participant 13)

The experts associated the rubric with two other ideas that were also mentioned. The first was their acknowledgement that rubrics were used little in higher education (category: Teaching innovation), while the second was their conviction that the situation would not change without institutional policies to provide technical and economic support as well as recognition for the faculty involved (category: Institutional support). Furthermore, if this is

the situation in student performance assessment, their use in the evaluation of curricula or course syllabuses would face even more obstacles. For the Teaching innovation category:

"[R]ubrics are not very common in university education. As a result, it is necessary to carry out a training session beforehand to serve at least as an initial point of contact and to provide guidance for a rubric's application [...]." (Participant 1)

For the Institutional support category:

"[Rubrics] are rare in the university and what's lacking is greater support from rector's offices and dean's offices." (Participant 11)

The managerial perspective.

From all of the above, it does not necessarily follow that the experts' feedback was homogeneous. Table 4 below shows the same categories sorted according to the different approaches identified by the researchers. While most of the experts gave feedback consistent with a single perspective, it was also possible to find a juxtaposition of different viewpoints.

Table 4

Categories sorted by dimension

Category	Dimension					
	Teaching		Management		Critique	
	n	%	n	%	n	%
Resources	9	9.18	2	2.04	2	2.04
Teaching improvement	12	12.24				
Reflection	6	6.12	3	3.06	1	1.02
Complexity of assessment	9	9.18				

Agreement	7	7.14			2	2.04
Professional development	8	8.16				
Quality assurance			6	6.12	1	1.02
Institutional support	6	6.12				
Competencies			3	3.06	3	3.06
Teaching innovation	4	4.08				
Research and teaching					4	4.08
Accountability	2	2.04	1	1.02		
Social impact					2	2.04
University					2	2.04
Pilot experiences			1	1.02		
Student evaluation of teaching			1	1.02		
Partnership					1	1.02

Ranked below the teaching perspective discussed above were two additional viewpoints that received almost the same number of mentions. On one hand, the experts put forward a managerial viewpoint in which the rubric appears primarily to be a resource that may prove useful in pursuing quality assurance policies (categories: Reflection and Quality assurance). They argued that many universities today find themselves at a midway point in their implementation of competency-based higher education (category: Competencies) and that national quality agencies are driving new evaluation programmes that extend accountability to different academic units.

"The general impression is that competency-based teaching is not deeply entrenched, possibly because it has involved a 180-degree shift in teaching traditions and the university lacks the time to assimilate the changes. [...] [T]he university is under a great deal of pressure from ANECA and AQU[2] to demonstrate how bachelor's programmes and master's

[2] Acronyms of the Spanish and Catalan quality agencies.

programmes are working. Another part of this reality is that the policies and programmes of these agencies will ultimately have an effect on smaller university units." (Participant 15)

One expert remarked that the rubric-based assessment systems for curricula adds a further alternative to teaching evaluations based on student surveys, which enjoy little acceptance among faculty (category: Student Evaluation of Teaching).

"A rubric like this one could be interesting for evaluation, but perhaps even more importantly for reflecting on the quality of curricula. That is why rubrics should be used as a complement to the yearly surveys that students fill out. This type of rubric solves many of the problems for which student surveys have been criticised [...]." (Participant 14)

However, the expert also stressed the need to verify the benefits of this specific use of the rubric (category: Pilot experiences) before a university could decide to incorporate it within its educational quality system (category: Accountability).

"In my opinion, the use of rubrics along these lines is very new. So there is a need for in-depth studies on their proper function and viability, that is, whether they can be used by teachers en masse without colossal efforts. The academic authorities must be very mindful of the fact that teachers are fed up with all of their administrative tasks [...]." (Participant 14)

Also reappearing under this dimension was the crucial matter of the resources needed to tackle new challenges. When the Resources category was mentioned on this dimension, there was an emphasis on: a) the need for new evaluation systems that would be adapted to the distinctive features of each faculty or department, b) the need to review the funding

programmes for faculties and departments in order to give greater importance to the robustness of their internal quality assurance systems, and c) the need for specific training for faculty. For example:

"This model of rubric should occupy a place among the instruments that universities possess to ensure the quality of their curricula and to review the funding programmes for academic units, because it has the potential to be adapted for different contexts. Teachers believe that universities have accreditation processes that are insufficient for determining how competency-based higher education is working, and this rubric is fully focused on thinking through the fundamental characteristics of that approach." (Participant 17)

"The key is to strengthen the prestige of education in society, bolster the improvement and innovation of teaching with quality projects so that the university community has models, and enhance the in-depth, ongoing training of faculty members from their moment of joining through their entire academic career." (Participant 15)

The critical perspective.

On the other hand, the dimension labelled Critique included feedback that offered a more critical view of the university model. The Competencies category grouped doubts over the suitability of competency-based higher education because of the risk of subordinating the university's educational purpose to purely economic interests, which were regarded as contrary to the very idea of a university. One expert said:

"There are reasonable doubts about competency as a concept and therefore about an education grounded in competencies. [...] The advent of competencies, largely linked in

Europe to the Bologna Process, could have acted as a lever to spur thinking about a new educational approach. However, it was a missed opportunity because the competency model rapidly proved susceptible to an economicist conception, as shown for example by the OECD's intrusion into the educational sector." (Participant 18)

This distrust of the concept of competency was perhaps prompted by the origin of the term in the field of human resources development in the private sector. This same expert confessed some unease at the direction of universities today, at the same time acknowledging that it was necessary to deepen the synergies between universities and the private sector, provided that their increasingly close partnership reflected a mutual recognition of the two parties on equal terms (category: Partnership). This expert, however, did not take a positive view of the effects resulting from the debate over the relationship between basic and applied research in the university (category: Research and teaching).

"[T]he university was asked to square the circle. On one hand, through the concept of transfer, it was required to make a greater effort in the technological application of funded projects. On the other hand, the industrial sector exerted pressure on the university to take over the bulk of basic research so that it would be developed with public moneys and industry could thus avoid an amount of funding that was as large as it was essential. The necessary collaboration between the public sector […] and the private sector is not something to criticise, but the absence of a more equitable scenario of efforts and benefits is." (Participant 18)

Experts also mentioned the importance of enhancing the social impact of the university, not only because this is one of its missions, but also as a necessary counterpoint to the rise of

managerialism in university governance systems (category: Social impact). Elsewhere, an expert from the area of the Arts and Humanities expressed concern that the educational function was often forgotten to be one of the main elements of the social impact of universities.

"There is still a long way to go before teaching and the social responsibility of the university enjoy the same recognition and resources that research activity receives. […] Just as the social impact of research is undeniable, so too is the educational function performed by the university. […] In addition, the social responsibility of the public university perhaps begins by offering quality educational programmes for free." (Participant 16)

From this more critical viewpoint, there was consensus that the university needs to develop and deepen the new educational paradigm in order to help young people to become better-prepared individuals and more active citizens (category: Research and teaching):

"[H]owever much talk there is, teaching improvement is not really at the top of the agenda [...]." (Participant 11) As a consequence, "the cornerstone is for the competencies analysed by the rubric to be redefined conceptually so that education serves the public interest and social well-being." (Participant 18).

The proposed rubric and rubrics in general are helpful in the dialogue among teachers to clarify the fundamental concepts and ideas of their professional milieu (categories: Reflection and Agreement). However, rubrics are only one aspect among a host of factors that must be considered holistically:

"Rubrics have gained ground in teaching innovation, but you cannot take them as a panacea. There is still a long way to go: there are many poorly designed rubrics, rubrics alone don't drive a greater understanding by students of what is really important, the effort that is required for their preparation, among other issues." (Participant 16)

From the analysis of the distribution of the percentages for each category (Table 4), it is possible to identify the structural relationship that exists among the three descriptive dimensions that comprise the outcome space, and this is a crucial aspect of phenomenography. First, the Teaching dimension concentrated 64.26% of the total frequencies of the categories. This figure confirms that it was the dominant perspective among the experts, far ahead of the Critique dimension (18.36%) and the Management dimension (17.34%). Second, the Teaching and Critique dimensions shared the same number of associated categories (nine) and almost the same number of exclusive categories (five and four respectively). This suggests that the two dimensions represent constructs of quite similar conceptual richness. The third matter concerns the degree to which each dimension interacts with the outcome space. The Teaching and Management dimensions share two categories, Management and Critique share four, and Teaching and Critique share three. Therefore, Critique is the most interconnected dimension, followed by Teaching and Management.

Discussion

From the perspective of unified construct validity theory, both empirical evidence and theoretical rationales are required to attribute validity consistently. And the content aspect of validity is the sole facet that directly examines the degree to which an assessment system consists only of the most relevant and representative components of the construct at issue and is not undermined by other irrelevant elements. In addition to the previously cited studies by

Messick, Cronbach (1989) emphasises the importance of the theoretical dimension when defining the construct in a strong programme of validation. The consistency of the theoretical rationales and the scope of application of an assessment system can only be determined by means of an analysis of the specialised literature and the judgement of experts. The latter is even more important when the literature available is scarce, which is the case with the introduction of new assessment systems.

The structural and external aspects of validity only cover the matter indirectly. An analysis of the consistency of the structure of the scores obtained and the construct itself requires that the experts have previously defined the construct, while the structure itself is highly dependent on the case at issue. A comparison of the results of the assessment with the results obtained using other valid and reliable assessment systems depends on the similarity of their respective objects of analysis. At the same time, it should also be asked what these alternative assessment systems for course syllabuses entail.

In spite of the importance of any study focusing exclusively on the content aspect of validity in the assessment of the curriculum design of degree programmes, it is rare to find a detailed explanation of the method and results obtained. In their study of a rubric-based assessment system for the alignment of assessment practices to international criteria in higher education, Alonzo et al. (2018) set out a robust analysis of the structural aspect of validity, but offered very little information on the process of content validity in which eight researchers with expertise in assessment took part. Similarly, in his study of a rubric-based meta-assessment system at one university, Orem (2012) grounded the content aspect of validity in a review of the specialised literature and examined the structural aspect and reliability. Koh (2013) focused her analysis of a rubric for assessing meaningful learning with ICT on a study of

reliability and agreement, and merely noted the participation of two faculties in the expert review of content validity. Although Halim (2008) examined the structural aspect of validity and reliability in a rubric-based assessment of teaching performance, his study, like Koh's, made only a passing reference to the participation of a panel of experts and the refinement of the rubric in response to their suggestions.

The sole exception is the study conducted by Stanford et al. (2016) on a rubric-based assessment system to examine the propagation of education innovations. Their paper presents a detailed description and rationale for the rubric and data collection procedures in the examination of content validity, but does not do so for the data analysis or the findings. The same tendency can be observed in research on the use of rubrics to assess student performance. With the exception of the study carried out by Alsina et al. (2017), the area is dominated by studies that address the structural and external aspects of validity in greater or lesser detail, but their treatment of the content aspect is very much secondary (e.g., Allen & Knight, 2009; Baily, Ryan, Astolfi, & Pollock, 2017; Docktor et al., 2016; Erlich & Russ-Eft, 2012; Simper, 2018; Thaler, Kazemi, & Huscher, 2009).

Consequently, one of the main contributions of the present paper is the presentation of a mixed method for the analysis of the content aspect of validity for rubric-based assessment systems. On the quantitative dimension, it is important to highlight the estimation of the extent of agreement among experts using the content validity index, which is a method that is not widely used in the field of Education Sciences. In relation to the qualitative dimension of the model, the authors are likewise unaware of any examples featuring the use of phenomenography to study the perceptions of the users of rubrics or the experts that have taken part in their validation.

The content aspect of validity

The aspects of the rubric that raised problems were, first, its fitness for purpose and for the educational context and/or discipline (consistency component) and, second, the wording of the qualitative descriptions (grammar component), whereas the mismatch of the other two components (structure and content) did not exceed 9%. However, these problems were minor, as is shown by the fact that the content validity indexes for scales were greater than .80 (Table 2). The study concurs with previous research on the validity of rubric-based assessment systems in maintaining that the wording is the most recurrent shortcoming (e.g., Alonzo et al., 2018; Alsina et al., 2017; Stanford et al., 2016; Young, James, & Noy, 2016). While there is always a risk of errors in spelling or syntax, the present study suggests that the problem is more complex.

The only questionnaire item included in the grammar component that did not meet the acceptable item-level content validity index set by Lynn (1986) was item 18 (.76, Table 2), which asked experts about the presence of technical terms in the rubric. The second lowest rated item in the grammar component was item 12, which asked about the presence of ambiguous wording (.78, Table 2). The hypothesis is that these two aspects, which concern the clarity of wording, could be related to the perception of certain shortcomings in the rubric arising from the variable presence of rubrics – and especially of their use in the assessment of educational programmes – in the academic context of the experts. This hypothesis is supported by the fact that the consistency component received the lowest content validity index for scales together with the grammar component (.83, Table 2).

In other words, the clarity and precision of the wording of rubrics are also dependent on the traditions that define the context in which they are to be applied. This assertion is in line with the study conducted by Stanford et al. (2016), which saw a need to adapt the wording for individuals without a strong background in the theory underpinning the rubric. It is also related to the need to understand the details of the rubric in order to gain the most benefit from its use and not to confuse the complexity of the instrument with a lack of familiarity with it (see Young et al., 2016). In the case of the present study, there is no doubt that the different academic traditions of the experts who took part in the validation process had a negative effect on the two rubric components mentioned above – namely, consistency and grammar – and, for this reason, the extent of agreement in their evaluation takes on greater value.

The experts' perceptions of rubric-based assessment

In relation to the qualitative focus of the study, seven of the 18 experts that added feedback in the open-ended text field only offered technical comments on the rubric. There are three possible reasons for this result: the other experts might have understood that technical issues were covered by other questionnaire items, they might have had no pertinent objections and/or they might have used the text field for other types of reflections that were not covered by the questionnaire. The second possibility relates to the fact that the structure and content components, which bring together the items on technical aspects of the rubric, obtained the highest scale-level content validity indexes (.90 and .87, Table 2). The third possibility lies in the fact that the experts' feedback connected the rubric to university educational policy. Most of the feedback came from a teaching perspective, but even this specific feedback reflected this interest in educational policy.

Whereas the majority of the feedback concerned teaching (Teaching dimension, Table 4), there was no lack of feedback on the use of the rubric in academic management (Management dimension) or of reflections on various aspects of the transformation that universities are now undergoing and the trends that appear most likely (Critique dimension). Common to all of this feedback was the expression of a combination of commitment to educational improvement, doubts over the possibilities of achieving such improvement and concern about the improvement process that is now occurring. As a whole, the feedback reflects an awareness of the complex situation that the university in general faces today and of the transition in which its educational mission is now immersed — a transition that produces different effects and proceeds at different speeds across different national systems. The use of rubrics in student evaluation and curriculum evaluation is viewed as an academic practice that is recent, but that will become well-established in the near future.

The distribution of the various types of feedback into three dimensions shows that there are no rigid professional profiles in the academic area, but rather a plurality of emphases that individuals adopt depending on the perspective from which they analyse a phenomenon according to the circumstances. As a result of the difficulties in assessing a rubric from within an uncertain professional context but also, very likely, as a result of the conditions imposed by the data collection procedure and instrument, the experts shifted from one focus to another as they shared their feedback. By way of illustration, six of the 17 categories (or 48.96% of frequencies, see Table 4) appeared either across three dimensions (Resources and Reflection) or across two of them (Agreement, Quality assurance, Competencies, and Accountability).

Indeed, the category of Accountability exemplifies the experts' diverse and complex views of the rubric. Specifically, one expert, whose feedback was grouped under the Teaching

dimension, defended the rubric's potential to empower faculty in the exercise of their teaching activity, while another expert, who held a view coded under the Management dimension, maintained that the rubric could be useful in curriculum evaluation provided that pilot experiences demonstrated its effectiveness. These two examples reflect bottom-up and top-down models of accountability respectively, but both arguments concur in their positive assessment of the rubric's potential for educational improvement. As the feedback attests, the consolidation of a rubric-based assessment system for course syllabuses depends on the commitment of scholars, because they will be responsible for research into its benefits and for its widespread use. However, it also requires leadership from academic authorities to promote implementation across the entirety of the university.

The distribution of the various types of feedback shows by inference that the predominant academic activity for most of the experts is teaching and research or that the experts consider that these activities constitute the identity of academic staff. This appears to be proven by the difference in the frequencies between the Teaching dimension (64.26%) and the Management dimension (17.34%). Then there is the fact that the Critique dimension has presented a larger number of relationships with the other dimensions. From the diverse interaction among the three dimensions, it is apparent that the experts take a critical stance toward how the new educational paradigm is developing in universities and, even more so, toward the current university management policy.

The experts' comments point not only to an unease about the current state of universities, but also to an awareness of the complexity of finding appropriate solutions. Their statements reflect significant paradoxes rooted in the university institution. For example, faculty face growing pressure to improve educational quality even when there remains a need for teacher

training programmes that are more in line with the new educational paradigm and when the activity of teaching has not yet received the same recognition that research receives. Another examples lies in the lack of specific systems to evaluate the quality of degree programmes in a context in which departments and faculties face increased responsibility for their quality. A further example lies in the mismatch between, on one hand, the criteria and standards set by quality agencies in collaboration with the highest decision-making authorities in the university and, on the other hand, the reality that students and teachers actually encounter in departments and faculties.

While unease and concern can be observed among the experts, no clear signs of discouragement are evident. The Critique dimension itself shows an intellectual position taken in reaction to events. Of all the proposals that emerged in the experts' critical comments, the ones most directly related to this study assert the commitment of groups of teachers to teaching innovation. These communities of practice could serve as an ideal setting in which to make the voices of faculty members heard in the assessment of their own professional practice, and to do so concretely by proposing, discussing, agreeing on and implementing criteria, procedures and instruments of which this rubric is but one example.

Limitations and further research

The analysis of the content aspect of validity for this rubric-based assessment system brought together a number of experts that exceeded the range of seven to ten that has been common in similar studies. The only exceptions are the study conducted by Erlich and Russ-Eft (2012), which drew on 19 experts, and the study carried out by Stanford et al. (2016), which succeeded in involving over 70 participants in the validation process. Consequently, the first limitation does not lie in the size of the expert panel, but rather in the fact that the 23

participants represented only the Social Sciences and the Arts and Humanities. The collaboration of experts from other branches of knowledge would have strengthened the level of validity.

Given that the rubric's dimensions reflect the customary sections in course syllabuses and that the quality indicators cover aspects widely recognised in the literature, it can be argued that this rubric-based assessment system could be applied in other contexts too. Because of this limitation, however, it is not possible to analyse any adaptations that would be necessary in such contexts. Consequently, the recommendation is to verify the rubric's content validity through future studies that include experts from branches of knowledge that are not represented in the present study.

The second limitation is that the phenomenographic analysis could have been supplemented with another sort of analysis drawn from the perspective of Phenomenology in order to add an examination of the structure and meaning of the rubric to an analysis of the distinct ways of experiencing it. The third and main limitation is that the study focused on the content aspect of validity, which is only one of the facets of the unified construct validity concept. Consequently, it is necessary to supplement the study with research on the other five aspects of construct validity: substantive, structural, generalizability, external and consequential.

Two new studies are now underway to remedy the second and third limitations. One of these studies combines Phenomenography and Phenomenology methods after applying the think-aloud technique with a group of teachers and students in an evaluation of course syllabuses. The aim is to examine how these users understand and assess the rubric, and to analyse the substantive, external and consequential aspects of validity. In the second study, the structural,

generalizability and again consequential aspects of validity are the object of a quantitative study in which several teachers use the rubric to evaluate the course syllabuses of a sample of degree programmes in different branches of knowledge.

Conclusion

This paper presents a study focusing on a rubric-based assessment system for course syllabuses in order to mitigate two gaps existing in the literature on rubrics: their use in curriculum evaluation and the limited attention that has been given to an analysis of the content aspect of validity. As an added value, the research also tackles its object of study with methodologies that are not widely used in the Education Sciences: specifically by combining a quantitative analysis of the content validity index to find the extent of agreement among a broad panel of experts with a phenomenographic study of their different ways of judging the rubric's usefulness.

The study shows that the proposed system did yield evidence of its content validity. This finding, together with the fact that the experts came from different countries, universities and disciplines, could facilitate the use of the system in different national systems once it has been adapted to different university traditions. The value of the study is further heightened because at present only a small number of publications address the use of rubrics in the examination of the alignment between teaching practices and competency-based higher education, and even fewer publications focus on the assessment of learning environments overall as they are set out in course syllabuses. As the experts themselves indicated in their feedback, the system might also be useful from the perspectives of teaching innovation and quality assurance. As such, it may be of interest to teachers, academic authorities at different

levels of decision-making, and specialists in charge of curriculum design or the evaluation of educational programmes.

Funding

This work was supported by the Institute for Professional Development-ICE [REDICE18-1980]; the Vice-rectorate for Teaching and Academic Planning and the Programme for Research, Innovation and Improvement of Teaching and Learning at the University of Barcelona [GINDOC-UB/103].

Declarations of interest

None.

References

Allen, S., & Knight, J. (2009). A Method for Collaboratively Developing and Validating a Rubric. *International Journal for the Scholarship of Teaching and Learning*: *3*(2). https://doi.org/10.20429/ijsotl.2009.0302100

Alonzo, D., Mirriahi, N., & Davison, C. (2018). The Standards for Academics' Standards-Based Assessment Practices. *Assessment & Evaluation in Higher Education*, *44*(4), 636–652. https://doi.org/10.1080/02602938.2018.1521373

Alsina, A., Ayllón, S., Colomer, J., Fernández-Peña, R, Fullana, J., Pallisera, M., Pérez-Burriel, M., & Serra, L. (2017). Improving and evaluating reflective narratives: A rubric for higher education students. *Teaching and Teacher Education*, *63*, 148–158. http://dx.doi.org/10.1016/j.tate.2016.12.015

Álvarez-Pérez, P.R., González Morales, O., López-Aguilar, D., Peláez Alba, M.P., & Peña Vázquez, R. (2018). Criterios e instrumento para la valoración del modelo de Guía docente de la Universidad de La Laguna [Criteria and instrument for the evaluation of

the Teaching Guide model of the University of La Laguna]. In A. Vega Navarro & D. Stendardi (Eds.), *De la innovación imaginada a los procesos de cambio* (pp. 239–256). La Laguna: Servicio de Publicaciones de la Universidad de La Laguna.

Bachan, R. (2017). Grade Inflation in UK Higher Education. *Studies in Higher Education*, *42*(8), 1580–1600. https://doi.org/10.1080/03075079.2015.1019450

Baily, C., Ryan, Q.X., Astolfi, C., & Pollock, S.J. (2017). Conceptual assessment tool for advanced undergraduate electrodynamics. *Physical Review Physics Education Research*, *13*(2), 1–10. https://doi.org/10.1103/PhysRevPhysEducRes.13.020113

Barrie, S., Hughes, C., Crisp, G., & Bennison, A. (2014). *Assessing and Assuring Australian Graduate Learning Outcomes: Principles and Practices within and across Disciplines, Final Report*. Sydney: Office for Learning and Teaching.

Bergsmann, E., Klug, J., Burger, C., Först, N., & Spiel, C. (2018). The Competence Screening Questionnaire for Higher Education: Adaptable to the needs of a study programme. *Assessment & Evaluation in Higher Education*, *43*(6), 537–554. https://doi.org/10.1080/02602938.2017.1378617

Bers, T.H., Davis, B.D., & Taylor, B. (2000). The Use of Syllabi in Assessments: Unobtrusive Indicators and Tools for Faculty Development. *Assessment Update*, *12*(3), 4–7.

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. https://doi.org/10.1080/02602938.2012.658155

Blumberg, P. (2009). *Developing Learner-Centered Teaching: A Practical Guide for Faculty*. San Francisco, CA: Jossey-Bass.

Blumberg, P., & Pontiggia, L. (2011). Benchmarking the Degree of Implementation of Learner-Centered Approaches. *Innovative Higher Education*, *36*(3), 189–202. https://doi.org/10.1007/s10755-010-9168-2

Boyle, P., & Bowden, J.A. (1997). Educational Quality Assurance in Universities: an enhanced model. *Assessment & Evaluation in Higher Education*, *22*(2), 111–121. https://doi.org/10.1080/0260293970220202

Brancaccio-Taras, L., Pape-Lindstrom, P., Peteroy-Kelly, M., Aguirre, K., Awong-Taylor, J., Balser, T., Cahill, M., Frey, R.F., Jack, T., Kelrick, M., Marley, K., Miller, K.G., Osgood, M., Romano, S., Uzman, J.A., & Zhao, J. (2016). The PULSE Vision & Change Rubrics, Version 1.0: A Valid and Equitable Tool to Measure Transformation of Life Sciences Departments at All Institution Types. *Life Sciences Education*, *15*(4), 1–15. http://dx.doi.org/10.1187/cbe.15-12-0260

Brownell, S.E., & Tanner, K.D. (2012). Barriers to faculty pedagogical change: lack of training, time, incentives, and tensions with professional identity? *CBE–Life Sciences Education*, *11*, 339–346. https://doi.org/10.1187/cbe.12-09-0163

Caffarella, R. (2002). *Planning Programs for Adult Learners*. *A Practical Guide for Educators, Trainers, and Staff Developers* (2nd ed). San Francisco, CA: Jossey-Bass.

Carrión Martínez, J.J. (2006). La guía docente: ¿burocracia o reflexión?. In *I Jornadas sobre experiencias piloto de implantación del crédito europeo en las universidades andaluzas* [I Conference on pilot experiences of implementation of European credit in Andalusian universities] (pp. 647–650). Cádiz: Servicio de Publicaciones de la Universidad de Cádiz.

Chowdhury, F. (2018). Grade Inflation: Causes, Consequences and Cure. *Journal of Education and Learning*, 7(6), 86–92. https://doi.org/10.5539/jel.v7n6p86

Cronbach, L.J. (1989). Construct validation after thirty years. In R. L. Linn (Ed.), *Intelligence: Measurement, Theory, and Public Policy: Proceedings of a symposium in honor of Lloyd G. Humphreys* (pp. 147–171). Urbana, IL: University of Illinois Press.

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. https://doi.org/10.1080/02602930903540991

Cullen, R., & Harris, M. (2009). Assessing Learner-Centredness through Course Syllabi. *Assessment & Evaluation in Higher Education*, *34*(1), 115–125. https://doi.org/10.1080/02602930801956018

Davis, L.L. (1992). Instrument review: Getting the most from a panel of experts. *Applied Nursing Research*, *5*(4), 194–197. https://doi.org/10.1016/S0897-1897(05)80008-4

Dawson, P. (2017). Assessment rubrics: towards clearer and more replicable design, research and practice. *Assessment & Evaluation in Higher Education*, *42*(3), 347–360. https://doi.org/10.1080/02602938.2015.1111294

Docktor, J.L. (2009). *Development and Validation of a Physics Problem-Solving Assessment Rubric* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses Global.

Docktor, J.L., Dornfeld, J., Frodermann, E., Heller, K., Hsu, L., Jackson, K.A., Mason, A., Ryan, Q.X., & Yang, J. (2016). Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research*, *12*(1), 1–18. https://doi.org/10.1103/PhysRevPhysEducRes.12.010130

Ellis, R.A., Jarkey, N., Mahony, M.J., Peat, M., & Sheely, S. (2007). Managing quality improvement of eLearning in a large, campus-based university. *Quality Assurance in Education*, *15*(1), 9–23. https://doi.org/10.1108/09684880710723007

Erlich, R.J., & Russ-Eft, D.F. (2012). Assessing Academic Advising Outcomes Using Social Cognitive Theory: A Validity and Reliability Study. *NACADA Journal*, *32*(2), 68–84.

Finefter-Rosenbluh, I., & Levinson, M. (2015). What Is Wrong with Grade Inflation (If Anything)? *Philosophical Inquiry in Education*, *23*(1), 3–21.

Fink, S.B. (2012) The many purposes of course syllabi: which are essential and useful? *Syllabus*, *1*(1), 1–12.

García-Ros, R. (2011). Analysis and validation of a rubric to assess oral presentation skills in university contexts. *Electronic Journal of Research in Educational Psychology*, *9*(3), 1043–1062.

Gerretson, H., & Golson, E. (2005). Synopsis of the Use of Course-Embedded Assessment in a Medium Sized Public University's General Education Program. *The Journal of General Education*, *54*(2), 139–149. https://doi.org/10.1353/jge.2005.0020

Goldstein, N. (2010). *The program manager's guide to evaluation* (2nd ed.). Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

Good, J., Osborne, K., & Birchfield, K. (2012). Placing data in the hands of discipline-specific decision makers: Campus-wide writing program assessment. *Assessing Writing*, *17*(3), 140–149. https://doi.org/10.1016/j.asw.2012.02.003

Goodwin, A., Chittle, L, Dixon, J.C., & Andrews, D.M. (2018). Taking stock and effecting change: curriculum evaluation through a review of course syllabi. A*ssessment & Evaluation in Higher Education*, *43*(6), 855–866. https://doi.org/10.1080/02602938.2017.1412397

Grainger, P., Adie, L., & Weir, K. (2016). Quality assurance of assessment and moderation discourses involving sessional staff. *Assessment & Evaluation in Higher Education*, *41*(4), 548–559. https://doi.org/10.1080/02602938.2015.1030333

Grunert O'Brien, J., Millis, B.J., & Cohen, M.W. (2008). *The course syllabus: A learning-centered approach* (2nd ed.). San Francisco, CA: Jossey-Bass.

Guerra García, J.M. (2013). La burocracia: un factor limitante en la investigación [Bureaucracy: a limiting factor in research]. *Chronica naturae*, *3*, 7–10.

Halim, S.M.A. (2008). *The Effect of Using Some Professional Development Strategies on Improving the Teaching Performance of English Language Student Teacher at the Faculty of Education, Helwan University in the Light of Pre-Service Teacher Standards* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses Global.

Hixon, E., Barczyk, C., Buckenmeyer, J., & Feldman, L. (2011). Mentoring University Faculty to Become High Quality Online Educators: A Program Evaluation. *Online Journal of Distance Learning Administration*, *14*(5), 1–5.

Iudica, A.M. (2011). *University Educational Leadership Technology Course Syllabi Alignment with State and National Technology Standards* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses Global.

Jobin, P., & Turale, S. (2019). Choosing the right qualitative approach: Is Phenomenography a design for my study? *Pacific Rim International Journal of Nursing Research*, *23*(4), 314–319.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Koh, J.H.L. (2013). A Rubric for Assessing Teachers' Lesson Activities with Respect to TPACK for Meaningful Learning with ICT. *Australasian Journal of Educational Technology*, *29*(6), 887–900. https://doi.org/10.14742/ajet.228

Larsson, J. & Holmström, I. (2007) Phenomenographic or phenomenological analysis: does it matter? Examples from a study on anaesthesiologists' work. *International Journal of Qualitative Studies on Health and Well-being*, *2*(1), 55–64. https://doi.org/10.1080/17482620601068105

Legon, R. (2015). Measuring the Impact of the Quality Matters Rubric: A Discussion of Possibilities. *American Journal of Distance Education*, *29*(3), 166–173. https://doi.org/10.1080/08923647.2015.1058114

Lynn, M.R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382–386.

Marton, F. (1986). Phenomenography: A Research Approach to Investigating Different Understandings of Reality. *Journal of Thought*, *21*(3), 28–49.

Marton, F. (1996). Cognosco ergo sum – Reflections on reflection. In G. Dall'Alba, & B. Hasselgren (Eds.), *Reflections on Phenomenography: toward a Methodology?* (pp. 163–187). Göteborg: Acta Universitatis Gothoburgensis.

Marton, F., & Booth, S. (1997). *Learning and awareness.* Lawrence Erlbaum.

Mathers, C.E., Finney, S.J., & Hathcoat, J.D. (2018). Student Learning in Higher Education: A Longitudinal Analysis and Faculty Discussion. *Assessment & Evaluation in Higher Education*, *43*(8), 1211–1227. https://doi.org/10.1080/02602938.2018.1443202

[dataset] Menéndez-Varela, J.-L., & Gregori-Giralt, E. (2020). Content validity questionnaire. *Mendeley Data*, v1. doi:10.17632/7hr3swbh8c.1.

Menéndez-Varela, J.-L., & Gregori-Giralt, E. (2018). Rubrics for developing students' professional judgement: A study of sustainable assessment in arts education. *Studies in Educational Evaluation*, 58, 70–79. https://doi.org/10.1016/j.stueduc.2018.06.001

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: American Council on Education and MacMillan Publishing Company.

Messick, S. (1990). Validity of Test Interpretation and Use. Advance online publication. Retrieved from https://files.eric.ed.gov/fulltext/ED395031.pdf

Messick, S. (1995). Validity of Psychological Assessment: Validation of Inferences from Persons' Responses and Performances as Scientific Inquiry into Score Meaning. *American Psychologist*, *50*(9), 741–749. https://doi.org/10.1002/j.2333-8504.1994.tb01618.x

National Agency for Quality Assessment and Accreditation (2013). *Guía de apoyo para la redacción, puesta en práctica y evaluación de los resultados del aprendizaje* [Guide for the wording, implementation and evaluation of learning outcomes]. Madrid: ANECA.

O'Malley, G. (2010). Designing induction as professional learning community. *The Educational Forum*, *74*(4), 318–327. https://doi.org/10.1080/00131725.2010.483915

Orem, C.D. (2012). *Demonstrating Validity Evidence of Meta-Assessment Scores Using Generalizability Theory* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses Global.

Parkes, J. & Harris, M.B. (2002). The purpose of a syllabus. *College Teaching*, *50*(2), 55–61. https://doi.org/10.1080/87567550209595875

Perera, S., Babatunde, S.O., Zhou, L., Pearson, J., & Ekundayo, D. (2017). Competency Mapping Framework for Regulating Professionally Oriented Degree Programmes in

Higher Education. *Studies in Higher Education*, *42*(12), 2316–2342. https://doi.org/10.1080/03075079.2016.1143926

Polit, D.F, & Beck, C.T. (2006). The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Research in Nursing & Health*, *29*(5), 489–497. https://doi.org/10.1002/nur.20147

Raybon, S.P. (2012). *An Evaluation of Best Practices in Online Continuing Theological Education* (Doctoral dissertation). Retrieved from Proquest Dissertations and Theses Global.

Reddy, Y.M., & Andrade, H. (2010). A Review of Rubric Use in Higher Education. *Assessment & Evaluation in Higher Education*, *35*(4), 435–448. https://doi.org/10.1080/02602930902862859

Reed, T.E., Levin, J., & Malandra, G.H. (2011). Closing the Assessment Loop by Design. *Change: The Magazine of Higher Learning*, *43*(5), 44–52. https://doi.org/10.1080/00091383.2011.606396

Romkey, L., Chong, A., and El Gammal, L. (2015). Using Student Focus Groups to Support the Validation of Rubrics for Large Scale Undergraduate Independent Research Projects. *Proceedings 2015 the Canadian Engineering Education Association Conference* (pp. 1–8). https://doi.org/10.24908/pceea.v0i0.5778

Sefcik, L., Bedford, S., Czech, P., Smith, J., & Yorke, J. (2018). Embedding external referencing of standards into higher education: collaborative relationships are the key. *Assessment & Evaluation in Higher Education*, *43*(1), 45–57. https://doi.org/10.1080/02602938.2017.1278584

Simper, N. (2018). Rubric authoring tool for supporting the development and assessment of cognitive skills in higher education. *Teaching & Learning Inquiry*, *6*(1), 10–24. http://dx.doi.org/10.20343/teachlearninqu.6.1.3

Singham, M. (2007). Death to the syllabus! *Liberal Education*, *93*(4). Retrieved from https://www.aacu.org/publications-research/periodicals/death-syllabus

Slattery, J.M. & Carlson, J.F. (2005). Preparing an effective syllabus: Current best practices. *College Teaching*, *53*(4), 159–164. https://doi.org/10.3200/CTCH.53.4.159-164

Stanford, C., Cole, R., Froyd, J., Friedrichsen, D., Khatri, R., & Henderson, C. (2016). Supporting sustained adoption of education innovations: The designing for sustained adoption assessment instrument. *International Journal of STEM Education*, *3*(1), 1–13. https://doi.org/10.1186/s40594-016-0034-3

Stanny, C., Gonzalez, M., & McGowan, B. (2015). Assessing the Culture of Teaching and Learning through a Syllabus Review. *Assessment & Evaluation in Higher Education*, *40*(7), 898–913. https://doi.org/10.1080/02602938.2014.956684

Storey, V., & Asadoorian III, M. O. (2014). The Political Sense of Urgency for Educational Leadership Preparation Programs to Show Impact Data. *International Journal of Educational Leadership Preparation*, *9*(1), 1–13.

Sue, W. (2020). Exploring how rubric training influences students' assessment and awareness of interpreting. *Assessment & Evaluation in Higher Education*, *29*(2), 178–196. https://doi.org/10.1080/09658416.2020.1743713

Thaler, N., Kazemi, E., & Huscher, C. (2009). Developing a Rubric to Assess Student Learning Outcomes Using a Class Assignment. *Teaching of Psychology*, *36*(2), 113–116. https://doi.org/10.1080/00986280902739305

Thomas, D., Moore, R., Rundle, O., Emery, S., Greaves, R., te Riele, K., & Kowaluk, A. (2018). Elaborating a framework for communicating assessment aims in higher education. *Assessment & Evaluation in Higher Education*, *44*(4), 546–564. https://doi.org/10.1080/02602938.2018.1522615

Tractenberg, R.E., Umans, J.G., & McCarter, R.J. (2010). A Mastery Rubric: Guiding Curriculum Design, Admissions and Development of Course Objectives. *Assessment & Evaluation in Higher Education*, *35*(1), 17–35. https://doi.org/10.1080/02602930802474169

Trigwell, K. (2000). Phenomenography: variation and discernment. In C. Rust (Ed.), *Improving Student Learning: Improving Student Learning through the Disciplines. Proceedings of the 1999 7th International Symposium* (pp. 75–85). Oxford, UK: Oxford Centre for Staff and Learning Development, Oxford Brookes University.

Tucker, V.M. (2012). Listening for the Squeaky Wheel: Designing Distance Writing Program Assessment. *Online Journal of Distance Learning Administration*, *15*(4), 1–10.

Uribe, R.V. (2013). Measurement of Learning Outcomes in Higher Education. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Ed.), *Modeling and Measuring Competencies in Higher Education* (pp. 137–146). Rotterdam, The Netherlands: Sense.

Veltri, N.F., Webb, H.W., Matveev, A.G., & Zapatero, E.G. (2011). Curriculum Mapping as a Tool for Continuous Improvement of IS Curriculum. *Journal of Information Systems Education*, *22*(1), 31–42.

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. https://doi.org/10.1080/02602938.2012.661842

Ward, H.C., & Selvester, P.M. (2012). Faculty Learning Communities: Improving Teaching in Higher Education. *Educational Studies*, *38*(1), 111–121. https://doi.org/10.1080/03055698.2011.567029

Whitlock, B., & Ebrahimi, N. (2016). Beyond the Library: Using Multiple, Mixed Measures Simultaneously in a College-Wide Assessment of Information Literacy. *College & Research Libraries*, *77*(2), 236–262. https://doi.org/10.5860/crl.77.2.236

Wijngaards-de Meij, L., & Merx, S. (2018). Improving Curriculum Alignment and Achieving Learning Goals by Making the Curriculum Visible. *International Journal for Academic Development*, *23*(3), 219–231. https://doi.org/10.1080/1360144X.2018.1462187

Willingham-McLain, L. (2011). Using a University-wide Syllabus Study to Examine Learning Outcomes and Assessment. *Journal of Faculty Development*, *25*(1), 43–53.

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

Young, K., James, K., & Noy, S. (2016). Exploration of a reflective practice rubric. *Asia-Pacific Journal of Cooperative Education*, *17*(2), 135–147.

To cite this article: Gregori-Giralt, E. & Menéndez-Varela, J.-L. (2021). The content aspect of validity in a rubric-based assessment system for course syllabuses. *Studies in Educational Evaluation*, *68*, 100971. https://doi.org/10.1016/j.stueduc.2020.100971