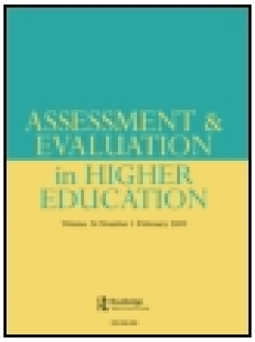

Assessment & Evaluation in Higher Education

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caeh20

Disclosing own reasoning while appraising the students' reasoning: implications for developments in formative assessment in science-engineering education

Mariana Orozco

To cite this article: Mariana Orozco (2023): Disclosing own reasoning while appraising the students' reasoning: implications for developments in formative assessment in science-engineering education, Assessment & Evaluation in Higher Education, DOI: 10.1080/02602938.2023.2196008

To link to this article: https://doi.org/10.1080/02602938.2023.2196008

Published online: 03 Apr 2023.

ASSESSMENT & EVALUATION IN HIGHER EDUCATION
https://doi.org/10.1080/02602938.2023.2196008

Routledge
Taylor & Francis Group

OPEN ACCESS
Check for updates

Disclosing own reasoning while appraising the students' reasoning: implications for developments in formative assessment in science-engineering education

Mariana Orozco

University of Twente, Enschede, The Netherlands

ABSTRACT

When instructors assess students' laboratory reports to appraise the underlying scientific reasoning, they disclose their own concerns, epistemological assumptions and beliefs about science. The analysis of such assessments (i.e. rubric-centred scores and corresponding justificatory comments) offer a wealth of insights that can be re-engaged in further improvements of the assessment tool and procedure, and in developments in formative assessment more generally. Such insights include concerns exceeding the rubric's descriptions (about meaningfulness, exhaustiveness, implicitness, connectivity, true inquiry, relevance), while differences among assessors are exposed (regarding epistemic values, approaches to scoring, sensitivity). This contribution is part of a broader effort to promote students' conducive scientific thinking and deep-learning in science and engineering education. It addresses the question(s): what does the assessors' reasoning tell us about the ways in which formative assessment is conducted, and could ideally be? The empirical investigation connects to existing knowledge, and discusses issues of representativeness and granularity in formative assessment. The paper elaborates on the design and use of the assessment tool, and presents evidence supporting context-bound recommendations and general conclusions. It is proposed that developments in formative assessment will benefit from reconceptualisation of assessment criteria, as the result of a co-design activity that engages with the assessors' epistemological concerns.

KEYWORDS

Students' scientific reasoning; formative assessment using rubrics; science-engineering education; epistemological concerns

Introduction

Research on formative assessment has shown that 'explicit' assessment criteria can be beneficial for students in various ways, while there is a risk that sharing detailed assessment criteria with the students may inhibit their learning (Andrade 2019). There seems to be controversy about making assessment criteria explicit (e.g. in the form of rubrics) and sharing them with the students, because it is not fully understood under which circumstances this practice is productive or not (Reddy and Andrade 2010). While some instructional designers and instructors are reluctant to use rubrics, others think that rubrics are the solution to all problems and, mistakenly, tend to use them as a pedagogical tool (Panadero and Jonsson 2013, 2020). It has been

CONTACT Mariana Orozco mariana.o@telenet.be; m.orozco@utwente.nl

proposed that the challenges posed by the use of rubrics relate to the difficulties, or even impossibility, in making assessment criteria fully explicit. Although rubrics aim to be transparent, they remain opaque to a great extent (Bearman and Ajjawi 2021) due to, allegedly, issues of representativeness and granularity. Such challenges appear even greater when the assessment regards not only students' mastery of content knowledge but also their reasoning skills.

When assessing students' scientific reasoning, assessment tools such as rubrics do not seem to fully capture the many criteria and issues that the assessors are concerned with. The grade resulting from the evaluative use of the rubrics will inevitably be informed by the assessors' preferences, styles and assumptions. When it comes to the formative use of the rubrics (e.g. for diagnosis and feedback), it can be expected that the resulting guidance offered by the instructors to the students responds just as much to the true students' needs, and to the assessors' epistemological perspective on such students' needs. These problems connect to the representativeness and granularity issues, and it appears that they cannot be straightforwardly resolved. Rather than trying to discover a quick-win, or to dissolve the problem through rhetorical manoeuvres, it seems more fruitful to embrace the assessors' epistemological concerns as they surface. This means to find some way to capitalise on the assessors' reasoning, which is informed by their epistemological beliefs, assumptions and concerns. To guide this endeavour, we raise the following (double) research question: what does the assessors' reasoning tell us about the ways in which formative assessment is conducted, and could ideally be?

To address this research question, we studied the assessors' reasoning at play during the evaluation of students' process and output of learning in a learning environment that is largely inquiry-based. More specifically, we conducted an investigation in the context of science and engineering education. This work is part of a broader research effort (Orozco, Boon, and Susarrey Arce 2022b) to promote students' conducive ways of scientific thinking—including conceptual modelling skill (Boon and Knuuttila 2009; Knuuttila and Boon 2011)—and deep-learning (Marton and Säljö 1997).

In brief, the present work concerns reasoning during assessment, and its implications for sustained developments in formative assessment. It connects to existing knowledge on formative assessment (Reddy and Andrade 2010; Panadero and Jonsson 2013, 2020; Ragupathi and Lee 2020), and engages in current debates on the use of rubrics (Bearman and Ajjawi 2021). Our work is informed by this body of literature and, concurrently, aims to contribute to it by broadening our problem formulation to encompass the representativeness and granularity issues.

This paper briefly introduces prior knowledge of the use of rubrics for (formative) assessment, and the topic of epistemological beliefs. It further contextualises our study, while elaborating on the process of co-design of the assessment tool (including the work on assessment literacy for the involved teachers) and the assessment procedure (including the training of the assessors). On this basis, we explain our analytical steps and present our extended empirical findings. The paper discusses implications for educational research (e.g. contribution to ongoing discussions on the formative use of rubrics) and practice (including both context-bounded and generalised recommendations for further developments in formative assessment).

Theoretical framework

The use of rubrics for formative assessment

Student-centred education focuses on learning as a process and demands progressive means of assessment. It is commonly accepted that students' sustained advancement is possible only when they understand what constitutes excellent/poor performance, and receive timely and qualitative feedback about the quality of their work. Increasing interest and empirical research (Reddy and Andrade 2010; Tai et al. 2018; Pastore and Andrade 2019; Andrade 2019; Lui and Andrade 2022) on formative assessment has advanced the use of rubrics in higher education, as they give students a better idea on what is assessed, on what criteria grades are based and what standards are expected (Ragupathi and Lee 2020).

A rubric is an assessment tool that operationalises learning objectives into criteria (or 'performance indicators') for student work, while articulating the levels of mastery for each criterion (in terms of the skills and knowledge that students are expected to employ at each level). Another feature of rubrics is a scoring strategy, i.e. a rating scale to interpret the assessors' judgments (Ragupathi and Lee 2020).

The preceding characterisation of rubrics presupposes that it is feasible to make performance criteria explicit, and that sharing detailed criteria with students is beneficial per se. While the practice of sharing such detailed criteria is often promoted (e.g. to enhance transparency and students' use of self-regulated learning strategies), it is not adequately understood under which circumstances this practice is actually conducive for student learning, or whether it may restrain learning by limiting students' autonomy and creativity (Reddy and Andrade 2010).

'Transparency' is often invoked to describe appropriate assessment criteria in higher education (Bearman and Ajjawi 2021). However, in an attempt to create and ensure transparency, rubrics are commonly taken as fair representations of the requirements for associated assessment tasks, thus fuelling the illusion that everything can, and should, be explicated (Bearman and Ajjawi 2021). In problematising assessment standards in higher education, it has been argued that a standard is best conceived as a dynamic and emergent performance that is enacted by the students, rather than as a truthful and static representation of performance (Ajjawi and Bearman 2018). This issue of representativeness connects to the issue of 'granularity'. The latter refers to the impossibility to achieve an ever finer grained account of expected performance. Making 'learning' or 'performance' explicit (formulating learning outcomes, performance indicators or rubrics) is necessarily a reductionist enterprise, as 'we simply cannot represent and codify the world in all its complexities in our accounts' (Säljö 2009, 204). As Säljö argues, there are several intellectual platforms (e.g. the social, the communicative and the biological) to conceive and describe reasoning and performance during a learning task. Any of those platforms/perspectives can be legitimately taken as foundational, and the descriptions of performance they generate are not reducible to each other. There is a critical conceptual distinction between 'learning' and 'performance' and, too often, measures of performance are an unreliable index of whether the learning has taken place (Soderstrom and Bjork 2015).

Although the transparency metaphor may be useful in some senses, it also has limitations that result in rubrics being treated as mere recipes, thus becoming instrumental, rather than supporting learning. The 'invitation metaphor' (Bearman and Ajjawi 2021) suggests a different way of conceptualising assessment criteria, by which a rubric is designed as an invitation to activity. This new perspective proposes a design of rubrics that conveys teachers' intentions, while encouraging students to develop their own ways of working and learning. It is in this sense that assessment criteria could be reconceptualised 'as enacted rather than represented' (Bearman and Ajjawi 2021, 360).

Epistemological beliefs and beliefs about the nature of science

Epistemological beliefs, often referred to as 'personal epistemology', concern people's assumptions about the nature and origins of knowledge (Hofer and Pintrich 2012). Several related terms are used in this regard, including epistemic beliefs (e.g. specific beliefs about the nature of knowledge and knowing), personal epistemology (e.g. a system of epistemic beliefs) and epistemic cognition (e.g. mental processes associated with knowledge) (Barger et al. 2016). Our work focuses on the cognitive-psychological aspects of epistemological beliefs (Hofer and Pintrich 1997, 2012; Bromme, Pieschl, and Stahl 2010; Schommer-Aikins et al. 2000), as we are interested in students' and instructors' assumptions of, e.g. the stability of knowledge, how knowledge is generated, what is new knowledge, as well as their arguments to invoke knowledge validity. Concurrently, we are interested in the implications of such sets of assumptions for students' learning and development.

Prior research has addressed questions of how epistemological beliefs relate to learning strategies (Muis, Bendixen, and Haerle 2006), to learning outcomes (Schommer 1993), to motivation (Barger et al. 2016) and to metacognition (Bromme, Pieschl, and Stahl 2010). Based on their empirical findings, Greene, Cartiff, and Duke (2018) argue that epistemic cognition (measured predominantly in terms of epistemological beliefs, as the ways in which people construct, justify and use knowledge) is an essential predictor of critical thinking and scientific literacy, among other learning outcomes. Based on empirical studies on university students' profiles (Lonka, Ketonen, and Vermunt 2021), it has been proposed that the epistemological beliefs students hold have consequences both for their studying practices and for their success in higher education. Epistemic beliefs have been identified as both domain-specific and domain-general (Muis, Bendixen, and Haerle 2006).

Beliefs about the 'nature of science' are a particular case within one's broader personal epistemology; these have received much attention in science and engineering education. Several studies document students' and teachers' significant misconceptions concerning the nature of science, while the evidence also suggests that understanding of the nature of science assists students in learning science content (McComas, Clough, and Almazroa 2002). For example, Songer and Linn (1991) found that students with dynamic views of science (i.e. seeing scientific knowledge as tentative) acquired a more integrated understanding than those with static views (i.e. seeing science as a group of facts). Several major issues related to the nature of science (i.e. the so-called 'myths of science') have been identified as most problematic in the experience of many science educators. McComas (2002) elaborates on these myths along with a call for science instruction to rethink its goals, for the focus to be on the nature of science itself rather than on facts and principles alone.

The context

This study is embedded in science and engineering education, in the field of chemical science engineering. In such context, future professionals are expected to develop research skills as much as engineering skills, while reasoning (or 'thinking skills') takes a central place in the curriculum. A new course was designed aiming to promote students' deep-learning of disciplinary knowledge, along with conducive ways of scientific thinking—in terms of a conceptual modelling skill that potentially enables drawing connections between physical phenomena and the theoretical concepts representing them (Knuuttila and Boon 2011). This course relied on inquiry-based learning principles, and approached the teaching of thinking in an explicit and content-related fashion (Orozco, Boon, and Susarrey Arce 2022a). It consisted of five practicums focussing on distinctive, yet often interrelated, electrochemical phenomena (e.g. redox reactions, ion diffusion, acid-base equilibrium). The students worked in groups of three on the preparatory work, the laboratory experience and the reflective work. In the preparatory work, they built an initial conceptual model of the phenomenon, formulated research questions and (ideally) designed an appropriate experiment that would allow them to answer their question. During the laboratory experience they run their experiments, made observations, collected relevant data and (ideally) performed on-the-spot preliminary analyses and interpretations. As part of their reflective work, they built the revised and extended conceptual model of the phenomenon, integrating prior knowledge, their empirical observations, their interpretation of the collected data and (ideally) drawing meta conclusions about the phenomenon and about their new knowledge of the phenomenon.

The conceptual modelling activity was supported by the Boon and Knuuttila method (2009), as a cognitive scaffold for the learning and persistent use of conceptual modelling reasoning.

The overall learning process was facilitated by tutors (learning assistants), i.e. more advanced students who had a formative role during the series of laboratory practicums, and who were later involved in this piece of research as assessors.

Methods

How the assessment tool was built

Given that the course did not have a systematic assessment procedure in place, we had to create one (certainly if we also aimed to compare the attributable effects of changes in implementation on the students' learning process and outcomes). The assessment tool was built by the teacher and researcher in collaboration, based on the existing 'intended learning outcomes' (i.e. four ILOs and their breakdown, adapted to the specificities of each laboratory practicum). We decided to use a small, even number of categories, and to describe them both in qualitative and quantitative terms. Next to the conceptual work, we fine-tuned the descriptions using randomly selected students' reports. The resulting assessment tool took the form of rubrics with four categories representing levels of performance: 'outstanding', 'meeting the expectations', 'not meeting the expectations' and 'very poor'. Each category had a description and an associated numeric range of scores.

Some expected limitations for the further use of these rubrics were noticed while building it. As part of the validation of the tool we appealed to replication of assessments by two independent assessors.

Procedure for data collection

A group of six tutors (also called the 'assessors') were trained on how to use the rubrics to grade a sample of laboratory journals and reports. They received instructions to assess the students' documents in the light of the rubrics, by giving scores and a qualitative appraisal that would justify the scores, while noting their struggles in using the tool and suggestions for its further revision.

We collected the completed rubrics for two selected practicums: i.e. the forms containing scores and the comments by the assessors, as well as students' laboratory journals and reports that had been annotated by the assessors. Approximately 30% of the reports were assessed twice by independent assessors, so to compare the appraisals and investigate the meaning of any differences. In total 24 documents were assessed, i.e. laboratory journals and reports for the first and second years of implementation of this course, for which we had obtained informed consent from the students.

We recorded periodic meetings with the tutors (i.e. exchange of perceptions and perspectives, held after each practicum). These data sought to contribute better understanding of the tutors' views, allowing us to come up with a fuller interpretation and contextualisation of the research findings. Some assessors also participated in asynchronous member-check conversations in the form of guided further reflection.

Analytical steps

We mapped the quantitative data (mainly to make comparisons, yet without any statistical tests of significance or power, due to the small scale). Next, we proceeded to the thematic analysis of comments (in relation to the scores) to account for differences in the assessors' double scoring. We conducted a parallel thematic analysis (unrelated to the scores) to grasp the assessors' concerns and assumptions.

While performing the qualitative analysis, we came across several themes which we labelled 'assessors' concerns'. We found that the assessors were recurrently concerned about aspects of the reports they were evaluating that were non-obvious part of the rubrics. The comments the assessors made while appraising the reasoning of students (mainly in students' reports, but also in recalled conversations) disclosed the assessors' reasoning. Such findings shifted our attention from the students to the assessors, and we dived into the trails of reasoning left by the assessors in their comments when judging the reasoning of the students.

We analysed the additional research data (i.e. transcripts of meetings and in-depth member-check consultations) while searching for any support or contradiction to our interpretations. We extracted further elements enabling the contextualisation of our findings (e.g. expectations regarding the profile of a scientist/engineer).

Results

This section presents the results of the qualitative analysis. The point is made on the difference between literal and latent meanings of the rubrics. We present descriptions for thirteen themes that emerged from the analysis of the assessors' comments (while a collection of quotes illustrating the results is available on request). Next, we report results from the triangulation and extended analyses. Finally, we touch on the reproducibility of assessment in terms of the profile of the assessors.

Literal and latent meanings of rubrics

The comments provided by the assessors to justify their grading consist of judgements and considerations that were guided by the descriptions of the rubrics in the first place, while entailing a personal interpretation. For example, regarding an ILO that states 'Determine relevant concepts, formulate and defend your answer to the main scientific question(s)'; here one assessor would stay closer to a literal use of 'defending', while another one stretches its meaning to include exploration and conceptualisation.

Themes: the assessors' concerns

Expectation

Concern about too high expectations for the students, considering their stage in the programme. The description of the rubric leaves room for the assessors' own judgement on what, e.g. depth, thoroughness and/or complexity can be taken as sufficient.

Implicitness

Realisation that students' reports make (and can make) explicit only a part of the reasoning and the knowledge involved. Whenever there is a leap, the assessor needs to judge whether the students omitted some piece because they do not know, or because it is infeasible to explicate everything. Such judgement is often supported by other sources of information (either passages elsewhere in the same report, or external sources such as prior conversations with these students).

Exhaustiveness

Attention to the extent of completeness or coverage in relation to some whole or totality that may be obvious (e.g. answering all listed questions) or non-obvious (i.e. a 'full' explanation, the 'complete' phenomenon, 'all' prior knowledge, 'everything').

Connectivity

Concern about links (a) between the object of inquiry to prior knowledge, (b) between empirical observations and emergent research questions and (c) between main research question and sub-questions. The first, and most prominent, category mainly focuses on concepts addressed during previous laboratory practices and/or during lectures. The second category refers to explorative questions triggered by puzzling observations (in the realm of inquiry-based learning), but also predictions and hypotheses generated by observations and growing theoretical insights. The third category concerns a methodological issue of operationalisation. In all cases, the kind of link remains entirely unspecified by the rubric and by the assessor; we take these links to be inferential connections (e.g. what something follows from, what follows from something, what co-occurs, what are necessary/sufficient/contributing conditions for something to occur).

Vagueness

Concern about lack of precision, lack of depth or excessive implicitness (partly overlapping with a previously presented theme).

Labelling concepts and phenomena

Importance given to identification and correct labelling of known concepts and phenomena in electrochemistry, thermodynamics and chemical equilibrium. The attribution of a known term to an observed phenomenon is taken as an indicator of understanding and mastery of concepts. Most often, the assessors seem to conceive the naming or labelling of a concept as a crucial first step that needs to be followed by further connections (e.g. to other concepts). In a few cases the assessors fall into the trap of accepting the mere mention of a concept (i.e. the concept as bare verbal representation) as an indicator of comprehension.

Use of mathematical models

Attention to whether and how mathematical models are used to describe the electrochemical phenomenon. Concern about the connections of mathematical models to the phenomena under study, the interpretation of graphs and other results and the rationale for the selection of a particular model.

Research questions emerging from observations

Much attention to formulating a main scientific question that emerges from empirical observations of a natural phenomenon (that the students need to identify explicitly). The observations are expected to generate (research) questions that can further guide and promote the students' inquiry and learning. The students are expected to provide justifiable answers to these questions (after operationalisation). Such kind of observation-triggered research questions are taken to be crucial for enabling students to explore and to conceptualise. Other assessors' comments raise doubts about the need for explicit research questions, as long as a research goal is pursued.

True inquiry

Perspective (informed by the ILOs) that true scientific inquiry needs to start in some empirical observation(s), generate a main research question and use empirical findings to provide answers that include and surpass theoretical accounts. Concern about insufficient connections between the theoretical and the empirical, and about omission of one of them.

Relevance

Concern about students' uncritical selection of what to pay attention to, what to include (e.g. in their lists of observations and questions, mathematical models and story-like explanations) and what to weed-out.

Imagination and dealing with unknowns

Sensitivity to students' imagination and strategies to deal with unknowns (either due to unexpected empirical results or to affordances of the learning environment). The assessors acknowledged students' efforts to provide explanations, describe errors and suggest recommendations, and reformulate their research question.

Meaning

Concern about students' dealing with concepts, empirical data and results of calculations in a meaningful way. In the case of the meaning of concepts, the assessors attend to how concepts are used (e.g. described, explained, connected, illustrated) beyond mere mentioning. When it comes to the meaning of empirical data and results of calculations, often explanations and interpretations are expected from the students to evidence understanding.

Operationalisation

Informed by one of the course ILOs ['Identify sub-question(s) to provide an answer to the main scientific question(s)'], the assessors became sensitive to sub-questions, and to how well they 'cover' the main question. In methods terms, this is a step towards the operationalisation of the central question. Concurrently, some assessors raised doubts about sub-questions as a must, i.e. the students could skip the step of answering sub-questions as a means to synthesise/aggregate/induce the answer to the main question. Often, it remained unclear whether some students had actually skipped the step of sub-question answering, or had resorted to sub-questions implicitly.

Triangulation and extension

Both in the member-check and meetings with tutors, we found additional support for most themes, while we did not find any contradicting evidence suggesting misinterpretation. The meetings provided insights and raised questions about relevant topics, such as sequencing of learning activities, support to students (its extent, kind and timing) and 'exploration versus hypothesis-testing'.

In the member check, we found further depth in terms of specification, clarification and explanation. For example, why it is important to label phenomena (i.e. to avoid confusion between the phenomenon concerned and a measurable variable representing only one aspect of the phenomenon), and that the omission of such labelling relates to a poor understanding or lack of depth. The assessors also had difficulties in labelling phenomena, and were inconsistent in their distinction between a phenomenon and a 'parameter'. Hypotheses were valued for their (inferential) connecting power between 'the phenomenon' and 'the actual result' (probably referring to the empirical observations). Although explorative and hypothesis-testing research questions were considered useful/instructive in different ways, a hypothesis-containing question (based on the empirical observations) was preferred for enabling students to show their ability to design an experiment that is appropriate for testing.

The profile of the assessors

From a comparison of the replicated assessments, it appeared that some assessors are more rigorous while remaining nuanced, others tend to give overall quite the same scores regardless of the ILO, and yet others are ready to give extreme scores to particular ILOs within the same assessment. The preferences of the assessors may explain part of the variability in grades.

The meaning and value attributed to the rubrics' qualitative categories varied from one assessor to the other. Often quite different scores were given to the same item, while the justification read nearly the same. Assessors appeared to have implicit (personal) criteria in mind, e.g. being more lenient in the grading to acknowledge effort or difficult circumstances of a student or a group. Exceptionally, reflective assessors seemed able to make (some of) these criteria explicit.

Discussion

We started this article by proposing that we need to capture the epistemological concerns of instructors and assessors in assessment procedures that are used with a formative purpose in science and engineering education. In aiming to understand and capitalise on the assessors' reasoning, we raised the question of what this tells us about the ways in which formative assessment is conducted. Such inquiry implies a further aim to advance how a formative assessment approach, that attends to epistemological assumptions and beliefs, could ideally be.

What the assessors' reasoning tells

Our findings revealed that the assessors' reasoning was informed by the rubrics used as an assessment tool, as much as by their epistemological assumptions and beliefs about the nature of science. Thirteen themes emerged from the qualitative analysis of the assessors' comments; i.e. recurrent concerns that either went beyond the criteria described by the rubrics, or added further specificity.

Some of these themes related not only to concerns about the students and their deep learning of disciplinary content (Marton and Säljö 1997) but, probably even more, to concerns about students showing any deep learning. The assessors expected more explication, explicit (inferential) connections between concepts within/across topics, explicit (inferential) connections between empirical observations and theoretical knowledge, meaningful use of concepts (as well as empirical data, and results of calculations), exhaustiveness or coverage in relation to some (non-obvious) whole or totality, precision rather than vagueness and the appropriate labelling of concepts and phenomena. Other themes reflect the assessors' concerns about the students' scientific reasoning and epistemic cognition (Barger et al. 2016). We found recurrent considerations about relevance and students' (un)critical selection of what to pay attention to, imagination and strategies to deal with unexpected empirical results or any limitations, and what qualifies as true inquiry. The latter concern is the one that most prominently reveals the assessors' epistemological assumptions and beliefs (Hofer and Pintrich 2012), as well as their beliefs about the nature of science (McComas 2002).

The finding that the assessors (i.e. the tutors) also have some difficulties in labelling the natural phenomena under study points to a recurrent confusion (Orozco, Boon, and Susarrey Arce 2022a) of phenomena with methods, techniques and even materials that persists among students and instructors. This is an issue of classification that is worth attending to and that resonates with the conceptual modelling approach to promoting conducive ways of scientific thinking (Boon and Knuuttila 2009; Knuuttila and Boon 2011). The assessors' attempts to separate 'the phenomenon' from 'the actual result' suggest an issue of reification of the empirical observations as if they were 'facts' that are free from interpretation, and independent of the empirical methods used.

Despite our claims about the assessors' recurrent confusions, difficulties in classifying or inadvertent reification, their accounts witness sophisticated epistemological concerns on some occasions. They raise informed questions about the benefits and drawbacks of explorative versus hypothesis-testing research questions. While single hypotheses are valued for their connecting power (e.g. inferential connection between two concepts), and for enabling students to show their skill to design an appropriate experiment, hypotheses are also seen as limiting the students' sensibility to observe (in a more open and less biased way). However, in their (unspoken) preference for hypotheses, the deeper concern seems to be that assessors wish to understand what originates the students' question (and, while this is traceable in the observations-based hypothesis-containing question, this cannot be readily inferred from the open explorative question). We strongly suggest that this can be solved by requesting the students to formulate scenario-questions. Scenario-questions (or what-if-questions) are appropriate because they reveal genuine curiosity, as explorative questions also do, and they retain the benefits of hypothesis-testing (Zimmerman 2000) while enabling multiple 'rival hypotheses' or alternative explanations.

The assessors exhibited genuine interest for topics, such as: sequencing of learning activities; the extent, kind and timing of support (where rubrics-based formative assessment can play a role); and 'exploration versus hypothesis-testing'. These are not isolated topics; rather, they are connected by the more fundamental concern about how students reason and what the most appropriate way of reasoning is (e.g. any inductive, deductive, abductive or some kind of combination thereof).

In studying the discrepancies in scoring, we compared the comments made by assessors during replicated appraisal of the same documents. We found clear differences in focus and epistemic values (e.g. praising quality of connections between concepts above quantity of isolated concepts), approaches to scoring and its justification (e.g. analytical versus holistic, and literal versus interpretative), positionings (e.g. radical versus nuanced) and sensitivity (e.g. distinguishing added value from mere repetition or not, and noticing implicitness or not). There was quite a mismatch between scores and comments, which can be explained in terms of personal, idiosyncratic epistemic values and preferences in scoring (i.e. tendency to choose mid or extreme categories). Neither the assessors' personal preferences nor their personal implicit criteria may be significantly corrected by additional training on how to use this, or any other, assessment tool, or by increasing the internal validity of the tool. The latter solution would require 'sharp' descriptions of the rubrics, while we have argued that ever finer-grained explication of performance is an illusion and would be counterproductive.

Attending to epistemological concerns in formative assessment

We suggest that this work has implications for further developments of formative assessment. In line with existing knowledge and the call to shift from a transparency to an invitation metaphor (Bearman and Ajjawi 2021), we advocate for a conceptualisation of assessment criteria that, next to the teachers' and the students' expectations, embraces the tutors' concerns about deep learning, scientific reasoning and the nature of science. This entails a co-design activity of teachers with both tutors and the students, welcoming 'multiple enactments' of the rubrics, and allowing 'sophisticated ways of knowing'. For example, rather than assessing students on the exhaustiveness of an explanation, to appraise the skills used in the process of building such explanation. Or, rather than assessing the precision of a hypothesis, to value the explorative questions and how they generate unexpected knowledge.

Limitations and further (action) research

The rubrics built in the context of the present work had the immediate purpose of assisting our educational research, and the further purpose of serving as a mainly formative assessment tool (albeit subject to revisions) in subsequent implementations of the electrochemistry course. Particularly for the latter purpose, the use of the rubrics by a single assessor appraising a student's output carries the risk of yielding a too partial account of the student's growing reasoning skill. Indeed, if the aim is to appraise a student's scientific reasoning (in connection to electrochemical concepts) and provide them with quality feedback, their reasoning has to be observed where it happens (i.e. on multiple occasions using multiple sources), and by the instructor/assessor who has access to the process and/or output of such student reasoning (i.e. multiple observers using multiple sources). Conducting a multiple-assessor and multiple-source formative assessment is a demanding activity, and practitioners need to consider its feasibility in their particular contexts.

Further (action) research, as well as further revisions of the formative assessment procedure for actual implementation, need to investigate the use of the rubrics by multiple assessors who have exclusive/preferential access to multiple instances of the reasoning activity and the products of reasoning. This means not only the teacher assessing students' reports but also, e.g. tutors assessing students' exchanges during follow-up meetings, and laboratory assistants assessing students' activity in the laboratory. Finally, as an extension of the question of who is best placed and entitled to contribute to a more 'ecological learning' (Damsa and Jornet 2017) and 'ecological assessment', we suggest investigating how to introduce self-assessment and peer-assessment using the same rubrics. That is, to take into account past research on the topic of formative assessment, and further explore under which conditions it is beneficial to engage students as their own formative assessors, next to the more usual assessors.

Conclusion

This work has exposed a need to incorporate the epistemological concerns of instructors (and of everyone who is involved in the formative assessment of students) in the design and use of assessment tools that are intended to support students' learning. Considering the assessors' reasoning (deeply informed by epistemological assumptions, including beliefs about the nature of science) is to embrace the complexity and the challenges posed by any assessment and, in particular, by the formative assessment of students' simultaneous progress in scientific reasoning and in mastery of disciplinary knowledge in science and engineering education. Addressing the problem raised by this work matters. For example, it can be expected that the feedback students get will be more useful if the assessment procedure and criteria account for relevant epistemological concerns and, therefore, there is greater alignment to the learning objectives. This is certainly key in a learning environment seeking to promote students' conducive ways of scientific thinking. This means that we should strive to reason (about science and engineering) in the same way during assessment, as we expect students to reason during their electrochemistry course and beyond (rather than falling into 'myths of science'). The main implication for developments in formative assessment resides in a call for reconceptualisation of assessment criteria, meaning that such criteria are enacted, rather than represented, as the result of a co-design activity.

Some implications for educational practice in the area of science and engineering, and more generally, include recommendations for (re)design and implementation of (rubrics-based) formative assessment. A first suggestion is that educators exploit the power of rubrics for 'true formative assessment', rather than limiting them to a mere research tool or any other instrumental use. This can be realised by assisting students in becoming aware of their own epistemological assumptions, next to more commonly known interventions (such as providing timely and qualitative feedback, supporting the development of metacognitive skills, and promoting self-regulative learning). A further recommendation is to consider existing knowledge on the use of rubrics to be ahead of their potential drawbacks (e.g. falling into the 'criteria compliancy' trap), along with the idea to include multiple assessors and sources so that students' reasoning can be appraised where it happens.

Finally, we advocate that this work makes a novel contribution to educational research, in particular to ongoing discussions on the formative use of rubrics, by turning the representation and granularity issues into an opportunity to move towards more ecological and transformative ways of assessment of learning and for learning.

Ethical statement

This research project received approval (no. 210292) from the Ethical Committee of the University of Twente (BMS Faculty, Humanities and Social Sciences).

Disclosure statement

No relevant financial or non-financial competing interests.

ORCID

Mariana Orozco http://orcid.org/0000-0003-3101-0247

References

Ajjawi, R., and M. Bearman. 2018. "Problematising Standards: Representation or Performance?" In *Developing Evaluative Judgement in Higher Education*, edited by R. Ajjawi and P. Dawson, 41–50. Abingdon Oxon: Routledge.

Andrade, H. L. 2019. "A Critical Review of Research on Student Self-Assessment." *Frontiers in Education* 4 doi:10.3389/feduc.2019.00087.

Barger, M. M., S. V. Wormington, L. G. Huettel, and L. Linnenbrink-Garcia. 2016. "Developmental Changes in College Engineering Students' Personal Epistemology Profiles." *Learning and Individual Differences* 48: 1–8. doi:10.1016/j.lindif.2016.04.002.

Bearman, M., and R. Ajjawi. 2021. "Can a Rubric Do More than Be Transparent? Invitation as a New Metaphor for Assessment Criteria." *Studies in Higher Education* 46 (2): 359–368. doi:10.1080/03075079.2019.1637842.

Boon, M., and T. Knuuttila. 2009. "Models as Epistemic Tools in Engineering Sciences: A Pragmatic Approach." In *Philosophy of Technology and Engineering Sciences. Handbook of the Philosophy of Science*, edited by A. Meijers, vol. 9, 687–720. North-Holland: Elsevier.

Bromme, R., S. Pieschl, and E. Stahl. 2010. "Epistemological Beliefs Are Standards for Adaptive Learning: A Functional Theory about Epistemological Beliefs and Metacognition." *Metacognition and Learning* 5 (1): 7–26. doi:10.1007/s11409-009-9053-5.

Damsa, C., and A. Jornet. 2017. "Revisiting Learning in Higher Education–Framing Notions Redefined through an Ecological Perspective." *Front Learning Research* 4 (4): 39–47. doi:10.14786/flr.v4i4.208.

Greene, J. A., B. M. Cartiff, and R. F. Duke. 2018. "A Meta-Analytic Review of the Relationship between Epistemic Cognition and Academic Achievement." *Journal of Educational Psychology* 110 (8): 1084–1111. doi:10.1037/edu0000263.

Hofer, B. K., and P. R. Pintrich. 1997. "The Development of Epistemological Theories: Beliefs about Knowledge and Knowing and Their Relation to Learning." *Review of Educational Research* 67 (1): 88–140. doi:10.3102/00346543067001088.

Hofer, B. K., and P. R. Pintrich. 2012. *Personal Epistemology: The Psychology of Beliefs about Knowledge and Knowing*. New York: Routledge.

Knuuttila, T., and M. Boon. 2011. "How Do Models Give us Knowledge? The Case of Carnot's Ideal Heat Engine." *European Journal for Philosophy of Science* 1 (3): 309–334. doi:10.1007/s13194-011-0029-3.

Lonka, K., E. Ketonen, and J. D. Vermunt. 2021. "University Students' Epistemic Profiles, Conceptions of Learning, and Academic Performance." *Higher Education* 81 (4): 775–793. doi:10.1007/s10734-020-00575-6.

Lui, A. M., and H. L. Andrade. 2022. "Inside the Next Black Box: Examining Students' Responses to Teacher Feedback in a Formative Assessment Context." *Frontiers in Education* 7. doi:10.3389/feduc.2022.751549.

Marton, F., and R. Säljö. 1997. "Approaches to Learning." In *The Experience of Learning: Implications for Teaching and Studying in Higher Education*, 2nd ed., edited by F. Marton, D. Hounsell and N. Entwistle, 39–58. Edinburgh: Scottish Academic Press.

McComas, W. F. 2002. "The Principal Elements of the Nature of Science: Dispelling the Myths." In *The Nature of Science in Science Education. Rationales and Strategies*, edited by W. F. McComas. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers.

McComas, W. F., M. P. Clough, and H. Almazroa. 2002. "The Role and Character of the Nature of Science in Science Education." In *The Nature of Science in Science Education. Rationales and Strategies*, edited by W. F. McComas. New York, Boston, Dordrecht, London, Moscow: Kluwer Academic Publishers.

Muis, K. R., L. D. Bendixen, and F. C. Haerle. 2006. "Domain-Generality and Domain-Specificity in Personal Epistemology Research: Philosophical and Empirical Reflections in the Development of a Theoretical Framework." *Educational Psychology Review* 18 (1): 3–54. doi:10.1007/s10648-006-9003-6.

Orozco, M., M. Boon, and A. Susarrey Arce. 2022a. "Action Research on Electrochemistry Learning: Conceptual Modelling Intervention to Promote Disciplinary Understanding, Scientific Inquiry and Reasoning." Paper Presented at the SEFI 2022 Conference, Barcelona, September 19–22.

Orozco, M., M. Boon, and A. Susarrey Arce. 2022b. "Learning Electrochemistry through Scientific Inquiry. Conceptual Modelling as Learning Objective and as Scaffold." *European Journal of Engineering Education* 1–17. doi:10.1080/03043797.2022.2047894.

Panadero, E., and A. Jonsson. 2013. "The Use of Scoring Rubrics for Formative Assessment Purposes Revisited: A Review." *Educational Research Review* 9: 129–144. doi:10.1016/j.edurev.2013.01.002.

Panadero, E., and A. Jonsson. 2020. "A Critical Review of the Arguments against the Use of Rubrics." *Educational Research Review* 30: 100329. doi:10.1016/j.edurev.2020.100329.

Pastore, S., and H. L. Andrade. 2019. "Teacher Assessment Literacy: A Three-Dimensional Model." *Teaching and Teacher Education* 84: 128–138. doi:10.1016/j.tate.2019.05.003.

Ragupathi, K., and A. Lee. 2020. "Beyond Fairness and Consistency in Grading: The Role of Rubrics in Higher Education." In *Diversity and Inclusion in Global Higher Education*, edited by C. S. Sanger and N. W. Gleason, 73–95. Singapore: Palgrave Macmillan.

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

Säljö, R. 2009. "Learning, Theories of Learning, and Units of Analysis in Research." *Educational Psychologist* 44 (3): 202–208. doi:10.1080/00461520903029030.

Schommer, M. 1993. "Epistemological Development and Academic Performance among Secondary Students." *Journal of Educational Psychology* 85 (3): 406–411. doi:10.1037/0022-0663.85.3.406.

Schommer-Aikins, M., W.-C. Mau, S. Brookhart, and R. Hutter. 2000. "Understanding Middle Students' Beliefs about Knowledge and Learning Using a Multidimensional Paradigm." *The Journal of Educational Research* 94 (2): 120–127. doi:10.1080/00220670009598750.

Soderstrom, N. C., and R. A. Bjork. 2015. "Learning versus Performance: An Integrative Review." *Perspectives on Psychological Science: A Journal of the Association for Psychological Science* 10 (2): 176–199. doi:10.1177/1745691615569000.

Songer, N. B., and M. C. Linn. 1991. "How Do Students' Views of Science Influence Knowledge Integration?" *Journal of Research in Science Teaching* 28 (9): 761–784. doi:10.1002/tea.3660280905.

Tai, J., R. Ajjawi, D. Boud, P. Dawson, and E. Panadero. 2018. "Developing Evaluative Judgement: Enabling Students to Make Decisions about the Quality of Work." *Higher Education* 76 (3): 467–481. doi:10.1007/s10734-017-0220-3.

Zimmerman, C. 2000. "The Development of Scientific Reasoning Skills." *Developmental Review* 20 (1): 99–149. doi:10.1006/drev.1999.0497.