Perpustakaan Tun Abdul Razak
UNIVERSITI TEKNOLOGI MARA

SDI@PTAR

Selective Dissemination of Information

Bil 4/2020

Assalamualaikum Warahmatullahi Wabarakatuh dan Salam Sejahtera

YBhg. Datuk/ Datin / Prof/ Tuan/ Puan

SDI@PTAR adalah salah satu perkhidmatan penyebaran maklumat terpilih yang disediakan oleh Perpustakaan Tun Abdul Razak, UiTM Shah Alam untuk ahli Mesyuarat Senat UiTM. Perkhidmatan ini bertujuan untuk menyalurkan maklumat terbaharu mengenai isu-isu semasa di dalam dan luar negara yang memberi nilai tambah serta impak kepada pengajaran, pembelajaran dan penyelidikan UiTM ke arah menjadi Universiti Terkemuka Dunia.

Untuk keluaran kali ini, SDI@PTAR menampilkan artikel teks penuh mengenai Higher Education di dalam 4 bidang iaitu (i) Higher Education Leadership and Management, (ii) Higher Education Teaching and Learning, (iii) Higher Education Transformation dan (iv) Higher Education Technology and Digital Transformation. Diharapkan maklumat ini memberi manfaat kepada YBhg. Datuk/ Datin/ Prof/ Tuan/ Puan.

Sebarang cadangan dan maklumbalas mengenai perkhidmatan ini boleh disalurkan kepada En. Mohd Ismail bin Abidin, Timbalan Ketua Pustakawan (e-mel ismailabidin@uitm.edu.my) dan Puan Nik Zatihulwani binti Jamaludin, Pustakawan, (e-mel nikzatihulwani@uitm.edu.my), Bahagian Penyelidikan, Pembelajaran dan Rujukan, Perpustakaan Tun Abdul Razak, UiTM Shah Alam.

Sekian. Terima kasih.

Bahagian Penyelidikan, Pembelajaran & Rujukan
Jabatan Perkhidmatan Perpustakaan
Perpustakaan Tun Abdul Razak Utama UiTM Shah Alam

Higher Education Transformation

BIL	JUDUL	AUTHOR	h-INDEX	UNIVERSITY	RANKINGS
1	A longitudinal study of the impact of reflective coursework writing on teacher development courses: a 'legacy effect' of iterative writing tasks	McLean, Neil	2	The London School of Economics and Political Science (LSE)	No. 2 in Social Sciences

Higher Education
https://doi.org/10.1007/s10734-018-0312-8

A longitudinal study of the impact of reflective coursework writing on teacher development courses: a 'legacy effect' of iterative writing tasks

Neil McLean[1] · Linda Price[2]

Published online: 19 September 2018

Abstract
Studies into the efficacy of teacher development courses for early career academics point to graduates conceiving of their teaching in increasingly complex and student-focussed ways. These studies have used pre- and post-testing of conceptions of teaching to identify this finding. However, these studies do not identify what aspects of these courses contributed to these changes. This exploratory case study investigates this phenomenon through a longitudinal study of 16 academic teachers' reflective coursework writing. Discourse analysis was used to contrast causal reasoning statements in assignments completed during participants' first 2 years in-service, while they were completing a UK-based teacher development course. This analysis identified how reasoning about teaching and learning became more complex over time. A key element was the integration of experiences and earlier learning into more nuanced and multi-factorial later reasoning about teaching choices and effects. This 'legacy effect' provides new evidence for the efficacy of academic teacher development courses.

Keywords Impact of teacher development · Academic identities · Reflective writing · Identity positioning

The impact of teacher development courses

The increasing link between academic probation and the completion of teacher development courses has sharpened interest in the efficacy of these courses. However, providing evidence for this efficacy is not straightforward, since it is difficult to separate out the impact of training and development programmes from other influences (Chalmers and Gardiner 2015; Norton et al. 2005; Roxå and Mårtensson 2015; Saroyan and Trigwell 2015).

✉ Neil McLean
N.D.Mclean@lse.ac.uk

[1] LSE Academic and Professional Development Division, London School of Economics, London, UK

[2] Centre for Learning Excellence, University of Bedfordshire, Luton, UK

Gibbs and Coffey (2004) argue:

> 'We are still not in a position to demonstrate that it was the training itself that resulted in positive changes, merely that institutions that had training also had teachers that improved.' (2004, p. 99)

Chalmers and Gardiner (2015) concur arguing that although increasing numbers of academic staff are required to take teaching training and professional development, there is limited evidence of their effectiveness. A further challenge in evaluating and comparing programmes is their diversity and the ambivalence of participants' reported experiences. For many, participation was not valued (Fanghanel 2004). Some pedagogical training has been spurned where the input of centrally organised training and development programmes is argued to lack congruence with disciplinary perspectives (Gibbs and Coffey 2004) or departmental norms and practices (Knight and Trowler 2000; Trowler and Cooper 2002).

However, studies into the impact of teacher development courses have identified positive influences on academic teacher development. Increased self-efficacy has been found to be associated with longer-term pedagogical training (Butcher and Stoncel 2012; Postareff et al. 2007, 2008). Contrastive studies have assessed participants' conceptions of teaching pre- and post-course, principally through the use of the Approaches to Teaching Inventory (Prosser and Trigwell 1997). These studies have shown that more 'complex', student-centred conceptions of teaching were present among those who had completed teacher development courses (Hanbury et al. 2008; Lindblom-Ylänne et al. 2006; Postareff et al. 2007).

A possible explanation for this finding is that attendance on such programmes may lead to informal learning (Butcher and Stoncel 2012; Knight and Trowler 2000) and the creation of communities of practice (Sadler 2008).

> 'Pedagogical courses, the main goal of which is developing teaching skills, can also be regarded as communities when they enable interaction between colleagues. The courses provide opportunities for university lecturers to contemplate and discuss their teaching with colleagues and help reduce academic isolation.' (Remmik et al. 2011, pp. 188-189)

A further explanation is that change in conceiving of teaching is promoted by reflection in the form of the kind of coursework writing that participants are asked to complete on these courses (M. McLean and Bullard 2000). This kind of writing is perhaps the most ubiquitous feature of what can be very different approaches to teacher development. Reflective writing serves as a vehicle for developing 'reflective practitioners' (Schön 1987) among early career academic staff. This reflective practice is intended to develop participants' appreciation of the complexity of teaching and learning that pre- and post-testing studies have found. This longitudinal study of coursework writing investigated this link between iterative reflective writing and the complexity of causal reasoning about teaching and learning.

One way to explain why this link might exist is from the perspective of identity formation. Goffman's work on 'presentation of self' characterises identity as co-constructed through being 'performed' with others (1990). In the case of teacher development contexts, the requirement to 'perform reflection' is a form of 'presentation of self' that combines personal experience and educational theory and research. This repeated performance would predictably lead to the formation of a reflective professional identity in this context (Davies and Harre 1990; McLean and Price 2017). This study seeks to contribute to the literature on the efficacy of teacher development programmes by providing a longitudinal investigation of this process of identity formation by identifying participants' 'interpretative repertoires' (Potter and

Wetherell 1987; McLean and Price 2016). These repertoires are ways of conceiving of and describing phenomena that are common in the discourse of particular identity positions (McLean 2012). Causal reasoning is a key element of repertoires (Edley 2001), and this study identified how participants reasoned about teaching and learning over a 2-year period. Coursework writing on these courses is therefore not just a potential explanation for findings of the positive effects of teacher development (Hanbury et al. 2008). This writing was also a source of naturally occurring data for the analysis of the impact of teacher development.

Methodology

This exploratory case study presents an investigation into the finding that graduates of teacher development courses think about teaching in a more complex manner than before they took their courses (Hanbury et al. 2008; Postareff et al. 2007, 2008). The setting for this study was a 'single faculty' social science university. This university employs some PhD students to teach undergraduate classes and offers these class tutors the chance to enrol on an in-service, 2-year teacher development course (a Postgraduate Certificate in Higher Education). This course provides an introduction to teaching and learning in higher education, with a particular focus on social scientific study. Research participants for this study were selected from among two different year groups on this teacher development course. The first selection criterion was that participants had no previous teaching experience. This enabled the study to investigate development during tutors' first 2 years in-service. The second criterion was that participants were teaching qualitative social science disciplines. This was to enable comparison across broadly cognate teaching experiences. Sixteen novice social science class tutors were invited to take part by allowing their completed portfolio of coursework assignments to be analysed. This invitation to participate came after participants had completed the course. There was therefore no influence on their writing from the context of being part of a research study.

In order to explore the finding of teachers' developing awareness of teaching and learning, a longitudinal approach was adopted that examined their reasoning about teaching across the 2 years of the course. The data source was participants' coursework assignments. In total, 80 texts were analysed: five coursework texts for each of the 16 study participants. The first text was written pre-service. The other texts were module assignments, each of which was completed at the end of each of participants' first four teaching terms. The assignments were approximately 2000 words in length and each one had their own guidelines and assessment criteria. The analysis of these texts enabled the kind of 'theory-informed, contextualised investigation' of the impact of a teacher development course advocated by Bamber (2008; 107).

A general requirement of the module assignments was for participants to write about teaching and learning in ways that combined reflection on teaching experiences and participants' reading of relevant educational literature. Table 1 provides an overview of these tasks.

Our hypothesis for this case study was that this writing serves as a means of encouraging reflection, where reflection is seen as a vehicle for developing increasingly complex notions of teaching and learning (Schön 1984, 1987). If this is the case, then a longitudinal study of causal reasoning statements in these texts should identify increasing complexity in participants' successive assignments. In the context of this study, complexity in causal reasoning statements is understood as observable discursive practices such as reference to multiple factors, qualification of claims and integration of different sources to justify decision-making.

Table 1 Overview of coursework assignments

Coursework task	Word guide	Instructions and key concepts
Pre-service reflective task	2500	Prompt questions elicited participants' views on aspects of class teaching, student learning, assessment and feedback and course design
Small-group teaching assignment (end of term 1)	2000	Participants were firstly asked to place their own approach to class teaching within the traditions of teaching in the disciplines. The second section reported how participants had made changes as they planned and delivered their term's teaching and the final section asked participants to identify lessons they had learned
Student learning assignment (end of term 2)	2000	Participants profiled students in one of their classes and identified challenges they faced, explored how learning theory offered insights into these challenges and proposed responses
Assessment and feedback assignment (end of term 3)	2000	Participants were asked to comment on the assessment on a course they were teaching on in terms of key concepts such as validity and reliability, as well as to comment on their own marking of a set of essays, and then to evaluate their own feedback to their students in terms of principles of effective feedback
Course design assignment (end of term 4)	2000	Participants were asked to comment on the design of the course they were teaching in terms of constructive alignment, proposing changes if appropriate, as well as to design in outline a course that they could teach on the basis of their own disciplinary knowledge

The study's hypothesis was therefore that these behaviours would be increasingly evident in the reasoning statements found in participants' later writing on the course. This investigation was to explore whether our hypothesis was correct, and if so why.

To conduct the longitudinal analysis, a form of discourse analysis was developed from the tradition of Discursive Psychology (Edwards and Potter 1992; McLean 2012; Wetherell and Potter 1992). This tradition explores the 'interpretative repertoires' of speakers or writers (Edley 2001). 'Interpretative repertoires' are the ways in which a person understands and provides explanations for phenomena, in particular their frames of reference and causal reasoning. The analysis of 'interpretative repertoires' in this study identified causal reasoning statements about teaching and learning in each of the five texts written over the 2-year period. The first stage of analysis was to identify statements where a causative conjunction (or a syntactic structure such as an infinitive of purpose) explicitly linked main and subordinate clauses or where one of these clauses was implied by the surrounding sentences and it was possible to supply this clause from the context. This process yielded 2487 causal statements from across approximately 160,000 words of the 80 texts.

These statements were analysed thematically in two further stages (Braun and Clarke 2006). Firstly, statements for each tutor were analysed according to each module assignment, with explicit consideration of the assignment guidelines which framed their use, much as an interview question frame would structure an interview-based approach (McLean and Price 2016). In this stage, organising themes for reasoning statements were identified for each teacher and each assignment. The next stage was to contrast organising themes from the reasoning statements across the 16 tutors, still explicitly considering the effect of the assignment guidelines. This second stage led to organising themes across the assignments that were present in writing across the 16 participants (Table 2).

Table 2 Data gathering and analysis

	Actions taken to examine participants' expanding awareness of teaching and learning
Stage 1—identification of the causal reasoning statements	Identification of discursive practices associated with implicit and explicit causal reasoning statements about teaching and learning. These statements are identified through causative conjunctions (or a syntactic structure such as an infinitive of purpose) explicitly linking main and subordinate clauses, or where one of these clauses is elided but clearly implied by the surrounding context. In total, this initial coding yielded 2487 explanatory causal reasoning statements about aspects of teaching and learning.
Stage 2—analysis of causal reasoning statements by assignment	Organising themes arising from coding causal reasoning statements by assignment.
Stage 3—contrasting reasoning statement between the 16 participants	Identification of organising themes across the 16 participants, and done by assignment.

This methodological approach has two advantages. First, stage 3 of the analysis created the basis to compare what participants wrote at the same time, with the same instructions, across their first 2 years of learning to teach. This made this analysis genuinely longitudinal. It took time-series examples of tutors' writing that are comparable through tutors writing at the same stage of their course and in line with the same assignment guidelines and assessment criteria. Second, in explicitly accounting for the influence of the programme coursework tasks, it is possible to assess the role these tasks played in novice teachers' expanding awareness of the complexity of disciplinary teaching (Åkerlind 2003).

In line with pre- and post-testing studies, and other studies on the impact of pedagogical training, (Hanbury et al. 2008; Postareff et al. 2007, 2008; Remmik and Karm 2009), we found that teachers' reasoning about their teaching and their students' learning grew more complex as the course progressed. The longitudinal methodology enabled us to account for this change. Crucially, these teachers' conceptions of teaching and learning built progressively on, and qualified, earlier explanations and understandings. In this study, the process of building on and qualifying earlier conceptions of teaching and learning over time is termed the 'legacy effect'. The driver for this effect was that the participants were required to write iterative reflective assignments. The next section demonstrates how the course requirement to repeatedly write about teaching and learning, integrating educational reading with lived experiences in a reflective manner, enabled these teachers to develop their conceptions of teaching over time.

The 'legacy effect' of iterative coursework writing

The thematic analysis of the causal reasoning statements showed that each tutor's interpretative repertoire for explaining teaching and learning became increasingly complex over time. This complexity was measured in terms of reference to multiple factors, the qualification of claims and reference to multiple sources to justify decision making. We believe that this increasing complexity was the result of multiple influences on these tutors during this 2-year period. However, a critical influence seems to be a pattern that emerged in the stages of analysis of participants' writing. This is that certain organising themes appeared and then recurred (and developed) in teachers' writing over time.

As these themes recurred, changes were identifiable in how teachers wrote about them or combined them with other themes. It was the identification of recurring themes and the increasingly complex discussion of these organising themes that allowed us to identify the 'legacy effect' created by completing the coursework assignments for the course. As will be shown in the next section, most organising themes directly related to assignment guidelines and requirements. However, others did not. In pre-service writing, some organising themes related to teachers' expectations and study experiences. In later assignments though, organising themes that recurred came from the assignment guidelines of earlier assignments. This recurrence of themes was not a requirement of later assignment guidelines. Instead, teachers were building on their earlier assignment writing as they completed subsequent tasks, and their reflection became increasingly complex as a result. Table 3 provides an overview of how organising themes recurred in teachers' writing.

This 'legacy effect' worked in the following manner. Participants encountered new ideas of class teaching in the first module of the course. These ideas built on and reframed their pre-service expectations. This new input mingled with the experience of teaching and dissonance between pre-service expectations and the realities of their teaching experiences. Then, in the second module on student learning, participants wrote about new input on learning theory and student diversity, but they also referred back to ideas encountered in the first module, in discussion around how to support their students' learning. Participants did this even though it

Table 3 Legacy effect demonstrated through flow of organising themes

Assignment	Organising themes from coding causal reasoning statements
Pre-service writing	Teacher as 'guide'
	Teacher's passion and enthusiasm
	Learning from own study experiences
	Developing knowledge of the discipline
Writing at the end of tutors' first teaching term (on small group teaching)	Teaching the discipline
	Managing participation
	Planning and preparing
	Student behaviours
	Student direction
Writing at the end of tutors' second teaching term (on student learning)	Student diversity
	Motivation and student behaviours (*continuation from module 1*)
	Learning theory
	Influence of course structures and assessment on student behaviours (*continuation from module 1*)
	Student direction (*continued from module 1*)
Writing at the end of tutors' third teaching term (on assessment and feedback)	Principles of assessment and feedback
	The influence of assessment methods and practice on student behaviours (*extension from module 2 theme*)
	Learning theory and student motivation (*continued from module 2*)
	Skills development/student direction (*continued from module 1*)
Writing at the end of tutors' fourth teaching term (on course design)	Principles of course design
	The influence of course design and delivery on student behaviours (*continued from modules 1 and 3*)
	Pedagogic content knowledge (*including the themes of teaching the discipline, student behaviours and student direction from earlier modules*)
	Learning theory and student motivation (*continued from module 2*)
	Institutional and educational realities

was not required for the second assignment. In their second year of teaching and after their third term, in their assignment on assessment and feedback, participants wrote about principles of assessment and feedback, but frequently justified the points they made with reference to learning theory and diversity from their second module. Further, in discussions around how feedback can influence study, the first assignment theme of 'student direction' recurred. Again, this was not an assignment requirement. Finally, after participants' fourth term of teaching and in their final assignment on course design, themes from all the earlier assignments were integrated into explanations of different aspects of teaching and learning, and in justifying decisions made about course outlines. The outcome of this 'legacy effect' is that tutors' interpretative repertoires for teaching and learning were far more complex in their final module assignment than in early writing. This complexity can be explained through a process of accretion of ideas from earlier modules, as well how tutors integrated these themes with their lived experiences of teaching.

How tutors explained aspects of teaching and learning pre-service

As novices, tutors' writing was understandably dominated by focus on their own plans and actions. Causal statements focussed on the characteristics of a 'good' teacher, with enthusiasm and passion highlighted in particular. However, because these expectations were based more on the teacher than their actual students, they tended to lack an appreciation of teaching and learning difficulties that appeared in their later module assignment writing. The following quotes are examples of this underestimation of complexity:

> I hope to make the material as interesting as possible by being enthusiastic about it … so that students feel more motivated to study it (Participant 14)
> In this process, I will be a supportive teacher who sets high expectations for the class, as well as for me. For example, in order to show the class that I am ready to work with them, I actually read all the readings for the first class and have developed my own presentation for this week. (Participant 9)
> …my students are meant to learn about one theorist per week, and for the exam specialise in about 4 – 5 theorists. This should provide them with a sound understanding of basic normative principles pertaining to politics. (Participant 6)
> And yet I find it necessary to guide the students in class discussions and take an active part, since I have the knowledge of historiography and of the existing debates which they lack. (Participant 8)

These statements mirror previous findings in the literature in relation to teacher-focussed conceptions of teaching at the beginning of development courses (Hanbury et al. 2008; Postareff et al. 2007, 2008). This was the 'starting point' from which teacher's reasoning about disciplinary teaching and learning became increasingly complex.

Development from pre-service to first term teaching

The experience of teaching and completing this module and its assignment led to clear differences in how participants explained teaching compared with descriptions from their pre-service writing. Pre-service reasoning statements about the role of the teacher focussed

on content knowledge, passion and enthusiasm and being a 'guide'. Reasoning statements within these themes echoed those in pre-service writing, but what was added was how these beliefs and values could (and should) be operationalised, and how input on teaching had made a contribution to this. The following quotes illustrate this development:

> … the course I am teaching on covers a very broad area. In this context, my primary aims for the class were to provide a bridging role between the material of different weeks in order to specifically avoid the problem of 'all periphery and no core' (Piachaud 2007), to suggest specific literature and help the students negotiate the long reading list … (Participant 15)
>
> Many students were taking the course *for* credit at their home institution, so I felt a strong responsibility to prepare them for the mid-term essay and terminal exam with as much focus on the syllabus as possible (Participant 10)
>
> My first experience of teaching has been with a small (often really very small) group of mostly quiet students at 9am, which certainly throws up challenges … The GTA and PGCertHE workshops have been absolutely vital, summarising a move through knowledge to interpretation … What I have tried to do is to start with an exercise that serves as both warm-up and knowledge fixing, usually a list of questions summarising key points in the lecture and reading. (Participant 1)

Discussion of planning and preparing was absent in tutors' pre-service writing; however, early input on the teacher development course on planning was reflected in tutors' writing at the end of their first term:

> I have varied the teaching styles out of consideration for different styles of learning and different student needs. Kennedy (2007) makes the point that different styles benefit different students, in relation to class debates. (Participant 12)
>
> This term I have organised my classes through my own PowerPoint presentation. Initially, this was motivated largely by fear of 'drying up' in class. (Participant 7)

Teachers were surprised and frustrated when students did not behave as would have liked or expected. This experience of student resistance created reasoning statements that reflected an awareness of the limits of their control in overseeing their students' learning. This theme of constraint developed over time. It was particularly evident in participants' second module assignment and their responses to course input on student learning.

Integration of themes from the first to the second module assignment

An important theme absent from pre-service writing, but that emerged in the second module assignments, was the influence of course structures on learning. An example of this is discussion of assumed knowledge:

> As an introductory course, it does not assume any knowledge, but there is a steep learning curve for non-philosophy students to become familiar with philosophical jargon and knowledge of broad philosophical positions. (Participant 5)
>
> The two most common issues for students in my seminar group are the amount of assumed empirical knowledge, and their lack of familiarity with the specific modes of political-sociological argument and explanation which this course requires. (Participant 4)

Tutors wrote, largely for the first time, about their environments and what this meant for students, and therefore for them as teachers.

> It was clear that students in this class were intelligent and motivated by high exam performance, though it was a challenge to get them interested in the course material. I approached this challenge in two ways. First, I regularly introduced supplementary material from current events that demonstrated 'social policy' concepts in action … Second I acknowledged the performance orientation of the group by regularly demonstrating strategies to approach assigned readings that would facilitate high performance in the seminar discussions, essays and the exam. (Participant 13)

In contrast to first assignment writing, student behaviours were identified as challenging.

> Nevertheless, I noticed that in the second term, my students' preparation for the course decreased – they knew less due to having done fewer readings … I always said that I would understand that they had other obligations as well … However, I was strict with the ones who were not co-operating by, for instance, setting another deadline, and underlining the consequences for non-compliance … (Participant 16)

Similarly, reasoning statements about directing or guiding students carried forward from the first assignment, now included more reference to learning, rather than simply teaching:

> I will focus, therefore, on instructing my students in the modes of analysis and argument which are commonly used in political sociology. My reasons for this are partly practical: mastery of these modes is essential if they are to succeed… (Participant 4)

Another development was multiple instances of values espoused in the first module writing that were integrated into learning-theory-informed reasoning in the second module writing. For example, in module 1, participant 1 wrote about using essay-planning tasks in this way:

> I can see that the danger of being content with the evidence of intellectual understanding, foregoing the next stage of teaching the craft of application … which lies at the heart of formulating fully developed arguments. That is why I think the essay exercises … which I would like to reproduce in different variants, are so important.

In the same teacher's module 2 writing, this had become:

> But for some students who had clearly done inadequate reading or still had lacunae resulting from a non-IR background, I encouraged them to use the exercise of essay planning to build up a picture of what they needed to revise further, which speaks to the constructivist approach …

The distinction here is how the two quotes show increasing awareness of the likely influence of essay planning exercises. In the first quote, participant 1 presents these exercises almost as a panacea for a potential teaching challenge. There is no acknowledgement of student diversity, or practical challenges such as variable student reading. These considerations are though present in the second quote, as is the use of the verb 'encourage', which suggests that participant 1's sense of his own control of student learning is more qualified. This is a good example of a shift to an increasingly student-centred conception of teaching (Åkerlind 2003).

Integration of themes from the first two modules into end of their third term writing

There was no requirement for tutors to refer back to earlier modules when completing their assignment on assessment and feedback. Nonetheless, earlier themes were integrated into tutors' discussion and understanding of assessment. In particular, the influence of the second assignment on student learning was very clear in tutors' writing. As such, a shared topic in the second and third assignments was student motivation.

> Students who are inclined to be, for want of a better term, instrumentally rational, are more likely to ask questions about the exact requirements of the course so they can minimise the amount of work they have to do in relation to the marks they want to achieve. (Participant 4)
>
> Although the formative essays are designed to prepare students for the summative assessment, in practice, students do not always see the relationship between the two and how they might use feedback from the former to help with the latter. As noted by Brown 'students take their cues from what is assessed, rather than from what lecturers assert is important'. (Participant 6)

Also evident as an influence was the integration of learning theory into espoused good practice on feedback. Participant 5 provides a good illustration of this:

> It is a continuing problem that I have no office hours, so it is difficult to provide specific individual guidance. It would be helpful to have an office hour so that I could encourage more students to engage in one-to-one help, as individual guidance, in my experience, appears to have a positive effect on essay marks. However, the extent to which students will seek and act on advice will depend on their goals, as evidenced by learning theory. Some students are focussed on a full understanding of the subject, hence likely to act on advice such as widening their reading to develop their arguments. Students with performance-oriented goals, whose primary interest is to pass the exam, are less likely to act on such advice (Mattern 2005: 27).

It is also the case that the themes raised in the first module and carried into module 2 reasoning statements about teaching and learning and were also integrated here. A good example is how the early theme of directing or guiding students informed the identification of skills deficits in the assessment module assignments:

> Another common issue … is their inability or unwillingness to write in their own voice and make their own original set of arguments to a question. While I recognise that this is not an easy skill … I do encourage and support them, and do find that over the course of the year, many are able to make important strides in this area. I also encourage this focus on original argumentation because it is an important element in their summative assessment in this course. (Participant 9)
>
> I would be more conscious of the particular language used in the marking scheme when writing essay feedback. I would also end on a positive note in order to emphasise the skills that the student has shown. (Participant 2)

The change here in teachers' reasoning about teaching is that initially, they tended to offer simple teaching plans that would apply equally to all students. Over time, teachers focussed more on context and the multiple factors affecting their different students' learning. This led to

reasoning statements that differentiate between their students and that qualify the extent to which any teaching intervention will apply equally to all students. This is another example of 'expanding awareness' (Åkerlind 2003), where building on earlier ideas led to more complex later reasoning. The requirement to write reflective assignments was the driver for this increasing complexity.

Integration of themes from earlier module into fourth term writing

The intention of this final assignment on course design was to encourage tutors to draw together learning from across the course. Although this was not explicit in the guidelines, this is what happened in the causal reasoning statements identified in this set of 16 assignments. An example of this is the theme of student motivation, common in all of the final assignments. This was evident even though including learning from earlier assignments on student learning was not a requirement. Even so, participants used learning theory to justify design choices.

> Natural feedback from students who do not wish to take the course suggests that the irrelevance of the course for career aspirations, as well as a concern that the course will be too difficult, prevents the students from engaging in the course proactively and with interest … In order to ensure that they engage … the course needs to be designed in such a way as to ensure that, in order to pass, students engage with learning activities which are focussed around student-centred learning outcomes (Biggs 1996, 2004). (Participant 14)
>
> The variety of backgrounds means that there will always be significant variance in the initial understanding of the subject that students bring with them to the course. This in turn means that the way in which students 'construct' meaning out of what they are studying is likely to differ … (Participant 3)

Similarly, participant 4 used constructive alignment principles to explicitly ensure that '…deep learning is the best exam strategy'.

Assessment choices in course redesign and proposals incorporated reasoning from the earlier assignment on assessment (e.g. principles such as validity). The following quotes demonstrate this link between reasoning in the two modules:

> Exams challenge the necessary validity of assessment methods (i.e. whether an assessment tests what it wants to test), since exams tend to test "skills" outside of the ones practiced during term time … Indeed, according to information I collected for Module 4 on student assessment, I found that most students perform better in their essays than they do on the final exam. This leads me to the conclusion that there may be better ways by which the assessment methods could prepare students to succeed in the course. (Participant 2)
>
> My approach on both designed courses has been to have a diversity of assessment methods to maximise the validity of the course in terms of the students' diverse skills and, crucially, to assess the learning outcomes thoroughly. (Participant 1)

This final quote is an example of how participants integrated earlier themes from multiple modules. Discussion of input on constructive alignment (final module) includes reference to earlier input on learning theory (reference to Bloom and 'deep' learning) and first module input

on managing participation (reference to teacher's role, which in her earlier assignment focused on an interest in Paolo Freire's critical pedagogy and 'teacher as facilitator').

> Having clear outcomes is the first step in constructive alignment. In the revised (course code), students know that they are expected to learn the vocabulary used in the subject. According to Benjamin Bloom's taxonomy of cognitive levels (1984), merely learning vocabulary and what concepts mean constitutes first (or at most second) class cognition. The verbs used in the learning outcomes are consciously higher-order actions that encourage deep learning. Importantly, clear outcomes also shift the responsibility and thus power from the teacher to the student, thus facilitating student-based learning (O'Neill and McMahon 2005). The teacher's role then becomes that of a facilitator and resource person. (Participant 2)

Conclusion

This study contributes to the body of knowledge on the impact of teacher development by demonstrating the value of iterative reflective writing over time. Our longitudinal study indicates that the 'legacy effect' of building on earlier learning offers an explanation for how participating in a teacher development course develops teachers' conceptions of teaching and learning over time.

The impact of teacher development courses and programmes has previously been examined using pre- and post-testing (Coffey and Gibbs 2002; Hanbury et al. 2008; Lindblom-Ylänne et al. 2006; Postareff et al. 2007) and case studies (Butcher and Stoncel 2012; Ho et al. 2001). However, these methods are not well indicated to investigate how these courses contribute to changing teachers' conceptions of teaching. This exploratory case study presents complementary findings to pre- and post-testing studies by demonstrating how this group of novice academic teachers' reasoning about their teaching developed over time. This analysis shows that tutors' interpretative repertoires for their teaching became more complex through the integration of new ideas on education that they encountered as the course itself developed. Iterative reflective writing provided the means through which this change was achieved.

This suggests that a key contribution of the course was to provide the opportunity to write iteratively about teaching and learning. This form of meaning making seems to have encouraged tutors to integrate ideas about teaching that came from course input, reading educational literature and reflecting on their teaching experiences. This integration, over time, seems to have built-in narratives of self as academic teachers of the sort identified as 'identity trajectories' (McAlpine and Lucas 2011). A key feature of these trajectories is that individuals link past, present and future in a coherent way that accounts for who they have become (Acker and Armenti 2004). The mechanism for this integration took the form of a 'legacy effect' evident in the iterative process of writing about teaching and learning. This 'legacy effect' was evident even though it was not a formal requirement of the coursework tasks, which suggests that writing of this sort can provide a basis for identity formation through consistent interaction and presentation of self (Davies and Harre 1990).

This study provides the basis for further research into the effect of other teacher development programmes that use reflective writing. While the context of this study is tied to one particular course and institutional setting, it does provide a framework for longitudinal investigations in other contexts. We believe that further research will offer insights into the

development of teachers' conceptions of teaching and their expanding awareness of teaching and learning. An important additional element would be to also investigate the link between changing conceptions identified here and actual classroom practice. The context within which academics work is also worthy of further investigation: as Roxå and Mårtensson (2015) and Englund and Price (2018) point out, the degree to which academics 'apply' their teaching beliefs is influenced by their surrounding environment. This is implied in what participants wrote in this study, and evidence of teaching materials was provided, but it was not systematically investigated. On a practical level, this study suggests that there would be value in explicitly requiring backwards and forwards referencing in coursework reflective writing as part of teacher development coursework guidance.

References

Acker, S., & Armenti, C. (2004). Sleepless in academia. *Gender and Education, 16*(1), 3–24.

Åkerlind, G. S. (2003). Growing and developing as a university teacher: Variation in meaning. *Studies in Higher Education, 28*(4), 375–390.

Bamber, V. (2008). Evaluating lecturer development programmes: Received wisdom or self-knowledge? *International Journal for Academic Development, 13*(2), 107–116.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*, 77–101.

Butcher, J., & Stoncel, D. (2012). The impact of a postgraduate certificate in teaching in higher education on university lecturers appointed for their professional expertise at a teaching-led university:'It's made me braver. *International Journal for Academic Development, 17*(2), 149–162.

Chalmers, D., & Gardiner, D. (2015). The measurement and impact of university teacher development programs. *Educar, 51*(1), 53–80.

Coffey, M., & Gibbs, G. (2002). Measuring teachers' repertoire of teaching methods. *Assessment & Evaluation in Higher Education, 27*(4), 383–390.

Davies, B., & Harre, R. (1990). Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour, 20*(1), 43–63.

Edley, N. (2001). Analysing masculinity: Interpretative repertoires, ideological dilemmas and subject positions. In M. Wetherell, S. Taylor, & S. J. Yates (Eds.), *Discourse as data: A guide for analysis* (pp. 189–228). Milton Keynes: The Open University.

Edwards, D., & Potter, J. (1992). *Discursive psychology* (Vol. 8). London: Sage.

Fanghanel, J. (2004). Capturing dissonance in university teacher education environments. *Studies in Higher Education, 29*(5), 575–590.

Gibbs, G., & Coffey, M. (2004). The impact of training of university teachers on their teaching skills, their approach to teaching and the approach to learning of their students. *Active Learning in Higher Education, 5*(1), 87–100.

Goffman, E. (1990). *The presentation of self in everyday life*. London: Penguin.

Hanbury, A., Prosser, M., & Rickinson, M. (2008). The differential impact of UK accredited teaching development programmes on academics' approaches to teaching. *Studies in Higher Education, 33*(4), 469–483.

Ho, A., Watkins, D., & Kelly, M. (2001). The conceptual change approach to improving teaching and learning: An evaluation of a Hong Kong staff development programme. *Higher Education, 42*(2), 143–169.

Knight, P. T., & Trowler, P. R. (2000). Department-level cultures and the improvement of learning and teaching. *Studies in Higher Education, 25*(1), 69–83.

Lindblom-Ylänne, S., Trigwell, K., Nevgi, A., & Ashwin, P. (2006). How approaches to teaching are affected by discipline and teaching context. *Studies in Higher Education, 31*(03), 285–298.

McAlpine, L., & Lucas, L. (2011). Different places, different specialisms: Similar questions of doctoral identities under construction. *Teaching in Higher Education, 16*(6), 695–706.

McLean, N. (2012). Researching academic identity: Using discursive psychology as an approach. *International Journal for Academic Development, 17*(2), 97–108.

McLean, M., & Bullard, J. E. (2000). Becoming a university teacher: Evidence from teaching portfolios (how academics learn to teach). *Teacher Development, 4*(1), 79–101.

McLean, N., & Price, L. (2016). The mechanics of identity formation. In J. Smith, J. Rattray, T. Peseta, & D. Loads (Eds.), *Identity work in the Contemporary University* (pp. 45–57). Rotterdam: Sense Publishers.

McLean, N., & Price, L. (2017). Identity formation among novice academic teachers: A longitudinal study. *Studies in Higher Education* (pp. 1–14).

Norton, L., Richardson, T., Hartley, J., Newstead, S., & Mayes, J. (2005). Teachers' beliefs and intentions concerning teaching in higher education. *Higher Education, 50*(4), 537–571.

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2007). The effect of pedagogical training on teaching in higher education. *Teaching and Teacher Education, 23*(5), 557–571.

Postareff, L., Lindblom-Ylänne, S., & Nevgi, A. (2008). A follow-up study of the effect of pedagogical training on teaching in higher education. *Higher Education, 56*(1), 29–43.

Potter, J., & Wetherell, M. (1987). *Discourse and social psychology: Beyond attitudes and behaviour*. London: Sage.

Prosser, M., & Trigwell, K. (1997). Relations between perceptions of the teaching environment and approaches to teaching. *British Journal of Educational Psychology, 67*(1), 25–35.

Remmik, M., & Karm, M. (2009). Impact of training on the teaching skills of university lecturers: Challenges and opportunities. *Haridus, 11*(12), 20–26.

Remmik, M., Karm, M., Haamer, A., & Lepp, L. (2011). Early-career academics' learning in academic communities. *International Journal for Academic Development, 16*(3), 187–199.

Roxå, T., & Mårtensson, K. (2015). Microcultures and informal learning: A heuristic guiding analysis of conditions for informal learning in local higher education workplaces. *International Journal for Academic Development, 20*(2), 193–205. https://doi.org/10.1080/1360144X.2015.1029929.

Sadler, I. (2008). *Development of new teachers in higher education: Interactions with students and other influences upon approach to teaching*. (PhD), University of Edinburgh, Edinburgh.

Saroyan, A., & Trigwell, K. (2015). Higher education teachers' professional learning: Process and outcome. *Studies in Educational Evaluation, 46*, 92–101.

Schön, D. A. (1984). *The reflective practitioner: How professionals think in action*. New York: Basic Books.

Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco: Jossey-Bass.

Trowler, P., & Cooper, A. (2002). Teaching and learning regimes: Implicit theories and recurrent practices in the enhancement of teaching and learning through educational development programmes. *Higher Education Research and Development, 21*(3), 221–240.

Wetherell, M., & Potter, J. (1992). *Mapping the language of racism: Discourse and the legitimation of exploitation*. New York: Columbia University Press.

Bil 4/2020

Higher Education Leadership and Management

BIL	JUDUL	AUTHOR	h-INDEX	UNIVERSITY	RANKINGS
1	Preparing interdisciplinary leadership for a sustainable future	Bammer, Gabriele	27	The Australian National University	No. 29 in World University Rankings

Sustainability Science
https://doi.org/10.1007/s11625-020-00823-9

ORIGINAL ARTICLE

Preparing interdisciplinary leadership for a sustainable future

Christopher G. Boone[1] · Steward T. A. Pickett[2] · Gabriele Bammer[3] · Kamal Bawa[4,5] · Jennifer A. Dunne[6] · Iain J. Gordon[7] · David Hart[8] · Jessica Hellmann[9] · Alison Miller[10] · Mark New[11] · Jean P. Ometto[12] · Ken Taylor[13] · Gabriele Wendorf[14] · Arun Agrawal[15] · Paul Bertsch[16] · Colin Campbell[17] · Paul Dodd[18] · Anthony Janetos[19] · Hein Mallee[20]

Received: 21 August 2019 / Accepted: 22 May 2020

Abstract
Urgent sustainability challenges require effective leadership for inter- and trans-disciplinary (ITD) institutions. Based on the diverse experiences of 20 ITD institutional leaders and specific case studies, this article distills key lessons learned from multiple pathways to building successful programs. The lessons reflect both the successes and failures our group has experienced, to suggest how to cultivate appropriate and effective leadership, and generate the resources necessary for leading ITD programs. We present two contrasting pathways toward ITD organizations: one is to establish a new organization and the other is to merge existing organizations. We illustrate how both benefit from a real-world focus, with multiple examples of trajectories of ITD organizations. Our diverse international experiences demonstrate ways to cultivate appropriate leadership qualities and skills, especially the ability to create and foster vision beyond the status quo; collaborative leadership and partnerships; shared culture; communications to multiple audiences; appropriate monitoring and evaluation; and perseverance. We identified five kinds of resources for success: (1) intellectual resources; (2) institutional policies; (3) financial resources; (4) physical infrastructure; and (5) governing boards. We provide illustrations based on our extensive experience in supporting success and learning from failure, and provide a framework that articulates the major facets of leadership in inter- and trans-disciplinary organizations: learning, supporting, sharing, and training.

Keywords Interdisciplinary organization · Leadership · Lessons learned · Transdisciplinary

Introduction

Inter- and trans-disciplinary (ITD) research has expanded in recent decades and there is growing evidence that ITD research helps solve complex societal problems and achieve societal aspirations (Irwin et al. 2018; Frantzeskaki and Rok 2018). Interdisciplinary research integrates disciplinary knowledge to create new scientific understanding while transdisciplinary research also incorporates knowledge and participants from beyond science to engage in the research process and inform policy and practice (Lang et al. 2012; Buizer et al. 2015). Alongside the growth in ITD research and application, organizations are being established to foster ITD research and to train students for new ITD careers (Huutoniemi et al. 2010; James Jacob 2015). These organizations are helping to meet growing demands on universities and other research institutions to demonstrate meaningful impact in meeting complex societal and environmental concerns (Caves 2020).

Urgent sustainability challenges require ITD leadership. Future leaders can benefit from lessons learned (Reid and Mooney 2016; Annan-Diab and Molinari 2017). We assert that such lessons can benefit from diverse experience with both successes and failures of past and on-going ITD efforts. Despite progress in developing ITD research programs, young researchers are still confronted with traditional incentives that discourage ITD activities (Bark et al. 2016; Brister 2016). To succeed, new leaders should be trained to navigate the problem-oriented nature of ITD research and to transform academic and research institutions to encourage rather than discourage ITD approaches, which is especially crucial

Handled by Alexander Gonzalez Flor, University of the Philippines Open University Faculty of Information & Communication Studies Los Banos, Laguna 4031, Philippines.

✉ Christopher G. Boone
cgboone@asu.edu

Extended author information available on the last page of the article

Published online: 31 May 2020

for the solution-orientated realms of sustainability (Liu et al. 2015; McDaniels and Skogsberg 2017; Gordon et al. 2019).

Lessons described in this paper are based on the thoughts, reflections, and experiences of 20 leaders of ITD organizations from nine countries (Palmer 2018) elicited and synthesized over several workshops. The objective is to advise leaders across various ITD fields and provide helpful justifications for universities, funders, and governments to support ITD initiatives. This is not a comprehensive 'handbook' on successful ITD leadership. Rather, it distills three lessons that current and future leaders of ITD initiatives should recognize and marshal resources to address: (i) the multiple pathways to successful programs; (ii) cultivation of appropriate leadership; and (iii) resources necessary for success.

Pathways to inter- and transdisciplinary sustainability organizations

Pathways to successful ITD organizations generally fall into two categories: some were created as ITD organizations by design (Box 1) while others evolved over time, often merging disciplinary units together (Box 2). The descriptions in Boxes 1 and 2, (along with Boxes 3–5) show how different organizations view themselves in relation to interdisciplinarity and/or transdisciplinarity and how they operationalize those approaches. Many of us started as disciplinary scientists and followed different paths to ITD, in the process creating a range of programs that approach sustainability challenges in various ways.

Both kinds of ITD organizations can benefit from a real-world focus. The leap from interdisciplinary to transdisciplinary programs can be accelerated by focusing on the public good or the needs of external partners (Fig. 1). Mission-oriented science requires the integration of multiple forms of knowledge and the expertise of end users. To mitigate poor air quality, for instance, requires the integrated expertise of many scientists and stakeholders to comprehend the dynamics of air quality, effects on humans and environment, and to build viable solutions, including atmospheric scientists, transportation modelers, public health officials, environmental economists, automotive engineers, and communication specialists. In the United States, federal transportation funds are tied to air quality, which incentivizes functioning ITD teams to address this as a public health and economic issue (https://www.fhwa.dot.gov/environment/air_quality/). This example shows the value of a problem-oriented and solution-oriented ITD approach with stakeholders connected to specific public good outcomes (Miller et al. 2014).

ITD organizations are motivated in various ways. Several universities have developed 'grand challenges' to encourage ITD research, education, and partner engagement. These programs may be assembled across existing units within

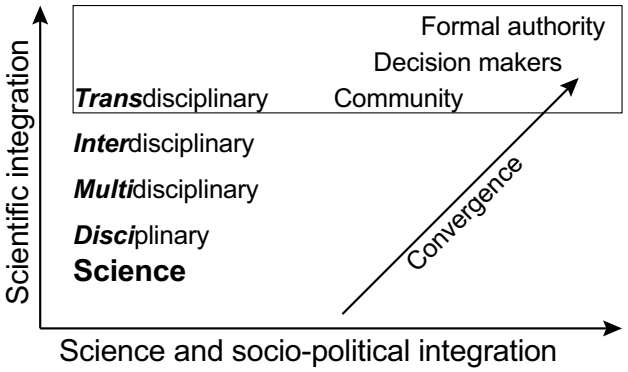

Fig. 1 Two dimensions of integration involving scientific research. Scientific integration, represented on the vertical axis, moves from disciplinary focus, through inclusions of multiple disciplines in a study, to the integration of those multiple disciplines in question asking, methodology, conclusions, and application. The final step of scientific integration is defined by its linkage with societal and political needs. Transdisciplinarity requires that various kinds of participants or stakeholders, here signified by communities, decision makers, and formal authorities (shown here on the horizontal axis), must be involved in posing questions, aligning methods, and assessing outcomes. The move from disciplinary science through transdisciplinary scientific-social research and intervention has been defined by the US National Science Foundation as convergence science. Some scholars use the term convergence to refer to deep integration in the scientific sphere as well (e.g., Irwin et al. 2018)

academic and research institutions or may bring together academic and mission-oriented partners. One example is Sustainable Los Angeles. Working across multiple colleges at UCLA, the university provided seed funds for research and education programs to help Los Angeles supply 100 percent renewable energy and 100 percent local water by 2050 while improving ecosystem health (Gold et al. 2015). The ambitious goals and long time horizon can inspire ITD collaboration because they address concerns that matter and have the potential of making a difference to the quality of life in a major city.

However, challenges do not have to be 'grand' to inspire ITD activities. Drawing more limited boundaries in space and time can encourage teams to tackle the inherently complex social–ecological–technical systems of sustainability challenges (Palmer et al. 2016) and short-term, smaller scale challenges can be equally energizing for researchers. Mitigating urban 'food deserts' is an example of a local need around which ITD researchers can band together for quick results, as food production in urban systems benefits from an ITD approach toward sustainability and social equity (Brinkley et al. 2017). For some researchers, the tangible, local, and immediate problems may be more motivating for ITD than global grand challenges. These fine-scaled ITD problems also allow flexibility, encouraging teams to form and reorganize according to the expertise needed rather than to maintain a persistent and potentially costly organization.

Rapid response to crises is another motivation for creating ITD projects and teams. Hurricane Sandy laid bare the inadequacies of New York City's preparation for extreme weather events (Rosenzweig and Solecki 2014). Academic institutions and local and state governments responded with an integrated resilience plan that joined expertise from research institutions, local and state agencies, community organizations, and the private sector with the explicit mission of making the metro area more resilient to major storms. The New York City Mayor's Office of Recovery and Resiliency was created in response to the devastating hurricane, which claimed 147 lives and caused $71 billion in damages (https://www.fema.gov/mat-results-hurricane-sandy). This office works closely with academics to develop and implement science-informed resiliency efforts to better prepare the city for future impacts of climate change. The ITD approach is reflected in the many dimensions of plans that go beyond physical infrastructure to include financial instruments, social vulnerability metrics, emergency planning with community organizations, and public health readiness. Advised by the New York City Panel on Climate Change, this office includes a scientific board that works in close partnership with the Center for Climate Systems Research within Columbia University's Earth Institute.

The variety of trajectories of ITD organizations suggests that a diverse roster of skills is needed for their leadership. Successful leaders must develop strategies and techniques for adapting to changing institutional situations and practical contexts. From our collective experiences, we summarize the skills below. In an earlier article (Gordon et al. 2019), we reviewed a broader range of skills involved in leading ITD organizations. Here we focus specifically on skills required for adapting to change, which is a major requirement for ITD organizations.

Cultivate appropriate leadership qualities and skills

Leaders of ITD organizations need the qualities that make any leader successful—creativity, humility, open-mindedness, long-term vision, and being a team player. In addition to these general qualities, ITD leaders require skills and attributes that are specific to inter- and trans-disciplinary interactions and that have the capacity to be transformative with real-world impacts. ITD leaders often must be more persuasive than other leaders to convince researchers to follow the unsettled and novel pathways of ITD research. Qualities that have been most transformative in our own journeys as leaders are the ability to create and foster: vision beyond status quo, collaborative leadership and partnerships, shared culture, communications to multiple audiences, appropriate monitoring and evaluation programs, and perseverance. It is important to note that these leadership qualities, skills, and attributes evolved over time. We did not begin our positions with each of these at hand; rather, as our roles and institutions grew, so did our leadership in these areas. Often, no individual has all of these qualities so it is also important to build a team that incorporates the full suite of these abilities.

Vision beyond status quo

Sustainability necessitates long-term vision that goes beyond the status quo (Matson et al. 2016). The complexity and scale of the challenges we confront require working and planning at time scales longer than the tenure of individual leaders. ITD leaders need the ability and creativity to see beyond existing conditions to imagine what is possible, what is needed, and how to get there, while integrating multiple stakeholder insights. We have operated in institutions that are sometimes slow to move and hesitant to change, yet we laid out strategic long-term plans that defied existing structures to facilitate the ITD goals we articulated. Ashoka Trust for Research in Ecology and the Environment (ATREE) in India provides an example of the vision and evolution required to move beyond the status quo (Box 3).

Collaborative leadership and partnerships

Leadership is a multidimensional process. It is important to know how to share leadership and to support the many roles required for sustainability work. Designated leaders must sometimes act as supporters, or as champions outside the organization. Appreciating and practicing different roles is a key cultural habit for leaders of ITD organizations. In some circumstances, ITD leaders must act as facilitators, 'de-centering' the role of academia to effectively prioritize the voices, concerns, and ideas of diverse stakeholders (Alonso-Yanez et al. 2019). Shared leadership may mirror necessities within ITD centers. Because of the multiplicity of leadership attributes, a team of more than one leader may be appropriate. The shared leadership model—as for example practiced by ZTG in TU Berlin and by the Wrigley Institute at ASU (Box 1)—also supports the idea of non-hierarchical working-structures, raising the credibility that partners outside of academia are fully accepted for their specific knowledge and perspectives.

Effective collaboration can catalyze problem analysis and address the broad range of elements that must be considered. Collaborative methods can be central for improving use of natural resources shared by society (Talley 2016) while also enhancing governance and accountability. Nevertheless, it is important to consider how and when to collaborate with partners. There is a tendency to want to partner with everyone who is interested, particularly in sustainability where the challenges are complex and

sense of urgency is strong. However, in our experience, the most effective leaders have developed clear processes for assessing whether to partner and how to measure success of partnerships. There are transaction costs to engaging partners as every partnership is a decision to allocate time and money. If not done carefully, partnerships can drain resources, taking intellectual and financial capital from other more fruitful activities. Before engaging with partners, it is important to ask key questions: Are the partner's objectives and proposed activity aligned with our strategy and operational plans? Can we establish and commit to a clear governance structure and resourcing? Is there enthusiasm from faculty and researchers? Is the proposed engagement intellectually interesting and impactful in the field? When the answers are yes, strong leaders invest to build participation, trust, excitement, and outcomes. Two examples of effective partner engagement are described in Box 4.

If an ITD organization identifies a strategic partner, it is important to engage them as much as possible from the beginning of the research process (Herrero et al. 2019). However, such participatory processes have challenges that need to be crystal clear to everyone from the outset, thereby avoiding frustrations from results that might not meet expectations (Stokols 2006; Disterheft et al. 2015). Clear articulation of the possible trade-offs between the scientific ideas and participatory methods is important to establish. A transparent set of scientific tools, visualized well across research phases, and a clear integration of different ways of expressing knowledge, including the follow-up of the results and the feedback to the stakeholders or to the practitioners, are of central importance (Mielke et al. 2017). Effective stakeholder engagement requires open access to data and knowledge so that key information is not restricted to the academic team members (Kondo et al. 2019). This approach provides informed options for decision processes while also using feedback from stakeholders to advance a specific research agenda. The development of the research or solution should be co-planned with stakeholders as this facilitates a way to effectively design and to measure outcomes. Determining outcomes with stakeholders increases the chance that results will be taken seriously and be implemented, while also incentivizing communities to help with gathering data (Heinzmann et al. 2019). However, lack of a concrete framework or model for carrying out a transdisciplinary sustainability project can increase potential for failure or reduce effectiveness of implementation (Smetschka and Gaube 2020). The risk associated with failing to meet anticipated objectives can be minimized by regularly revisiting goals and progress with all interested parties within an agreed upon evaluation framework (Williams and Robinson 2020; Turner and Baker 2020).

Shared culture

Because sustainability and ITD science are relatively new, attention to culture is crucial for future leaders (Longino 1990; Johnson and Xenos 2019). Culture includes norms and habits of mind that affect problem selection, research approaches, pathways of application (Pickett et al. 2007) and adapted solutions. Norms can limit or promote specific research and outcomes. Indeed, the traditional culture of science has promoted narrow disciplinary and academic outcomes (Capra 1983). Even tacitly adopting a familiar scientific culture may thwart the interdisciplinarity that sustainability requires.

Culture usually exists in the background, yet to succeed, leaders of ITD organizations must promote a new scientific culture that values and promoted ITD research and activities. They may have to guide their organizations through articulating and establishing new norms, finding ways to reward appropriate collaborative behaviors, and discouraging lapses into cultural norms of a narrow disciplinary past (Brown et al. 2019). Among the most significant cultural features supporting ITD success is a sharing attitude. This feature may be difficult for those trained in science as an individual, rule-based pursuit. In particular, the traditional idea that an individual researcher owns data can impede robust ITD research (Willig and Walker 2016). Consequently, sharing data in clear, well-documented, understandable formats is an important cultural norm for interdisciplinarity and transdisciplinarity.

Communications with multiple audiences

Communication is respectful listening coupled with clarity of exposition. Oral, written, quantitative, and visual modes may be combined in many ways. Conducive places for discussion, scheduled and serendipitous meetings, and access to multiple tools are all parts of effective communications in ITD organizations. Effective communication requires deep respect for other ways of knowing and social practices, especially as ITD endeavors engage increasingly diverse stakeholders. Because sustainability problems are complex, successful ITD leaders find it helpful to have a clear understanding of the logic of constituent or partner institutions and the incentives that drive stakeholders and find ways to mediate, resolve conflicts, and develop common ground priorities (Barrett et al. 2019).

Effective communication within the organization is also required to build and maintain networks uniting disciplinary expertise for ITD challenges. Communication with senior leadership of larger organizations that may host ITD centers is required to sustain buy-in while minimizing institutional friction. Leaders should adopt a variety of participation methods to integrate local expertise. Communication

requires the ability to convene and engage across disciplines, to convince others, and understand how to excite researchers to participate in ITD when doing so is outside their norms (Box 5).

Appropriate monitoring and evaluation

Properly evaluating ITD research remains a challenge. It may be tempting to set over-ambitious goals. Failure to achieve such goals demotivates researchers, distances stakeholders, and disappoints funders and clients of ITD organizations. Some examples of overpromising include fundraising across too broad a scope of activities, with none funded adequately; trying to do too many things, which leads to 'dropped balls' and disappointed partners; priming junior faculty for leadership, when such positions are not available; and relying on students to produce deliverables, but not informing the funder that this necessarily includes an education component that differs from a consultancy. Back-up support also needs to be available if students fail to complete a project. Ambitious goals can be valuable in motivating innovative ITD work, but appropriate expectations need to be set from the beginning and revisited frequently with internal and external stakeholders. Establishing a flexible, dynamic evaluation and monitoring framework as close as possible to the beginning phases of programs can greatly assist the management of ITD programs, freeing up time for leaders to pursue other responsibilities. In addition to evaluating program outputs and outcomes, the framework should evaluate the effectiveness of ITD processes themselves so that learning and development can take place in ITD teams (Holzer et al. 2018).

Perseverance

As sustainability programs and ITD research inherently challenge the status quo, effective leaders must be able to articulate a shared strategy and persevere against a tendency to regress to traditional, disciplinary approaches. The normative, practical nature of sustainability, its breadth of concerns, and its shifting or inexact definitions can invite skepticism from established scientific disciplines. The tendency for scientists to believe their own disciplines have higher value than other disciplines can also fracture ITD programs. All of these dynamics are acute in the early days of ITD program development.

Leaders who persevere and continuously communicate the value and role of ITD programs and research provide time for skepticism to erode, for disciplinary scientists to develop empathy for other ways of knowing, and for the creation of shared research, education, and outreach products that demonstrate the value of ID and TD (Kelly et al. 2019). Examples from Columbia University's Earth Institute, Arizona State University, and the University of Minnesota's Institute on the Environment illustrate the necessary perseverance around the establishment of new structures and celebration of their achievements, whereas the example from Baltimore Ecosystem Study (BES) illustrates perseverance within team processes (Box 5).

Resources for success

Resources needed to enable success in positions of leadership within ITD organizations fall into five categories: (1) intellectual resources; (2) institutional policies; (3) financial resources; (4) physical infrastructure; and (5) governing boards. First, leaders need to build and sustain mechanisms for recognizing and engaging intellectual expertise outside the disciplinary academic discourse (Bammer et al. 2020). This includes engaging all partners—those within one's home institution, other academics, and a broad array of stakeholders. Such engagement elicits new ideas, perspectives, and initiatives, contributing to the dynamism that is so important to ITD research. Tapping outside experts for short engagements through visiting appointments, internships, fellowships, post-docs, speakers, or program evaluators provides concentrated value and broadens reach and scope without the long-term budget commitments of adding permanent staff (Trimble and Plummer 2019).

Secure funding to support early career researchers, including doctoral students, post-doctoral fellows, and junior faculty is central for the longevity and success of ITD research. Many junior scholars, some trained in ITD, are attracted to the mission-oriented nature of ITD programs and institutes. They want to help solve sustainability problems and need roadmaps to consult. Traditional departmental training will not be sufficient to succeed in ITD scholarship without strong mentoring, explicit incentives to engage, and guidance on best practices. Graduate students and post-doctoral fellows should be given opportunities to share leadership, especially when their ITD training can facilitate multi-investigator and stakeholder projects that involve individuals with traditional, disciplinary training or single-issue agendas (Fam et al. 2020).

Second, leaders must be aware of the role of institutional infrastructure and how to foster policies that result in collaborative relationships, non-traditional outputs and outcomes, engagement with practitioners, celebration of ITD work, and career progression from recruitment to promotion. Columbia University's Earth Institute, for example, developed practice-oriented guidelines for appointment and promotion for its research scientists, with explicit guidance on new metrics and criteria for activities outside the scope of traditional research and how to judge them. Spokespersons for ITD must not be seen as competing for funds within the organization but as adding value to existing programs. Linking ITD

activities to the core culture of the institution can promote ITD work. As an example, courses co-taught by faculty from different disciplines or courses co-taught by tenured faculty and industry or non-profit professionals can lead to the co-production of novel approaches to solving topical, real-world problems.

Third, leaders need to operate based on the reality that many ITD research organizations are soft-money institutions. Long-term grants for ITD research are rare, so developing nimble ways to leverage limited budgets is critical. Experimenting with different seed funding for interaction and collaboration, such as those tied to specific outputs, can help expand into larger programs and broaden participation. Buying out faculty time or borrowing individuals for part of a year for leadership or collaborative activities can relieve constrained funding. Utilizing non-financial resources, such as staff time for proposal support, project management, or communications assistance, can also attract ITD participants from across and between institutions (Cundill et al. 2019).

However, it is important to be aware that proponents of disciplines may be openly hostile to ITD programs because they see them as direct competitors for funding. Attempts to compensate by 'buying' contributions from researchers in discipline-based departments are not always successful. Short-term income generation and time pressure are often achieved at the expense of longer term relationship building. Some organizations have found endowments to be key in allowing them to function, but maintaining a funding stream through endowments can bring its own challenges, depending on investment returns and broader economic conditions.

Fourth, the physical place and space of an ITD organization is vitally important. Co-location of scholars from different disciplines sparks serendipity—encouraging the hallway conversations and spontaneous brainstorming over coffee breaks—that inspires ITD work and reduces the need for formal meetings, seminars, and workshops (Lyall 2019). Where co-location is not possible, technology to engage distant partners electronically is an important aspect of the physical place. Co-location with external stakeholders can generate easy access to policymakers and facilitate the co-production of knowledge and solutions to real sustainability problems. One example is the Sustainable Cities Network, housed in the ASU Wrigley Institute, which brings together sustainability officers and other practitioners from municipalities and tribal governments from across the State of Arizona (https://sustainability.asu.edu/sustainable-cities/). The network identifies real-world sustainability problems as opportunities for research, education, and outreach. An example of an established ongoing program that resulted from this network is Project Cities, which links courses from across Arizona State University to solve specific community solutions, with monetary and other support from the participating cities (https://sustainability.asu.edu/project-cities/).

Finally, trustees, governing boards, or members of advisory bodies are important ITD resources. Supportive boards can advocate across their networks and help leaders motivate employees. However, if the Board is anchored in the past, represents legacy organizations, or is loyal to narrow disciplines, a leader must be steadfast in developing ITD strategy. Board members are often eminent leaders with large networks. However, their diversity and power require a subtle hand. They can be aloof, moderately engaged, or deeply involved depending on their defined responsibilities, individual interest, and how well the leader engages them. For example, leaders of Ashoka Trust for Research in Ecology & the Environment (https://www.atree.org/) have been deeply involved with board members as advisors, sounding boards, and fundraisers. Consequently, the organization has built a healthy endowment supporting core staff and functions. This endowment, partly gifted by the board, has allowed the institution to attract reputed faculty, take risks, and be innovative.

Conclusion: an inclusive framework for sustainability leadership

The work of ITD organizations is informed by theory and practice. Sustainability science has a rich and evolving canon and its work is equally motivated by practical concerns. Governments, non-governmental organizations, community groups reflecting different cultural backgrounds, and advocacy organizations all need ITD understanding of sustainability (Kates 2011).

The insights from our collective experience are tempered by the knowledge that the world is complex and rapidly changing. While we draw on diverse past trajectories, we acknowledge that the challenges of the future cannot be met based on past experience alone. The rapid proliferation of the coronavirus pandemic in early 2020 is a case in point. Surprises happen and ITD leaders need to be prepared to pivot, sometimes quickly, to meet changing priorities.

Our aggregate experience reflects many institutional contexts, practical motivations, and career paths. In addition, the variety of issues in sustainability we have addressed has exposed us to a wide range of approaches to education, research, engagement, and application. Our insights have also drawn on both our failures—addressed anonymously—and our successes, often summarized in the examples (Boxes 1–5). We hope this richness of experience can help those who will lead, or plan to organize, a transdisciplinary organization in the future. Our experience by no means reflect the full breadth of ITD challenges and successes, but the diversity of experiences represented in this group and the case studies we present in the boxes we believe has very real value.

The practical motivations of ITD work demand extensive consultation and stakeholder engagement. While an academic foundation is important, it is not enough for success. Indeed, the transdisciplinary practice of sustainability must be action-oriented, focusing on what people and institutions care about. ITD research and its implications must be understandable to all participants. Transparency, co-production of research and interventions, and communication that is effective for all stakeholders, are key attributes of the framework (Newton and Elliott 2016). At the same time, inter- and trans-disciplinary approaches provide opportunities for engaging diverse stakeholders and viewpoints, with the potential of increasing success of research to action by creating buy-in for a broad scope of participants (Belcher et al. 2019).

Inter- and transdisciplinary work must operate on various timeframes. Some participants may require near term actions, while other organizations may desire medium- to long-term outcomes. All participants should be aware and informed about the long-term implications of their sustainability decisions. Accordingly, inter- and transdisciplinary work must link multiple time scales.

Finally, the structures and practices of ITD work are not chiseled in stone. It must be possible to modify institutional goals and processes as needs change. Flexibility, a learning attitude, and open-mindedness focused on the future complete the framework for leadership of ITD organizations that can meet the challenges for a sustainable future.

Boxes for preparing interdisciplinary leadership for a sustainable future

Box 1

New ITD centers can be created by design, or established de novo to engage in ITD research activity. One case is the ZTG-Center for Technology and Society at the Technische Universität Berlin in Germany. It exemplifies an institution expressly designed to link important fields of research across disciplinary boundaries. It integrates social perspectives into the innovation and application of technology. The University has developed a strategy to foster transdisciplinary research supported by the ZTG.

A second example exists at Arizona State University (ASU), where President Michael Crow brought together leading thinkers in sustainability to a retreat to design a cross-university research institute dedicated to solving grand challenges. Following the retreat, the Julie Ann Wrigley Global Institute of Sustainability was founded and since 2004 has built a community of 540 Sustainability Scientists and Scholars spanning all 17 colleges at ASU. This transdisciplinary community is supported by staff trained in preparing ITD proposals. To underscore the mission-orientation of the institute, the Sustainability Scientists & Scholars are identified by strength of affiliation with the 17 UN Sustainable Development Goals (https://sustainability.asu.edu/sustainable-development-goals/). Although the institute has evolved over time, its success stems from careful and deliberate design from the beginning.

A third example is the Institute on the Environment at the University of Minnesota. In this case, faculty led an initiative to create a center for interdisciplinary scholarship, recognizing that solutions to environmental problems require collaboration across disciplines and with partners outside the university. That group of 11 senior faculty created the structure and placement of the institute within the university, and the proposal was supported and adopted by the university administration. More than a dozen years later, the institute now supports and enables more than 150 faculty from across the university—and select experts from outside the university. In addition to seeding research, it has taken on responsibility for developing skills in interdisciplinary and translational research, helping scholars of all ages and stages move beyond research on environmental *topics* to scholarship that affects environmental *outcomes*. Over time, the institute has embraced an active mission: to help build a future where people and planet prosper together.

Box 2. Origin by permanent or temporary merging of existing organizations

Organizations can also arise from mergers. Some may be permanent as in the case of the James Hutton Institute, founded in 2011 by merging two natural science institutes, one of which had some social and economic sciences. The vision for the more inclusive, new institute was one that fully embraced both natural and social sciences to tackle complex questions in new ways. It now has disciplines ranging from cell and molecular biology, through ecology, environment, geography, computational, social and economic sciences. Such a mix needs an understanding of what languages different groups use. One of the first leadership projects was to understand what everyone meant by 'interdisciplinarity' and how it represents many views. The internal project called 'Developing an Interdisciplinary Culture of Excellence (DICE)' (https://www.hutton.ac.uk/research/projects/dice) was aimed to improve understanding of interdisciplinary science

within the Institute and build capacity to undertake such research. The DICE project helped a great deal in surfacing views and setting a way forward by providing tools and examples of how to do ID research. There have also been experiments with structures ranging from matrix or cross-functional management to what is now a project-based organization. Our development of ID science is also driven by funders in the Scottish Government who demand interdisciplinary projects and even monitor outputs in terms of how many research products result from a combination of natural and social sciences. This helps in messaging the need to do things differently. The institute is known for its breadth and interdisciplinary work has been highly successful with other funders seeking ID solutions such as the EU Horizon 2020 programme.

An example of what are effectively temporary, 10 year mergers across existing organizations comes from New Zealand. In 2014 the government established eleven national science challenges to provide the science required to address complex long-term, national issues for New Zealand. These were intended to be mission-led, collaborative, and cross-institutional initiatives with a strong focus on science excellence and impact. Furthermore, they recognized a requirement for science to participate in transformational change if those fundamental national issues were to be resolved. In the case of the Our Land and Water (OLW), one of the eleven national science challenges, this means finding ways to decouple agricultural land use from adverse environmental impacts, recognizing that the country faces serious declines in land and water quality, and that agriculture, which is critical to New Zealand's economy, is not returning its maximum potential value to the country.

The drive for transformational impact has forced OLW to reflect on and respond to some key concepts and preconditions in the design and delivery of its research portfolio. Not the least of these has been the need to develop a better understanding of the economic, social, and cultural aspects of change, with an increasing emphasis on transdisciplinary methodologies. The Challenge has recognized that the way it undertakes research is fundamental to its relevance, accessibility, and to the speed of implementation. The leadership of the Challenge is embedding three facets of ITD thinking in research practice:

- The importance of co-design in problem definition and research design, and co-innovation in implementation to deliver greater impact faster;
- The critical part that Mātauranga, or indigenous knowledge systems and methodology, plays in enriching research and learning;
- The role of scientists in synthesizing, integrating and translating multiple strands of knowledge in ways that are meaningful to stakeholders and communities.

Challenge governance and management structure has evolved to encourage these practice shifts, with the development of cross-disciplinary leadership teams that have specific accountabilities for their delivery. They are also reinforced by the government funder of the Challenge, through a formal performance reporting system.

Box 3. Institutional and leadership evolution to move beyond the status quo

During its 23-year history, there have been two important transitions at Ashoka Trust for Research in Ecology and the Environment (ATREE). First, it expanded its initial focus on biodiversity to the interrelated themes of water and climate change. Second, in order for the knowledge ATREE generates to have an impact upstream on policy and downstream on action on the ground, the organization has developed two additional centers, a center for policy research and actions and a center for socio-environmental innovation and leadership. The purpose is to bridge the boundaries between research and policy on the one hand and research and action at the grassroots level on the other. These centers facilitate solution-oriented research. Developing consensus for both changes was not easy, and often it seemed that differences within and among faculty, the board, and the executive staff might tear the organization apart. But the ability of leadership to be patient, have open discussions, and respect various points of view had marked effect on changing minds and allowing the organization to keep its eyes on its mission and long term impact.

Box 4. Two cases of partner engagement

The close connection of the Earth System Science Centre and the Ministry of Science and Technology in Brazil has been instrumental in the implementation of the Brazilian Network for Climate Change Research (RedeCLIMA) and the System for Information and Analysis on Impacts of Climate Change (ImpactaClima), both scientific mechanisms to inform policy processes. Further, the Brazilian Platform on Biodiversity and Ecosystem Services emerged from a broad debate across

government (which was already engaged with the Global Intergovernmental Science Policy Platform on Biodiversity and Ecosystem Services), NGOs and the private sector. A series of meetings was held in which the expectations and potential use for the platform's research were discussed in depth. The platform is proving instrumental for the implementation of the first Biodiversity Synthesis Center in the country, the SINBIOSE (https://agencia.fapesp.br/brazil-to-have-a-biodiversity-synthesis-center-by-the-end-of-2018/29016/).

The Baltimore Ecosystem Study (BES) was based on mutually respectful partnerships from its inception. BES included not only social and natural scientists but leaders of the non-profit Parks & People Foundation, the Baltimore departments of Recreation and Parks, Public Works, and Planning. Additional partners included the community-based watershed associations in the Baltimore Region. Partners in the Baltimore County Department of Environmental Protection and Management, and the Maryland Department of Environment were also closely involved. As the issue of sustainability became of greater public concern, several government partners changed scope and mission. BES scientists were involved in the civic process driving evolution of these agencies, and the partnerships continue to be crucial.

Box 5. Communication and perseverance

As the examples demonstrate, communication and perseverance often go hand-in-hand. The Earth Institute at Columbia University was established to work across departments and schools throughout the university in order to address issues of sustainable development. Institute leaders have become skilled in navigating the operating structure of the university. This required communication and collaboration with deans, the provost, and other leaders to constantly advocate for and deliver the Institute's value-added to each constituency. This ongoing process of communication allows the Institute to attract students, faculty, and funding that departments might not have attracted on their own. Examples include developing and implementing a new major in Sustainable Development for the undergraduate college at the heart of the university, and fundraising for endowed chairs for faculty that reside in units other than the Institute.

When the School of Sustainability at ASU was established in 2006, there was excitement for what this new pursuit could bring. Yet there was also a good deal of skepticism on campus, ranging from the belief that sustainability was just a buzzword that lacked definition to the belief that students receiving a degree in sustainability would not get jobs. Continuous support from the university's president, the founding director of the school, external donors, and many committed faculty across campus gave the school the necessary time to create innovative programs not beholden to old disciplinary ways. When the degree programs opened, students flooded in, validating the school's value. The first group of graduates were nearly fully employed with many in sustainability-related careers. As sustainability programs expanded at other universities, the skepticism about the value of a sustainability college at ASU faded away. Without the perseverance of leadership, the school as a bold, transdisciplinary endeavor would not have had the chance to demonstrate its value.

Working within university power structures—to both challenge them and live within them—is a difficult part of running an ITD institute. Like Columbia's Earth Institute and ASU, the Institute on the Environment at the University of Minnesota has found communications essential to building a durable and effective interdisciplinary community. Those communications should celebrate the accomplishments of participants as a way to draw attention to the innovative ways they do their work and to increase their recognition and acclaim. Without this celebration, interdisciplinary achievements have a difficult time standing alongside more traditional approaches and standards. Further, to sustain the incentives for interdisciplinary and translational scholarship, institutes must have recurring and reliable funds, or else incentives for risk-taking and experimentation are lacking and the institute will fail to push the university in new, transformative directions.

Finally, interdisciplinary research is said to require a common language. The Baltimore Ecosystem Study (BES) found that shared terminology can sometimes be deceptive, tacitly connoting disparate ideas to those from different disciplines. Terms must be unpacked to reveal the disciplinary biases, different theoretical structures, and even the divergent practical motivations. BES participants found that it simply takes time to achieve this unpacking. Ultimately, the ITD project has produced shared meanings rather than a shared language. Perseverance through respectful, mutually open dialog among those who may come from different disciplines is the deep requirement.

Acknowledgements This work was supported by the National Socio-Environmental Synthesis Center (SESYNC) under funding received from the National Science Foundation DBI-1639145. The first meeting of leaders was hosted at SESYNC and we thank Professor Margaret Palmer, Director of SESYNC, and Dr Jonathan Kramer, Director for Interdisciplinary Science at SESYNC, for their participation and facilitation. The Santa Fe Institute supported and

hosted a follow-up workshop on "Tackling Complex Sustainability Issues: Lessons from Inter- and Transdisciplinary Organizations" which led to the production of this paper.

References

Alonso-Yanez G, House-Peters L, Garcia-Cartagena M et al (2019) Mobilizing transdisciplinary collaborations: collective reflections on *de* centering academia in knowledge production. Glob Sustain 2:e5. https://doi.org/10.1017/sus.2019.2

Annan-Diab F, Molinari C (2017) Interdisciplinarity: practical approach to advancing education for sustainability and for the Sustainable Development Goals. Int J Manag Educ 15:73–83. https://doi.org/10.1016/j.ijme.2017.03.006

Bammer G, O'Rourke M, O'Connell D et al (2020) Expertise in research integration and implementation for tackling complex problems: when is it needed, where can it be found and how can it be strengthened? Palgrave Commun 6:5. https://doi.org/10.1057/s41599-019-0380-0

Bark RH, Kragt ME, Robson BJ (2016) Evaluating an interdisciplinary research project: lessons learned for organisations, researchers and funders. Int J Proj Manag 34:1449–1459. https://doi.org/10.1016/j.ijproman.2016.08.004

Barrett MJ, Alphonsus KB, Harmin M et al (2019) Learning for transdisciplinary leadership: why skilled scholars coming together is not enough. Bioscience 69:736–745. https://doi.org/10.1093/biosci/biz072

Belcher BM, Claus R, Davel R, Ramirez LF (2019) Linking transdisciplinary research characteristics and quality to effectiveness: a comparative analysis of five research-for-development projects. Environ Sci Policy 101:192–203. https://doi.org/10.1016/J.ENVSCI.2019.08.013

Brinkley C, Raj S, Horst M (2017) Culturing food deserts: recognizing the power of community-based solutions. Built Environ 43:328–342

Brister E (2016) Disciplinary capture and epistemological obstacles to interdisciplinary research: lessons from central African conservation disputes. Stud Hist Philos Sci Part C Stud Hist Philos Biol Biomed Sci 56:82–91. https://doi.org/10.1016/j.shpsc.2015.11.001

Brown SA, Leinen MS, Strathdee SA (2019) Building a cross-disciplinary culture in academia through joint hires, degree programs, and scholarships. Strategies for team science success. Springer, Berlin, pp 489–494

Buizer M, Ruthrof K, Moore SA et al (2015) A critical evaluation of interventions to progress transdisciplinary research. Soc Nat Resour 28:670–681. https://doi.org/10.1080/08941920.2014.945058

Capra F (1983) The turning point: science, society, and the rising culture. Bantam Books, New York

Caves LSD (2020) On the emergence of interdisciplinary culture: the York Centre for Complex Systems Analysis (YCCSA) and the TRANSIT Project BT from astrophysics to unconventional computation: essays presented to susan stepney on the occasion of her 60th birthday. Springer International Publishing, Cham, pp 363–395

Cundill G, Harvey B, Tebboth M et al (2019) Large-scale transdisciplinary collaboration for adaptation research: challenges and insights. Glob Chall 3:1700132. https://doi.org/10.1002/gch2.201700132

Disterheft A, Caeiro S, Azeiteiro UM, Filho WL (2015) Sustainable universities: a study of critical success factors for participatory approaches. J Clean Prod. https://doi.org/10.1016/j.jclepro.2014.01.030

Fam D, Clarke E, Freeth R et al (2020) Interdisciplinary and transdisciplinary research and practice: balancing expectations of the 'old' academy with the future model of universities as 'problem solvers'. High Educ Q 74:19–34. https://doi.org/10.1111/hequ.12225

Frantzeskaki N, Rok A (2018) Co-producing urban sustainability transitions knowledge with community, policy and science. Environ Innov Soc Transit 29:47–51. https://doi.org/10.1016/J.EIST.2018.08.001

Gold M, Rauser C, Herzog M, Lueders J (2015) Sustainable LA Grand Challenge Five-Year Work Plan (Full Report). UCLA: Sustainable LA Grand Challenge. Retrieved from https://escholarship.org/uc/item/7v39j2xt

Gordon IJ, Bawa K, Bammer G et al (2019) Forging future organizational leaders for sustainability science. Nat Sustain. https://doi.org/10.1038/s41893-019-0357-4

Heinzmann J, Simonson A, Kenyon DB (2019) A Transdisciplinary approach is essential to community-based research with american indian populations. Am Indian Alsk Native Ment Health Res 26:15–41. https://doi.org/10.5820/aian.2602.2019.15

Herrero P, Dedeurwaerdere T, Osinski A (2019) Design features for social learning in transformative transdisciplinary research. Sustain Sci 14:751–769. https://doi.org/10.1007/s11625-018-0641-7

Holzer JM, Carmon N, Orenstein DE (2018) A methodology for evaluating transdisciplinary research on coupled socio-ecological systems. Ecol Indic 85:808–819. https://doi.org/10.1016/j.ecolind.2017.10.074

Huutoniemi K, Klein JT, Bruun H, Hukkinen J (2010) Analyzing interdisciplinarity: typology and indicators. Res Policy 39:79–88. https://doi.org/10.1016/j.respol.2009.09.011

Irwin EG, Culligan PJ, Fischer-Kowalski M et al (2018) Bridging barriers to advance global sustainability. Nat Sustain. https://doi.org/10.1038/s41893-018-0085-1

James Jacob W (2015) Interdisciplinary trends in higher education. Palgrave Commun 1:15001. https://doi.org/10.1057/palcomms.2015.1

Johnson J, Xenos MA (2019) Building better bridges: toward a transdisciplinary science communication. Tech Commun Q 28:112–123. https://doi.org/10.1080/10572252.2019.1583378

Kates RW (2011) What kind of a science is sustainability science? Proc Natl Acad Sci. https://doi.org/10.1073/pnas.1116097108

Kelly R, Mackay M, Nash KL et al (2019) Ten tips for developing interdisciplinary socio-ecological researchers. Socio-Ecolog Pract Res 1:149–161. https://doi.org/10.1007/s42532-019-00018-2

Kondo Y, Miyata A, Ikeuchi U et al (2019) Interlinking open science and community-based participatory research for socio-environmental issues. Curr Opin Environ Sustain 39:54–61. https://doi.org/10.1016/J.COSUST.2019.07.001

Lang DJ, Wiek A, Bergmann M et al (2012) Transdisciplinary research in sustainability science: practice, principles, and challenges. Sustain Sci. https://doi.org/10.1007/s11625-011-0149-x

Liu J, Mooney H, Hull V et al (2015) Systems integration for global sustainability. Science. https://doi.org/10.1126/science.1258832

Longino HE (1990) Science as social knowledge: values and objectivity in scientific inquiry. Princeton University Press, Princeton

Lyall C (2019) Being an interdisciplinary academic: how institutions shape university careers. Palgrave MacMillan, Cham

Matson P, Clark WC, Andersson K (2016) Pursuing sustainability: a guide to the science and practice. Princeton University Press, Princeton

McDaniels M, Skogsberg E (2017) The scholars we need: preparing transdisciplinary professionals by leveraging the scholarship of practice. New Dir High Educ 2017:71–83. https://doi.org/10.1002/he.20235

Mielke J, Vermaßen H, Ellenbeck S (2017) Ideals, practices, and future prospects of stakeholder involvement in sustainability science. Proc Natl Acad Sci 114:E10648–E10657. https://doi.org/10.1073/pnas.1706085114

Miller TR, Wiek A, Sarewitz D et al (2014) The future of sustainability science: a solutions-oriented research agenda. Sustain Sci. https://doi.org/10.1007/s11625-013-0224-6

Newton A, Elliott M (2016) A typology of stakeholders and guidelines for engagement in transdisciplinary. Participatory Processes Front Mar Sci 3:230

Palmer L (2018) Meeting the leadership challenges for interdisciplinary environmental research. Nat Sustain 1:330–333. https://doi.org/10.1038/s41893-018-0103-3

Palmer MA, Kramer JG, Boyd J, Hawthorne D (2016) Practices for facilitating interdisciplinary synthetic research: the National Socio-Environmental Synthesis Center (SESYNC). Curr Opin Environ Sustain 19:111–122. https://doi.org/10.1016/j.cosust.2016.01.002

Pickett STA, Kolasa J (2007) Ecological understanding: the nature of theory and the theory of nature. Academic Press, San Diego

Reid WV, Mooney HA (2016) The Millennium Ecosystem Assessment: testing the limits of interdisciplinary and multi-scale science. Curr Opin Environ Sustain 19:40–46. https://doi.org/10.1016/j.cosust.2015.11.009

Rosenzweig C, Solecki W (2014) Hurricane Sandy and adaptation pathways in New York: lessons from a first-responder city. Glob Environ Chang 28:395–408. https://doi.org/10.1016/j.gloenvcha.2014.05.003

Smetschka B, Gaube V (2020) Co-creating formalized models: participatory modelling as method and process in transdisciplinary research and its impact potentials. Environ Sci Policy 103:41–49. https://doi.org/10.1016/J.ENVSCI.2019.10.005

Stokols D (2006) Toward a science of transdisciplinary action research. Am J Community Psychol 38:63–93

Talley J (2016) A simplified approach to stakeholder engagement in natural resource management: the five-feature framework. Ecol Soc. https://doi.org/10.5751/ES-08830-210438

Trimble M, Plummer R (2019) Participatory evaluation for adaptive co-management of social–ecological systems: a transdisciplinary research approach. Sustain Sci 14:1091–1103. https://doi.org/10.1007/s11625-018-0602-1

Turner JR, Baker R (2020) collaborative research: techniques for conducting collaborative research from the science of team science (SciTS). Adv Dev Hum Resour. https://doi.org/10.1177/1523422319886300

Williams S, Robinson J (2020) Measuring sustainability: an evaluation framework for sustainability transition experiments. Environ Sci Policy 103:58–66. https://doi.org/10.1016/J.ENVSCI.2019.10.012

Willig MR, Walker LR (2016) 1 Changing the nature of scientists. Long-term ecological research. Oxford University Press, Oxford

Affiliations

Christopher G. Boone[1] · Steward T. A. Pickett[2] · Gabriele Bammer[3] · Kamal Bawa[4,5] · Jennifer A. Dunne[6] · Iain J. Gordon[7] · David Hart[8] · Jessica Hellmann[9] · Alison Miller[10] · Mark New[11] · Jean P. Ometto[12] · Ken Taylor[13] · Gabriele Wendorf[14] · Arun Agrawal[15] · Paul Bertsch[16] · Colin Campbell[17] · Paul Dodd[18] · Anthony Janetos[19] · Hein Mallee[20]

1 School of Sustainability, Arizona State University, 800 S. Cady Mall, Tempe, AZ 85284, USA

2 Cary Institute of Ecosystem Studies, Millbrook, NY, USA

3 Integration and Implementation Sciences, Research School of Population Health, The Australian National University, Canberra, ACT, Australia

4 University of Massachusetts, Boston, MA, USA

5 Ashoka Trust for Research in Ecology and The Environment, Bangalore, India

6 Santa Fe Institute, Santa Fe, NM, USA

7 Fenner School of Environment and Society, The Australian National University, Canberra, ACT, Australia

8 Senator George J. Mitchell Center for Sustainability Solutions, University of Maine, Orono, ME, USA

9 Institute On the Environment, University of Minnesota, Saint Paul, MN, USA

10 The Earth Institute, Columbia University, New York, NY, USA

11 African Climate and Development Initiative, University of Cape Town, Cape Town, South Africa

12 Earth System Science Centre (CCST-INPE)/Rede-Clima, São José dos Campos, Brazil

13 Our Land and Water National Science Challenge, Christchurch, New Zealand

14 Center for Technology and Society, Technische Universität Berlin, Berlin, Germany

15 International Forestry Resources and Institutions, University of Michigan, Ann Arbor, MI, USA

16 CSIRO Land and Water and Queensland Chief Scientist, Brisbane, QLD, Australia

17 The James Hutton Institute, Aberdeen, Scotland, UK

18 Office of Research, Interdisciplinary Research and Strategic Initiatives, University of California, Davis, CA, USA

19 Frederick S. Pardee Center for the Study of the Longer-Range Future, Boston University, Boston, MA, USA

20 Research Institute for Humanity and Nature, Kyoto, Japan

Perpustakaan Tun Abdul Razak
UNIVERSITI TEKNOLOGI MARA

SDI@PTAR
Selective Dissemination of Information

STAY CONNECTED
ANYWHERE | ANYTIME | ANYONE

Bil 4/2020

Higher Education Technology and Digital Transformation

BIL	JUDUL	AUTHOR	h-INDEX	UNIVERSITY	RANKINGS
1	Technology Intervention: Rethinking the Role of Education and Faculty in the Transformative Digital Environment	Katherine Rosenbusch	4	George Mason University, Fairfax Campus	No. 452 in World University Rankings

Article

Advances in Developing Human Resources
2020, Vol. 22(1) 87–101

DOI: 10.1177/1523422319886297
journals.sagepub.com/home/adh

Technology Intervention: Rethinking the Role of Education and Faculty in the Transformative Digital Environment

Katherine Rosenbusch[1]

Abstract

The Problem

Technology has affected almost every aspect of our lives, including education. Higher education is shifting the dynamics of delivery methods from traditional face-to-face to online to blended modes. Many universities are reaching a physical space capacity and therefore are attempting to increase online enrollment and geographical footprint. These changes are shifting the nature of higher education and how faculty are being viewed, evaluated, and, to some degree, hired. This article will focus on highlighting the challenges and opportunities of utilizing technology within universities, especially human resource development (HRD) programs.

The Solution

Technology is transforming higher education. Institutions can serve as an incubator to reimagine and redesign education altogether for the good of society. Online, mobile, and blended learning have become a part of our future. An important step is tracking how these models are actively enriching learning outcomes. Universities must be at the forefront of advancing progressive learning approaches and understanding the impact of technology on faculty and students.

The Stakeholders

The key stakeholders for this article include faculty, students, and university administrators. It will also affect businesses and human resource professionals for talent acquisition.

Keywords

technology, human resource development, blended and online learning

[1]George Mason University, Fairfax, VA, USA

Corresponding Author:
Katherine Rosenbusch, George Mason University, 4400 University Drive MSN 5F5, Fairfax, VA 22030, USA.
Email: krosenbu@gmu.edu

Technology has affected almost every aspect of our lives, including education. In this digital age, we are bombarded by technology. As Bates (2015) discussed in his book *Teaching in the Digital Age*, "technology is leading to massive changes in the economy, in the way we communicate and relate to each other, and increasingly in the way we learn" (p. 13). No longer is education about a teacher standing up in a classroom and lecturing; the format and access points have been transformed.

Higher education is shifting the dynamics of delivery methods from traditional face-to-face to online to blended (mixed) modes. More than 86% of traditional residential colleges and universities now offer online course options. One third of all degrees are now offered online (Online Schools Center, 2018). Universities are implementing new learning management systems (LMSs) and artificial intelligence (AI) is altering education as we know it. The changing of the guard is shifting from traditional instruction to technology-driven interactive online education. Faculty, students, and administration are all being affected by the digital age.

Electronic learning (E-Learning) is no longer just trendy, but a necessity. E-learning can be defined as "the use of computer network technology, primarily over an intranet or through the Internet, to deliver information and instruction to individuals" (Welsh et al., 2003, p. 246). Many universities are reaching a physical space capacity and consequently are attempting to increase online enrollment and geographical footprint. This strategy has implications for andragogical techniques, sustainability, and the redefinition of resources. These changes are shifting the nature of higher education and how faculty are being viewed, evaluated, trained, and, to some degree, hired. Human resource development (HRD) programs are no exception.

Many HRD faculty are feeling the crunch in terms of making decisions on how to best invest their time and money to stay competitive in the field, given the technology that is inundating them. Faculty are at a critical juxtaposition in this digital age to determine what factors are instrumental for learning and what components could be detrimental to the future of education. HRD faculty are being challenged to assess the value of HRD academic programs and determine the technological advancements needed to help drive the program's mission. Bing et al. (2003) stated that "HRD academic programs must continuously redesign curricula to ensure that HRD graduates leave with the most current knowledge and skills as well as a commitment to lifelong learning" (p. 347). This is now significantly being affected by technological design, delivery of instruction, and dissemination of research.

This article will focus on literature related to the technological trends that are affecting education and HRD, highlighting the challenges and opportunities of utilizing technology within universities, especially HRD academic programs. This article will build off the *Advances in Developing Human Resources Issue* by McWhorter and Bennett (2014) who discussed Virtual Human Resource Development (VHRD) and surveyed the limited research that has taken place regarding the impact of technology on students and faculty. In conclusion, concrete solutions will be offered for overcoming the challenges presented by the ever-changing platforms that faculty and students must navigate and discuss the key implications for HRD practice.

Technological Trends in Academia

Throughout the last decade, significant technological changes have shaken up the very essence of the field of HRD. In 2014, McWhorter and Bennett (2014) unveiled new perspectives that showcased some of the challenges that HRD professionals would face with the use of VHRD. Some of these included workplace technology, intranet opportunities to provide a learning organizational culture, sociomaterial perspective on technology, and trust in virtual teams. Since then, many of the perspectives have come to light and new high-tech initiatives have been developed. This article will showcase some of the trends that academic institutions have faced with the virtual world and technology. It will also highlight some of the challenges and opportunities that come with this technological movement and VHRD.

Blended Learning and Online Platform Designs

Over the past several years, perceptions of online learning have been shifting favorably as more learners and educators see online learning as a viable alternative to some forms of face-to-face learning (Adams Becker et al., 2017). Many universities and HRD programs are looking toward best practices to enhance learning and curriculum design.

Blended learning has become a popular trend as a way to leverage the digital platform for educational purposes. Blended learning is a "coherent design approach that openly assesses and integrates the strengths of face to face and online learning to address worthwhile educational goals" (Garrison & Vaughan, 2008, p. X). The advantages of blended learning include that learning can become more efficient and effective; students are able to pace themselves; teachers and students are more engaged; and retention of the content can increase (Lynch, 2018). This approach allows for ease of access, flexibility, and the integration of sophisticated multimedia and technologies. In addition, blending learning develops a webbed environment in which the new digital system becomes part of the organization (Bennett, 2014). If properly implemented, the dynamic relationship between faculty and students can be enriched and enhanced through the use of digital objects (audio, video, and text).

Research on blended and online learning has shifted to understanding the impact of digital modes on students. One of the primary purposes of blended education is to fuel learning both inside and outside the classroom. Like VHRD, blended learning creates an informal learning environment that allows the student to connect learning experiences in between the formal learning events (McWhorter, 2014). Current findings show an increase in creative thinking, independent study, and the ability for the student to tailor learning experiences to meet their individual needs (Adams Becker et al., 2017). Rasid and Asghar (2016) found that the use of technology has a direct positive relationship with students' engagement and self-directed learning but no significant effect on the student's academic performance. Joseph (2012) indicated that blended learning can lead to active learners who master their learning content and increase learner modes of critical thought ensuring students' growth at their own

level and use of techniques of multimedia applications and video application have greater collaboration skills and research skills. Most of the research indicates that it is too early to fully understand the long-term impact that technology is having on our students and there is a need to further research this area, especially in higher education settings.

New Technology Platforms

All faculty will need to continuously learn new skills in the face of an increasingly technological workplace (Sorcinelli, 2007). One question that comes to mind: How will higher education transform as AI and new digital platforms overtake some of the traditional functions? Ma and Siau (2018) discussed two major areas that will be affected in higher education by AI: curriculum and enrollment.

Universities are already using AI algorithms to personalize learning and deliver content that is suited to the students' needs and pace of learning (Alam & Kendall, 2018). Some faculty are using Augmented Reality (AR) in the classroom to revolutionize the learning experience and blend physical and digital reality (Delello et al., 2015). A few universities are providing machine learning to computer science majors, business students, education students, and corporate executives. Curriculum will continue to transform with emerging technology and many HRD faculty are working to keep up with the fast-paced change. AI may also affect the basis of university enrollment. More liberal arts and humanities majors will emerge because they are less susceptible to "AI-invasion" (Ma & Siau, 2018). Because of the potential of increased unemployment due to AI, higher education may no longer be affordable to many students. Petropoulos (2018) discussed the need for education and training programs to be redesigned so that they provide the right qualifications for work to interact and work efficiently alongside machines and boost relevant digital skills. The strength of AI is speed, accuracy, and consistency while it is weak on creativity, innovation, critical thinking, leadership, and empathy. As HRD's strength comes from the interpersonal skill base, it will be important to provide opportunities and training to students to enhance these components for future job placement. It is important for university students, academic institutions, and the field of HRD to remain abreast of the technology development with AI.

Many new avenues of education, including massive open online courses (MOOCs) and open educational resources (OER), are opening up possibilities to make higher education more available, affordable, and responsive to audiences that would otherwise not have access. An MOOC is defined as an online course aimed at large-scale interactive participation and open access via the web (Allen and Seaman, 2013). MOOCs have been one of the emerging themes in online learning in higher education. They have given rise to the online learning environment and increased the exposure to millions of people who might not have been touched by the curriculum. Many institutions are not likely to have MOOCs in the coming decades, but they are more likely to feel pressure to adopt MOOCs as a new instructional approach over time (Allen and Seaman, 2013). These tools could serve as

access point for outreach and community-engaged scholarship throughout the world.

OER are a different story. Many university libraries are finding OER to be a powerful tool for students and faculty. OER are defined as freely accessible and open licensed intellectual properties for teaching and learning, such as documents and media (Allen and Seaman, 2013). More than two thirds of academic leaders believe that OER have the potential to add value and reduce costs for their institution by saving the time and effort of developing new course materials (Allen et al., 2016; Lokken & Mullins, 2014). However, many academic leaders, faculty, and staff still lack understanding on how to find and utilize OER. Librarians are a great resource for both students and faculty when it comes to OER. Many universities are applying for national grants to assist with the implementation of OER into their curriculum. This area still remains to be discovered regarding the impact on institutions.

LMSs have become a critical tool for nearly all institutions of higher education, and a driving force of technology intervention. According to a 2017 report by the EDUCAUSE Center for Analysis and Research, 99% of higher education institutions have an LMS in place, and the LMS is used by 85% of faculty and 83% of students. In 2017, only six systems accounted for over 90% of LMS adoption by colleges and universities. These include Blackboard, Moodle, Canvas (Instructure), Brightspace (Desire2Learn), Sakai, and LearningStudio (Pearson) (Rhode et al., 2017). Studies of university students and their technology preferences have noted that nearly all students use an LMS and that the LMS is identified by students as among the most important instructional technologies for their academic success (Brooks, 2016; Dahlstrom et al., 2013). Many education futurists call for LMS tools and platforms to be more agile to support emerging instructional practices. There is a need to unbundle the components of a learning experience to remix open content and educational apps (Adams Becker et al., 2017; Anshari et al., 2016). M. Brown et al. (2015) believed that LMS platforms are too limited and propose "next-generation digital learning environment" (NGDLE), to support more personalized and flexible learning experiences. The evolution of tools to transform the learning experience will continue to advance and it will be up to institutions to keep up with the times to maintain their competitive advantage.

Challenges and opportunities. Many universities are turning to online and blended learning to compensate for decreased funding, increased enrollments, and technology growth and development. Some academic leaders expect that online education will compensate for a decrease in traditional course offerings by saving costs and improving the effectiveness of learning (Allen et al., 2016). The issue is that many universities are still experimenting and are in the infancy stages of implementing educational technologies and have failed to understand the key implications of these practices for the future of their educational programs. Moskal et al. (2013) cautioned educators about adopting too much too fast. They stated that a reliable and robust infrastructure must be in place to support students and faculty to be efficacious. HRD is no exception. Several programs have instituted portions of their traditional programs to blended learning while others have fully embraced online programs.

Anecdotally, many HRD faculty have experienced challenges with these new online and blended learning models (Rosenbusch et al., 2018). Some have been forced by university administrators to comply while others were given free rein to design and deliver their curriculum how they see fit. One of the challenges that has been faced by many programs is the generation gap that exists of early adopters versus laggards to technology. Some faculty members continue to resist the change that technology is bringing to institutions. Some new faculty are having to embrace the shift to meet the demands of the academic programs (Rosenbusch et al., 2018). HRD faculty have a stronger pedagogical base to work on embracing the technology and using it as an intervention to learning more than some other disciplines.

Studies have shown that blended learning and online education provide more flexibility and convenience than traditional educational opportunities. Students can utilize online assignments for on-demand practice and immediate feedback, and faculty can use student performance data from the assignments to tailor instruction (Horn et al., 2015). In the past, students had to travel to centers of learning but now massive amounts of information are available at one's fingertips through the internet, podcasts, MOOCs, Khan Academy, and traditional online degree programs (Purdue University, 2018). Distance delivery modes extend education to a global audience that might not have had the opportunities available to them from experts around the world.

One of the challenges with the increase of technological tools available to students, faculty, and institutions is the amount of support they receive to integrate the new platforms. Zheng et al. (2018) studied the impact of organizational and technical support in the faculty perceived benefits of using an LMS. They found that universities could increase the use of LMS and achieve more effective outcomes from faculty for online learning by structuring their organizations in a more succinct way to support faculty in technical areas. If universities are going to adopt state of the art technology, they must provide training and create support mechanisms for all involved.

Another opportunity may exist in the data universities are collecting in all these new technological systems. There is a likelihood that the components universities are implementing into the classroom will have predictive capabilities of student success. At this time, many faculty do not even know the information that is being acquired through their LMS platform. They do not know how to analyze the data or use it to improve their curriculum and predict student outcomes. Predictive analytics has been used for decades in the business world but only recently has been adopted by institutions of higher education. Predictive analytics is the process of analyzing and interpreting meaningful patterns from large amounts of data (Patil, 2015). Daniel (2014) addressed the need for higher education to begin to use Big Data to address the complex problems that they are facing for the future. He cautioned that institutions must begin to consolidate data in a more succinct manner to truly be able to utilize it for effective decision-making. It will become important for higher education to create data warehouses to retrieve information. HRD professionals can serve as the mediator for data scientists, faculty, and administration to bring transparency to learning analytics.

Shift in Research Dissemination

Other technological trends are affecting how research is being disseminated around the world. Academics have traditionally circulated their research through peer-reviewed journals but technology is opening up many new paths to publication and dissemination. Swist and Magee (2017) explored the constraints and potentials for academic publishing in the digital age. They discussed how the advancement of digital platforms amplifies the underlying tensions of institutional and scholarly change (Swist & Magee, 2017). This section will present some of the trends that are underway that have affected our research dissemination.

One trend that was established in 1998 by David Wiley was the "Open Content Project" (García-Peñalvo et al., 2010). The idea was aimed at the academic world and proposed a process to make sharing intellectual creations easier. The open movement began with the notion of sharing information and knowledge with the rest of the world. "Open" refers to "the fact of granting copyright permissions beyond those offered by standard copyright law" (García-Peñalvo et al., 2010, p. 521). Open practices in research have been described by Weller (2011) as "digital scholarship" or by Scanlon (2014) as "open scholarship." Open access to research is not just about disseminating resources but also about an opportunity to broaden and deepen the collective understanding of teaching, research, and practice (Iiyoshi & Kumar, 2008). This movement has created many new platforms online for academics and practitioners to share their knowledge with one another.

Another trend in this area is the use of social media to spread research. Social media is posing an interesting approach to enable researchers to communicate with one another around the world and spread their findings to viable research centers (Schnitzler et al., 2016). Greenhow and Gleason (2014) proposed to reconceptualize *social scholarship*, which is a new set of practices being discussed by several disciplines based off of Boyer's original framework which looked at scholarship in four dimensions: discovery, integration, teaching, and application. Researchers are beginning to see the value of avenues like Twitter, Facebook, Academia.edu, Google Scholar, ResearchGate, and LinkedIn to reach new audiences and share knowledge at an exponential rate (Schnitzler et al., 2016). Faculty are beginning to gauge their scholarly impact through social media platforms.

The origin for assessment of scholarly impact dates back to the 1920s when librarians used citation index factors to manage collections (LaBorie & Halperin, 1976). We have since evolved from this time. Some thought leaders are proposing a new framework to measure scholarly impact. A.Brown et al. (2016) developed a social network analysis to evaluate faculty productivity and scholarly impact that are promising for institutional administrations. The graphic depiction of research through the social network utilizes modern technology to measure faculty performance and productivity.

There is still much to be learned about this path of dissemination but many faculty are utilizing it to create their academic brand. Studies have shown that early career scholars are using social media in their professional lives for communication with peers and outside contacts to strengthen relationships and disseminate information and findings. They are also using various internet platforms to gauge feedback on their

research from the public (Gruzd et al., 2012). Social media will continue to be an opportunity for faculty to capitalize on to reach a larger audience and increase their scholarly impact.

Self-publishing is also a new trend where the digital scholar produces a range of informal, non-peer-reviewed papers of their work and research. Some of the venues for this have included blogs, tweets, and LinkedIn white papers (Scanlon, 2014; Schnitzler et al., 2016). Many scholars now have their own website and post the research they are working on to build collaboration and idea sharing. It has become an intriguing tool to capture research collaboration and propagation.

Research dissemination is an interesting challenge for academics. Many want their work to be widely accessible, but are bound by the constraints to obtain tenure, promotion, and raises. Many institutions still base their evaluations on peer-reviewed publications, and they rely on the publishers themselves not only to disseminate research but also to maintain a credible peer-review system. Self-publishing will continue to grow once the evaluation model changes in universities.

Challenges and opportunities. Technology has affected how we disseminate research. Faculty are being challenged in the publication realm by having more availability of open access journals. This change brings into question the credibility of the review process and the quality and rigor of the research. With the increase of technology and online social networks, journals are having a hard time competing with one another for authors' research. The reliability of some journals is also called into question. Some researchers are finding it harder and harder to get their results published in reputable venues.

Another challenge is circulation of one's research through social media, including Twitter, Facebook, LinkedIn, Google Scholar, and Academia.edu. Journals often limit to what extent an author can share their work through online platforms. This brings into question how we are truly measuring the impact of this research. One of the issues within the journals is the ability of the author to disseminate the research because of copyright laws. Faculty have to be careful when posting their papers to social media sites because of the infringement clauses they have agreed to with the publishers.

Digital scholars face the challenge of the proper protocol for open science (Masterman, 2016): "Practising open approaches in one's research includes openly licensing the methods, data and other artefacts that can enable others to reproduce the results reported" (Masterman, 2016, p. 34). Another issue that has arisen is the intellectual property rights of the author. Who actually owns the research the academic has produced? Many universities claim that because the researcher is employed by the institution, the findings are owned by them. It then becomes tricky on who can disseminate the findings on technology platforms.

Impact of Technology on Faculty and Students

Each of these technology trends places new demands on HRD and higher education. For example, what is the role of HRD in supporting and adopting technology? How

can HRD professionals blend technology with human processes to maximize learning? How does social responsibility influence the design and delivery of HRD educational programs and the dissemination of research?

Rapid developments in educational technologies mean that faculty and instructors need a strong framework for assessing the value of different technologies, new or existing, and for deciding how or when these technologies make sense for them and their students to use. This is a perfect opportunity for HRD scholars to connect their insights on instructional design and learning with change management strategies to incorporate the new trends of technology.

Building off of Bennett's (2014) IGO-Time model, HRD programs could integrate technology at multiple levels, various activities, and time perspectives to enhance the complexity in VHRD, making it more robust for future needs. Creating the optimal learning environment will take existing technology with rich media linked with the human elements at the individual, group, and organizational level. Because virtual environment is so complex and changing rapidly, it is critical that faculty and administration keep up with the necessary skills. It will be one of the roles of HRD professionals to act as a bridge between technologists and users (McWhorter & Bennett, 2014).

Daniel (2014) contended that the Big Data framework may be a way to address some of the key issues currently facing higher education, including the technological shift. Big Data can influence higher education practice, from enhancing the student experience to improved academic programming, to more effective evidence-based decision making, and to strategic response to changing global trends. It promises to turn complex, often unstructured data into actionable information. Daniel and Butson (2013) proposed a conceptual framework to describe Big Data in higher education along four components: institutional analytics, ITS analytics, learning analytics, and academic analytics. With the large volumes of student information—including enrollment, academic, and disciplinary records—universities could benefit from targeted analytics. Big Data and analytics in higher education could be transformative, altering the existing processes of administration, teaching, learning, academic work. It will take a different approach that must be embraced by various departments throughout an institution which could be complicated. HRD faculty could serve as facilitators to the change management process to link the data scientists to administration and faculty.

Based off the data analytics and digital structural changes in the classroom, faculty, staff, and administers will be challenged to maintain an influential learning environment. Blended and online learning, social media, and open learning are all developments that are critical for effective teaching in a digital age (Bates, 2015). However, HRD faculty must assess the overall impact the development of these courses and programs has on students, faculty, and universities.

Academic leaders in higher education institutions with online course offerings have consistently maintained a more positive view of the effectiveness of online education than those of institutions with no online course offerings (Allen et al., 2016). This reveals that there are positive correlations between exposure to and a positive view of online education. Over three quarters of academic leaders at public institutions report

that online is as good as or better than face-to-face instruction (compared with only 55.4% of private nonprofits and 67.0% of for-profits (Allen and Seaman, 2013).

There is still much to be learned about how the technology is helping or impeding the learning that is taking place on university campuses. Online learning does not always have the most positive outcomes like many institutions report. There is still room for improvement on how to determine whether the correct student is placed in the online environment. The use of technology is not for everyone, and as HRD professionals, we must be careful in how we measure the learning objectives in this new territory of education. Most of the current research in HRD has been in the practitioner realm with e-learning. It is now time to apply those very same principles to HRD programs.

HRD Implications and Intervention

It is inevitable that higher education will continue to change due to technology. It is how HRD faculty embrace that change to serve as an incubator to reimagine and transform education altogether for the good of society. Online, mobile, and blended learning have already begun to revolutionize our education system: "If institutions do not already have robust strategies for integrating these now pervasive approaches, then they simply will not survive" (Adams Becker et al., 2017, p. 2). Table 1 describes specific solutions to overcome the challenges associated with the latest technological trends specifically for the field of HRD and possible interventions for HRD professionals and academic institutions.

So what does this mean for the field of HRD and higher education? One important step in this transformation is tracking how these new models are actively enriching learning outcomes. Universities must be at the forefront of advancing progressive learning approaches. This often requires cultural transformation. Experts state that organizations should not implement technology unless they have a change management strategy (Biswas, 2018). It is important not to jump on the bandwagon of online education but be thoughtful in how the university undertakes this new initiative.

Institutions must be structured in ways that promote the exchange of fresh ideas, identify successful models within and outside of the campus, and reward teaching innovation—with student success at the center (Adams Becker et al., 2017). Faculty and administrators must be aware of the progressive views and changes regarding knowledge, skills, and abilities that the marketplace needs. HRD programs must stay at the forefront in designing curriculum that meets students' and employers' essentials for the shifting workplace demands. Students are expecting to graduate into gainful employment, which means that universities must prepare and develop real-world proficiencies to bolster their employability: "Institutions have a responsibility to deliver deeper, active learning experiences and skills-based training that integrate technology in meaningful ways" (Adams Becker et al., 2017, p. 2).

Educators must rethink how they design and deliver the curriculum. Faculty members often have to use different strategies than their traditional ways of teaching. Blended learning approaches are a good segue with technology before fully engaging

Table 1. Solutions to Overcome the Challenges Associated With Technological Trends.

Issue/challenge	Possible solution/intervention
Blended/online learning	• Ensure that HRD programs are adopting the new format for the right reason • Provide HRD faculty the necessary tools to properly redesign traditional courses • Investigate what digital tools are available at the university, department, and across other institutions • Collect data from HRD students to address impact of technology on learning • Understand implications for accreditation of HRD programs
New tools and technological platforms (i.e., MOOCs, OER, LMS)	• Explore possible MOOCs or OER materials to be embedded in the classroom and HRD curriculum • Explore strengths and weaknesses as a HRD instructor with the use of new technologies • Keep up to date on the latest trends in technology and how it can be used in HRD programs • Assess HRD students' level of knowledge of the digital platforms to ensure that all students have equal opportunity for success • Contact university teaching excellence centers to discover what resources are available • Stay abreast of Artificial intelligence (AI) and its impact on universities and the HRD field • Analyze how Big Data can help improve course design and instruction
Shift in research dissemination	• Begin formulating a digital footprint with research • Understand the university and journal policy for dissemination of HRD research. • Explore the impact of open source research versus traditional dissemination through peer-reviewed journals on evaluation process • If publishing in peer-reviewed journals, contact the publisher to see what can and cannot be posted to social media sites

Note. HRD = human resource development; OER = open educational resources; LMS = learning management system; MOOCs = massive open online courses.

online. This is also causing universities to redesign the learning spaces. Some of the changes have included the flipped classroom approach and educational settings using mixed reality technologies to increase active learning and incorporating innovation workspaces. To improve the overall system, many institutions are upgrading wireless bandwidth, and installing large displays that allow more collaborative spaces for student and faculty engagement. Faculty are having to make large pedagogical shifts through this rearrangement of physical spaces and must ramp up their technological skills to fully utilize the equipment.

Conclusion

The dynamics between the old guard (those slow to adopt technology) and the new guard (early adopters and pioneers of digital innovation) will continue to exist in university settings. Higher education institutions and administrators must work to bridge the rate of advancement with the needs of our students and faculty. How faculty respond to the technological trends will affect teaching, learning, and even research. HRD faculty are at a pivotal point to uncover the new dynamics of higher education as HRD professionals. We will be asked to embrace the new challenges and serve as facilitators to the learning process for many of our colleagues and students. It will not be easy but we could revolutionize how education is altered for the future of HRD.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Adams Becker, S., Cummins, M., Davis, A., Freeman, A., Hall Giesinger, C., & Ananthanarayanan, V. (2017). *NMC horizon report: 2017 higher education edition*. The New Media Consortium.

Alam, N., & Kendall, G. (2018). *Five ways artificial intelligence will shape the future of universities*. https://theconversation.com/five-ways-artificial-intelligence-will-shape-the-future-of-universities-94706

Allen, I. E., & Seaman, J. (2013). *Changing Course: Ten years of tracking online education in the United States*. Babso Survey Research Group. http://faculty.washington.edu/rvanderp/DLData/AllenSeaman2013.pdf

Allen, I. E., Seaman, J., Poulin, R., & Straut, T. T. (2016). *Online report card: Tracking online education in the United States*. Babson Survey Research Group. http://onlinelearningsurvey.com/reports/onlinereportcard.pdf

Anshari, M., Alas, Y., Yunus, N., Sabtu, N., Hamid, M., & Smith, M. (2016). Learning management system 2.0: Higher education. In M. M. Pinheiro & D. Simões (Eds.), *Handbook of research on engaging digital natives in higher education settings* (pp. 265–278). IGI Global.

Bates, A. W. (2015). *Teaching in a digital age: Guidelines for designing teaching and learning*. https://www.tonybates.ca/teaching-in-a-digital-age/

Bennett, E. (2014). Introducing new perspectives on virtual human resource development. *Advances in Developing Human Resources*, *16*(3), 263–280. https://doi.org/10.1177/1523422314532091

Bing, J. W., Kehrhahn, M., & Short, D. C. (2003). Challenges to the field of human resources development. *Advances in Developing Human Resources*, *5*, 342–351. https://10.1177/1523422303254671

Biswas, S. (2018). *HR technology in 2018: Opportunities, challenges and the future*. https://www.hrtechnologist.com/articles/digital-transformation/hr-technology-in-2018-opportunities-challenges-and-the-future/

Brooks, D. C. (2016). *ECAR study of undergraduate students and information technology, 2016* [Research report]. EDUCAUSE Center for Analysis and Research. https://er.educause.edu/~/media/files/library/2016/10/ers1605.pdf?la=en

Brown, A., Cowan, J., & Green, T. (2016). *Faculty productivity: Using social media and measuring its impact*. EDUCAUSE Review of Social Media for Teaching and Learning. https://er.educause.edu/articles/2016/5/faculty-productivity-using-social-media-and-measuring-its-impact

Brown, M., Dehoney, J., & Millichap, N. (2015, July/August). *What's next for the LMS?* [EDUCAUSE Review]. https://er.educause.edu/articles/2015/6/whats-next-for-the-lms

Dahlstrom, E., Walker, J. D., & Dziuban, C. (2013). *ECAR study of undergraduate students and information technology, 2013* [Research report]. EDUCAUSE Center for Analysis and Research. https://library.educause.edu/resources/2013/9/ecar-study-of-undergraduate-students-and-information-technology-2013

Daniel, B. K. (2014). Big Data and analytics in higher education: Opportunities and challenges. *British Journal Education Technology*, *46*(5), 904–920. https://doi.org/10.1111/bjet.12230

Daniel, B. K., & Butson, R. (2013). Technology enhanced analytics (TEA) in higher education. In Proceedings of the *international conference on educational technologies* (pp. 89–96).

Delello, J. A., McWhorter, R. R., & Camp, K. M. (2015). Integrating augmented reality in higher education: A multidisciplinary study of student perceptions. *Journal of Educational Multimedia and Hypermedia*, *243*(3), 209–233.

García-Peñalvo, F. J., García de Figuerola, C., & Merlo, J. A. (2010). Open knowledge: Challenges and facts. *Online Information Review*, *34*(4), 520–539. https://doi.org/10.1108/14684521011072963

Garrison, D. R., & Vaughan, N. D. (2008). *Blended learning in higher education: Framework, principles and guidelines*. John Wiley.

Greenhow, C., & Gleason, B. (2014). Social scholarship: Reconsidering scholarly practices in the age of social media. *British Journal of Educational Technology*, *45*(3), 392–402. https://doi.org/10.1111/bjet.12150

Gruzd, A., Staves, K., & Wilk, A. (2012). Connected scholars: Examining the role of social media in research practices of faculty using the UTAUT model. *Computers in Human Behavior*, *28*, 2340–2350. https://doi.org/10.1016/j.chb.2012.07.004

Horn, I. S., Kane, B. D., & Wilson, J. (2015). Making sense of student performance data: Data use logics and mathematics teachers' learning opportunities. *American Educational Research Journal*, *52*(2), 208–242. https://doi.org/10.3102/0002831215573773

Iiyoshi, T., & Kumar, M. S. V. (2008). Conclusion: New pathways for shaping the collective agenda to open up education. In T. Iiyoshi & M. S. V. Kumar (Eds.), *Opening up education: The collective advancement of education through open technology, open content, and open knowledge* (pp. 429–440). MIT Press.

Joseph, J. (2012). The barriers of using education technology for optimizing the educational experience of learners. *Procedia-Social and Behavioral Sciences*, *64*, 427–436. https://doi.org/10.1016/j.sbspro.2012.11.051

LaBorie, T., & Halperin, M. (1976). Citation patterns in library science dissertations. *Journal of Education for Librarianship*, *16*(4), 271–283.

Lokken, F., & Mullins, C. (2014). *Trends in e-learning: Tracking the impact of e-learning at community colleges*. Instructional Technology Council.

Lynch, M. (2018). *5 major benefits of blended learning*. https://blogs.edweek.org/edweek/education_futures/2018/05/5_major_benefits_of_blended_learning.html

Ma, Y., & Siau, K. L. (2018). Artificial intelligence impacts on higher education. In *MWAIS Proceedings* (p. 42). https://aisel.aisnet.org/mwais2018/42

Masterman, E. (2016). Bringing open educational practice to a research-intensive university: Prospects and challenges. *Electronic Journal of e-Learning*, *14*(1), 31–42.

McWhorter, R. (2014). A synthesis of new perspectives on virtual HRD. *Advances in Developing Human Resources*, *16*(3), 391–401. https://doi.org/10.1177/1523422314532126

McWhorter, R., & Bennett, E. (Eds.). (2014). Introducing new perspectives on virtual human resource development. *Advances in Developing Human Resources*, *16*(3), 263–280. https://doi.org/10.1177/1523422314532126

Moskal, P., Dziuban, C., & Hartman, J. (2013). Blended learning: A dangerous idea? *The Internet and Higher Education*, *18*, 15–23.

Online Schools Center. (2018). *The state and challenges of online education*. https://www.onlineschoolscenter.com/challenges-of-online-education/

Patil, S. (2015). Predictive analytics in higher education. *International Journal of Advanced Research in Computer and Communication Engineering*, *4*(12), 138–141. https://doi.org/10.17148/IJARCCE.2015.4123

Petropoulos, G. (2018). The impact of artificial intelligence on employment. In *Work in a digital age*. Rowman & Littlefield. https://bruegel.org/2018/07/the-impact-of-artificial-intelligence-on-employment/

Purdue University. (2018). *How has technology changed education?* https://online.purdue.edu/ldt/learning-design-technology/resources/how-has-technology-changed-education

Rasid, T., & Asghar, H. M. (2016). Technology use, self-directed learning, student engagement and academic performance: Examining the interrelations. *Computers in Human Behavior*, *63*, 604–612. https://doi.org/10.1016/j.chb.2016.05.084

Rhode, J., Richter, S., Gowen, P., Miller, T., & Wills, C. (2017). Understanding faculty use of the learning management system. *Online Learning*, *21*(3), 68–86.

Rosenbusch, K., Morrison, E., & Shuck, B. (2018, February 17). *The changing of the guard . . . How do we plan for the academy's future needs* [Conference session]. Presented at the Meeting of the Academy of Human Resource Development Conference, Richmond, VA.

Scanlon, E. (2014). Scholarship in the digital age: Open educational resources, publication and public engagement. *British Journal of Educational Technology*, *45*, 12–23.

Schnitzler, K., Davies, N., Ross, F., & Harris, R. (2016). Using Twitter™ to drive research impact: A discussion of strategies, opportunities and challenges. *International Journal of Nursing Studies*, *59*, 15–26. https://doi.org/10.1016/j.ijnurstu.2016.02.004

Sorcinelli, M. D. (2007). Faculty development: The challenge going forward. *Peer Review*, *9*(4). https://www.aacu.org/publications-research/periodicals/faculty-development-challenge-going-forward

Swist, T., & Magee, L. (2017). Academic publishing and its digital binds: Beyond the paywall towards ethical executions of code. *Culture Unbound: Journal of Current Cultural Research*, *9*(3), 240–259. http://www.cultureunbound.ep.liu.se/v9/a17/cu17v9a17.pdf

Weller, M. (2011). *The digital scholar: How technology is transforming scholarly practice*. Bloomsbury Academic. https://doi.org/10.5040/9781849666275

Welsh, E. T., Wanberg, C. R., Brown, K. G., & Simmering, M. J. (2003). E-learning: Emerging uses, empirical results and future directions. *International Journal of Training and Development*, *7*, 245–258. https://doi.org/10.1046/j.1360-3736.2003.00184.x

Zheng, Y., Wang, J., Doll, W., Deng, X., & Williams, M. (2018). The impact of organisational support, technical support, and self-efficacy on faculty perceived benefits of using learning management system. *Journal of Behaviour and Information Technology*, *37*(4), 311–319. https://doi.org/10.1080/0144929X.2018.1436590

Author Biography

Katherine Rosenbusch is an assistant professor of management in the School of Business, George Mason University. She has experience in training and development with private and nonprofit organizations, including 3 years living in Indonesia. It was this international experience that inspired her current research in cross-cultural adjustment, international human resources, and global leadership development. Her research has been published in *Human Resource Development International* and *Cross-Cultural Management: An International Journal*. She is currently researching the impact of technology. Her experience teaching online, face to face, and blended for the past decade peaked her interest in uncovering the impact and solutions to the online learning space. She received her bachelor's and master's from Texas A&M University and her doctorate from the George Washington University.

Bil 4/2020

Higher Education Teaching and Learning

BIL	JUDUL	AUTHOR	h-INDEX	UNIVERSITY	RANKINGS
1	Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform	Ruipérez-Valiente, José A.	13	Massachusetts Institute of Technology	No. 1 in World University Rankings

Computers & Education 154 (2020) 103894

Contents lists available at ScienceDirect

Computers & Education

journal homepage: http://www.elsevier.com/locate/compedu

Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform

Felix Garcia-Loro [a,*], Sergio Martin [a], José A. Ruipérez-Valiente [b], Elio Sancristobal [a], Manuel Castro [a]

[a] Spanish University for Distance Education, of Electrical and Computer Engineering Department, Juan del Rosal 12, 28040, Madrid, Spain
[b] Massachusetts Institute of Technology, Department of Comparative Media Studies, 77 Massachusetts Ave, Cambridge, MA, 02139, USA

ARTICLE INFO

Keywords:
Peer assessment
MOOCs
Krippendorff's alpha
Inter-rater reliability (IRR)
Reliability

ABSTRACT

Peer assessment activities might be one of the few personalized assessment alternatives to the implementation of auto-graded activities at scale in Massive Open Online Course (MOOC) environments. However, teacher's motivation to implement peer assessment activities in their courses might go beyond the most straightforward goal (i.e., assessment), as peer assessment activities also have other side benefits, such as showing evidence and enhancing the critical thinking, comprehension or writing capabilities of students. However, one of the main drawbacks of implementing peer review activities, especially when the scoring is meant to be used as part of the summative assessment, is that it adds a high degree of uncertainty to the grades. Motivated by this issue, this paper analyses the reliability of all the peer assessment activities performed as part of the MOOC platform of the Spanish University for Distance Education (UNED) UNED-COMA. The following study has analyzed 63 peer assessment activities from the different courses in the platform, and includes a total of 27,745 validated tasks and 93,334 peer reviews. Based on the Krippendorff's alpha statistic, which measures the agreement reached between the reviewers, the results obtained clearly point out the low reliability, and therefore, the low validity of this dataset of peer reviews. We did not find that factors such as the topic of the course, number of raters or number of criteria to be evaluated had a significant effect on reliability. We compare our results with other studies, discuss about the potential implications of this low reliability for summative assessment, and provide some recommendations to maximize the benefit of implementing peer activities in online courses.

1. Introduction

Last century's recent changes on educational paradigms have promoted the integration of new evaluation methods that intend to advance beyond the classical knowledge assessment (summative assessment) as its only grading goal. This new mindset aims to develop evaluation methods that are more embedded within the training and learning process in what is known as formative assessment (Dochy, Segers, & Sluijsmans, 1999; Earle, 2014; Guan-Yu Lin, 2018). Formative assessment can have a significant impact on the quality of learning that students experience by practicing the required skills in advance, and by helping them to be more self-aware of their current status, but also for instructors so that they can have just-in-time feedback regarding how the class is

* Corresponding author.
E-mail addresses: fgarcialoro@ieec.uned.es (F. Garcia-Loro), jruiperez@um.es (J.A. Ruipérez-Valiente).

https://doi.org/10.1016/j.compedu.2020.103894
Received 17 September 2019; Received in revised form 6 March 2020; Accepted 5 April 2020
Available online 29 April 2020

progressing (Topping, 2017; Van der Pol, Van den Berg, Admiraal, & Simons, 2008). In fact, assessment is now conceived as a central part of the learning process, of which the student has become more responsible (Black & Wiliam, 2009; Dochy et al., 1999; Dochy & McDowell, 1997; Kilic, 2016). This new paradigm can be interpreted as trying to shift from the consolidated idea of assessment as the final goal of the learning process, to a paradigm where assessment is just one of the many tools and options (Sluijsmans, Brand-Gruwel, van Merriënboer, & Bastiaens, 2002b). Furthermore, in today's society where information is easily available and where AI is called to take over tasks that are easy to automate, higher education institutions have acknowledged the need to train students to develop more transverse skills, given that they will face a more and more uncertain future carrying out work responsibilities that might still not exist (Boud, 2000; Marton & Bowden, 1999; Susskind & Susskind, 2015).

From the very beginning, the European Higher Education Area (AHEA) has been watching over the implications of this on-going educational shift. However, it did not start talking about student-centered learning until 2009, in a meeting which took place at Leuven/Louvain-la-Neuve (European Higher Education Area, 2009). Besides, AHEA's present educational model is based on competences (de Miguel, Alfaro, Apodaca, Arias, García, & Lobato, 2005), and so the current speech is focusing now on 'competence alignment' or 'constructive alignment'. The new emphasis on student-centered learning and competences, together with the Information and Communications Technology (ICT) democracy, has facilitated the creation of new pedagogical approaches or boosted the use of underused ones, by promoting a redesign of the learning scenario (Beldarrain, 2006); some examples that have received a lot of attention include collaborative learning (Van Den Bossche, Gijselaers, Segers, & Kirschner, 2006), self-regulated learning (Boekaerts & Corno, 2005), collaborative inquiry learning (Bell, Urhahne, Schanze, & Ploetzner, 2010), competence-based learning (Benlloch--Dualde & Blanc-Clavero, 2007), personalized learning (Chen, 2008), differentiated learning (Lawrence-Brown, 2004), active learning (Gauci, Dantas, Williams, & Kemm, 2009), flipped learning (Lukassen, Pedersen, Nielsen, Wahl, & Sorensen, 2014), instructional scaffolding (Quintana, Reiser, Davis, Krajcik, Fretz, Duncan et al., 2004), problem-oriented and project-based learning(Lehmann, Christensen, Du & Thrane, 2008), and so on. These approaches can be combined in order to achieve an effective metacognitive learning that can prepare better students for efficient lifelong learning (Cornford, 2002; Lüftenegger et al., 2012; Weinstein, Acee, & Jung, 2011). It is with the implementation of these new methodologies that evaluation has ceased being an isolated activity carried out at the end of the learning process and it is now frequently integrated more seamlessly in the learning process, and it is regarded as yet another tool for its success. According to Delgado, Borge, García-Albero & Salomón (2005), evaluation now intends to assess the quality of learning the student has developed; it is no longer based on products, but rather, on processes.

One of the tools favored by the new perspectives on educational plans has been peer assessment or peer review tasks. In this sense, Falchikov & Goldfinch consider that "peer assessment is grounded in philosophies of active learning and andragogy, and may also be seen as being a manifestation of social constructionism, often involves the joint construction of knowledge through discourse". According to Duran (2017) "the first reviews and meta-analyses on peer tutoring revealed evidence of learning by the tutor in their role of 'teacher'". Moerkerke (1996) and Dochy et al. (1999) share the idea that peer assessment activities are compatible with a society of lifelong learners.

The area of learning at scale presents massive online scenarios, such as MOOCs among others, that require alternative approaches in order to implement learning and assessment approaches that target many learners at the same time. In order to provide a learning design that is sustainable and can scale to large numbers of learners, formative assessment cannot be dependent on direct feedback from teachers. Therefore, for those classes where formative assessment is a crucial part of the learning process, peer assessment turns into a tool with huge potential to solve the issue of scale. This article analyses the reliability of peer assessments developed specifically under MOOC environments. It focuses on the consistency of students as raters, by studying Inter-Rater Reliability (IRR). In addition, we aim to assess the validity of the obtained evaluations in our specific framework, taking into account our limitations. For these analyses, we have gathered the data of all the peer assessment activities carried out on UNED's MOOC platform (http://coma.uned.es/). These courses are highly diverse, being related to different knowledge areas, subjects and levels (Capdevila & Aranzadi, 2014). MOOCs have proved to be successful non-formal open learning environments (Hood, Littlejohn, & Milligan, 2015), where students' motivation and self-regulation capabilities are key factors. For those reasons, MOOCs are an optimal resource for knowledge transference in our current society. Nevertheless, and in spite of the many developments on virtual tutoring, the massive nature of MOOCs limits the type of activities that can be implemented. Specifically, activities that do not scale to a high number of students (e.g., a teacher providing individualized feedback to each assignment), cannot be implemented in these environments (Suen, 2014). As many other learning activities, peer assessment generally implies receiving a score, which could potentially be used as part of the summative grade. Therefore, in this manuscript we explore the reliability and validity of scores generated through peer assessment activities, in order to evaluate whether it would be appropriate to use these scores as part of a weighted final grade. The data we analyze have been gathered based on the assessment that students performed on the activities of their peers. Both tasks, submitting an activity, and peer reviewing someone else's work, are mandatory on the platform. Consequently, our purpose was to obtain a data sample large enough to analyze the consistency of the assessments according to multiple observers in different courses and activities. For this purpose, we have collected a high number of valid submitted tasks (more than 27,000), reviews (more than 93,000) and criteria assessed (almost 334, 000), conferring a solid background to the results and conclusions obtained in this analysis. Overall, the research question that has concerned us in this study is the following:

RQ: Are peer assessments reliable in a typical MOOC environment like the one provided by UNED platform?

2. Literature review

Peer assessment can be described and implemented in many different ways. The number of studies and diversity of educational contexts suggest that peer assessment can be, practically, applied to all areas of knowledge (Topping, 1998). As an assessment

approach, peer assessment has traditionally been considered valid or not, by confronting students' and teachers' grades (Cho, Schunn, & Wilson, 2006; Falchikov and Goldfinch, 2000; Formanek, Wenger, Buxner, Impey, & Sonam, 2017; Jackson, 2014; Jones & Alcock, 2014; Stefani, 1994; Sung, Chang, Chang, & Yu, 2010), despite the fact that the core objective of peer assessment is to actually create opportunities for peers to learn from each other and to participate more in the learning process. This correction over students' evaluation has been called 'validity', while we use the term 'reliability' to determine the consistency among peer ratings (Jackson, 2014; Luo, Robinson, & Park, 2014; Richmond et al., 1992).

This section is meant to frame peer assessments and, more specifically, their reliability. It does so by starting from a general point of view up to its specific impact on MOOCs.

2.1. Definition of peer assessment

Several authors have provided broad definitions, conceptually talking, for peer assessment. For example, Fachikov & Goldfinch (2000) highlight that, when students use them, they "judge the work of their peers". This view is similar to Reinholz (2016) although he talks about evaluating others. Orsmond, Merry, and Reiling (1996) refer to peer assessment as a learning tool and Van Zundert, Sluijsmans, and Van Merriënboer (2010) focus their argument on its not necessarily bidirectional reciprocity. According to them, the goal is to "evaluate or be evaluated by peers". Topping (1998; 2009) includes the concept of learning through peer assessment in his definition: "Peer-assessment is an arrangement for learners to consider and specify the level, value, or quality of a product or performance of other equal-status learners". Van der Pol et al. (2008) provide a broad definition which includes every step carried out on peer assessments, described as an activity. They talk about the pre-established criteria that the student must stick to, as well as the requirements of a critical evaluation that includes feedback (formative assessment) for the evaluated student. In their words, "students engage in reflective criticism of the products of other students and provide them with feedback using previously defined criteria". De Grez Valcke & Roozen (2012a,b) use the term 'peer assessment' on a test in which they invited students from a more advanced course to act as raters. To some extent, they might be considered peers, but this implementation misses the point where a student is rating a piece of work the student has already completed. Consequently, the cognitive process that involves personal reflection and self-criticism is lost.

On this paper, we consider as 'peers' the students of each course who are registered and active in each evaluated tasks. This implies that they all have carried out the task before engaging in the peer assessment activity. They find themselves in a position of equality towards the task and hence we can effectively consider them as peers based on the previous definitions provided by Topping (1998, 2009) and based on the idea of "other equal-status learners".

2.2. Peer assessment and its integration in MOOCs. Implications for reliability and validity

MOOCs usually implement assessment methods that do not require manual correction by the instructors, usually, these are generally known as auto-graded tools (machine-assessment): single choice and multiple-choice items are particularly common; as well as fill-in the blanks, with a number, a word or even a sentence. Other more nuanced auto-graded items include programing environments where students code their solution and the system expects an specific function output, or specific tools that can be integrated with the MOOC platform through authentication protocols such as LTI protocol (Alcarria, Bordel, Andres, & Robles, 2018; Aleven, Sewall, Popescu, Xhakaj, Chand, Baker et al., 2015; Garcia-Loro, Sancristobal, Gil, Diaz, Castro. Albert-Gómez, 2016; Garcia-Loro, San Cristobal, Diaz, Macho, Baizan, Blazquez, et al., 2018; Mullen, Byun, Gadepally, Samsi, Reuther, & Kepner, 2017). There have also been some limited advances in auto-grading essays (Ambekar & Phatak, 2014). Auto-graded assessment instruments have high validity, but they are quite limited in what they can assess and the cognitive process of students solving them is very low, which can be especially critical in some areas of knowledge. In order to improve and support students' learning, it is essential to include feedback information that can help students understand where they are at in their learning process and their potential misconceptions.

Peer evaluation, besides the reliability and validity of its methodology, can provide this sort of beneficial personalized feedback to every single one of the otherwise unmanageable number of students in MOOCs. Furthermore, it is a well-aligned contribution to the current educational perspectives that locate the student in the center of the whole learning process (Suen, 2014; Van Hattum-Janssen & Lourenço, 2008). Finally, the exercise of acting as an evaluator can enact more complex cognitive processes that favor deeper learning for students (Hsia, Huang, & Hwang, 2016).

With regard to the typical learning environments in MOOCs, while traditional learning contexts can assume a high similarity degree in the background of their learners, the 'Open' nature of MOOCs highly increments the diversity in learners' profiles, hence potentially breaking the equality among learners' condition. In MOOCs we find that learners have multiple backgrounds in content knowledge (especially those regarding STEM), diverse sets of skills related to writing, text comprehension, synthesis and very different intentions when enrolling in a MOOC (Alario-Hoyos, Pérez-Sanagustín, Delgado-Kloos, Parada, & Muñoz-Organero, 2014; Watson, Watson, Yu, Alamri, & Mueller, 2017). This characteristic heterogeneity in students' profiles collides even more with the assumption of equity among peers.

2.2.1. Feedback

Feedback is undoubtedly the core mechanism in peer assessment to become formative (Thelwall, 2000; Gipps, 2005; Miller, 2009; Nicol & Macfarlane-Dick, 2006; Ng, 2014). When correctly implemented, peer assessment involves students in both feedback roles: as evaluators, by contributing with ideas and comments to the assessed tasks, as well as evaluates, by receiving peers' observations with constructive comments to improve their own work (Ng, 2014). This sort of assessment usually coexists with the summative ones,

although it can appear on its own. Nevertheless, it is recommended that formative assessment goes alongside with the summative one (Gipps, 2005; Miller, 2009; Nicol & Macfarlane-Dick, 2006). In this sense, Ng (2014) highlights the importance of students receiving tailored feedback instead of just receiving scores. Feedback and feedforward strategies are used in critical learning (Cartney, 2010; Kilic, 2016) as well as in social learning (Guan-Yu Lin, 2018). These tools stand out in peer assessment because they help the student develop analytical thinking, critical thinking and deeper knowledge development. However, students must be well prepared and highly motivated to be capable of developing this task (Winstone, Nash, Parker, & Rowntree, 2017). On the other hand, students enrolled in MOOCs tend to be from a broad spectrum of educational backgrounds, they can have diverse levels of initial knowledge, different intended learning objectives and different self-regulated learning patterns. Such diversity in MOOC students, and, therefore, in raters, can undermine the underlying assumption of "equality" in peer assessment methodologies (Meek, Blakemore, & Marks, 2017).

2.2.2. Assessment criteria and rubrics

Dochy et al. (1999) highlight the importance of establishing clear assessment criteria: "it should be clear that students have to know the criteria clearly … criteria should include information about the area to be assessed, the aims to be pursued and the standards to be reached". In this sense, Falchikov and Goldfinch (2000) in their meta-analysis have found that the reliability and validity of peer assessment is positively correlated with the establishment of a clear assessment criteria. They also found that peer assessment tasks requiring several independent scoring dimensions were less valid than peer assessment tasks based on a global judgement. In this context, Sadler and Good (2006) as well as Meletiadou and Tsagari (2014) stated that "five or fewer criteria increase reliability". Nonetheless, studies like the one carried by Jones and Alcock (2014) based on comparative judgment (Thurstone, 1927), consider that evaluation criteria are not a necessary condition for reliable and productive peer assessment; instead, they consider that students feel stimulated as raters if they have more freedom to develop their own assessments. Furthermore, it would further promote their abilities, critical thinking and sense of responsibility.

Although traditionally teachers' and experts' grades are considered as the valid ones (Cho et al., 2006; Falchikov and Goldfinch, 2000; Formanek et al., 2017; Jackson, 2014; Stefani, 1994; Sung et al., 2010), authors such as Piech et al. (2013) state that the "true mark" is not necessarily the teachers' one; they propose to distance teacher's rubric and its validity. To avoid this dichotomy in the "true grade" (teachers' vs students' grading), and also to improve validity, several authors have highlighted the benefits of training in the reviewing mechanism (Formanek et al., 2017; Meletiadou & Tsagari, 2014; Sadler & Good, 2006; Sluijsmans et al., 2002b; Topping 2009, 2017; Van Zundert et al., 2010). Furthermore, many studies have involved students in the definition and development of the assessment criteria in order to improve assessment results and students' involvement in the activity (Falchikov, 2013; Falchikov & Goldfinch, 2000; Leenknecht, & Prins. 2018; Liu & Carless, 2006; Orsmond, Merry, & Reiling, 2000; Sluijsmans, Brand-Gruwel, & van Merriënboer, 2002a).

Different approaches to assessment criteria do not necessarily imply different points of view on whether they should be applied or not to MOOCs, as opposed to traditional learning environments. The way in which MOOCs are implemented develops new ways of student-teacher-course interaction. Several authors (Topping, 2009; Van Hattum-Janssen & Lourenço, 2008) point out the relevance of student implication and participation when designing evaluation criteria for peer assessment activities. Students get more involved in the task, and a two way path of understanding the activity is created. However, this proposal cannot be applied to MOOCs: (i) the 'open' nature of MOOCs brings together students with very different backgrounds and needs, and, consequently, with very different perspectives; and (ii) another common property of these courses is students' asynchrony when following the course. Student implication and participation in the design of criteria becomes complicated due to this factor. Strict submission dates can help overcome such issue. Many authors have highlighted the important effects of deadlines on formative actions that require feedback (Black & William, 2009; Epstein et al., 2002; Kulik & Kulik, 1988; McKeachie, Pintrich, Lin, & Smith, 1986; Ng, 2014; Webb, Stock, & McCarthy, 1994). Feedback delays can cause formative evaluations to be useless. Some studies have addressed through experimentation that immediate feedback leads to better learning than a delayed one (Kehrer, Kelly, & Heffernan, 2013). In this sense, MOOCs usually take place in fast paced contexts, and hence, deadlines times are usually tight.

2.2.3. Number of raters

The effect of the number of raters on peer assessment has been analyzed with different results depending on the study. Falchikov and Goldfinch (2000:312) hold that "singletons do not appear to be less reliable than others", however they refer to reliability by analyzing its correlation with instructor grades (validity), instead of analyzing the reliability of the raters. They also suggest that a large number of raters may cause a diffusion of responsibility in reviewing tasks. However, this may be caused due to the consequent higher number of required reviews for each student and, therefore, promote boredom in the reviewing process. The studies of Cho et al. (2006), Kilic and Cakan (2007), Xiao and Lucking (2008), Sung et al. (2010) and Chang, Liang, and Chen (2013) found that reliability increases by increasing the number of raters. The results obtained in the study carried out by Kulkarni, Wei, Le, Chia, Papadopoulos, Cheng et al. (2013) concluded that an increasing number of raters increases accuracy (they use accuracy to express the degree of proximity to the teachers'/experts' mark). To be more specific, the improvements experimented are decreasing as the number of reviewers increases following a logarithmic trend. In the model used by Li, Xiong, Zang & Mindy (2016) for their meta-analysis, the correlation between teachers' and peers' ratings was high for assignments with more than 10 reviewers, medium for assignments with 6–10 reviewers, and low for 5 or less reviewers. However, the results were not statistically significant at the 95% level.

2.2.4. Social factors

According to Topping (2009:24), "social processes can influence and contaminate the reliability and validity of peer assessments".

Social factors such as friendship, aversion, popularity, conflict avoidance and so on are present in peer assessments (Friedman, Cox, & Maher, 2008; Topping, 2009). They particularly show up when peer assessment activities are carried out on face to face methodologies. Therefore, these are not a critical factor in MOOCs due to geographical distance, online anonymization, and even because of asynchrony.

Onset education often chooses to keep the assessed tasks double-blinded (Ng, 2016). This is often the case in MOOCs, where users are only identified by the nickname or just the identification number that the platform assigns to each one. However, factors such as anxiety are present at any educational scenario for both reviewer and reviewee (Topping, 2017). MOOC anonymity and distance environments diminish the assessment subjectivity caused by these social factors. However, many others social factors, such as the inevitable sympathy towards peers, the use of a foreign language, different culture, economic factors, gender, etc. (Kizilcec, Davis, & Cohen, 2017; Kizilcec, Saltarelli, Reich, & Cohen, 2017; Suen, 2014) cannot be avoided nor controlled.

Havnes, Smithe, Dysthe & Ludvigsen (2012) identified another factor that affects peer assessment marginally. They tested it in six different high schools in Norway. Students perceive feedback as more or less useful depending on the manners and the terms used as well as on the classroom's atmosphere. In this way, critical feedback is taken as constructive under the appropriate circumstances and a correct choice of words. Peer evaluation promotes this sort of contexts because students are often acquainted to each other. Furthermore, Hovardas, Tsivitanidou, and Zacharia (2014) hold that peer feedback entails more improvements for learners than expert feedback. Initially, this factor does not affect the reliability or validity of the assessment process as it involves the way the students perceive the feedback in the assessment.

We can conclude that social factors can also play some role in MOOC peer assessment, since "peer assessment is a multifaceted process … affected by a number of psychological and personality traits" (AlFallay, 2004, p. 419).

2.3. Measuring reliability in peer assessments

The core aim and benefit in peer assessment is the learning that students experience during the peer assessment process, both as assessors and assessees. However, summative assessment may be considered as a possibility in some cases. Traditionally, the resulting grades from peer assessment have been considered valid or not by confronting them with teacher's/expert's ratings (Chang, Tseng, & Lou, 2012; Cho et al., 2006; Falchikov and Goldfinch, 2000; Formanek et al., 2017; Kilic & Cakan, 2007; Li et al., 2016; Stefani, 1994; Sung et al., 2010; Tsai, Lin, & Yuan, 2002). This comparison of students' evaluation with the teachers' ratings has been referred to as 'validity', while the term 'reliability' is used to determine the consistency among peer ratings (Jackson, 2014; Luo et al., 2014; Richmond et al., 1992). The results obtained in terms of validity and reliability of peer assessment vary from one study to another.

Cho et al. (2006) point out that both reliability and validity studies always leave aside students' point of view, in favor of the teachers'. Students and teachers perceive reliability and validity differently: "the instructor can take into account the effective reliability of ratings generated by a set of peers, whereas each student is restricted to a consideration of the reliability of individual peer ratings"; hence, students' opinion is based on the criterion that "the greater the spread of grades, the less reliable".

No matter the rater or the group of raters chosen for a specific task, Hayes and Krippendorff (2007) talk about the inherent presence of the human condition: "When relying on human observers, researchers must worry about the quality of the data". Classical test theory is based on the assumption that every grade can be understood as the sum of 'true score' (Lord & Novick, 1968; Novick, 1966), this is, "the expectation of an individual's observed score" (Zimmerman, Williams, Zumbo, & Ross, 2005), plus the error score.

The level of agreement or consistency among the evaluations or judgments carried out by the raters or 'graders' is known as IRR (Lange, 2011; Lavrakas, 2008). Krippendorff (2011) defines reliability as "the extent to which different methods, research results or people arrive to the same interpretations or facts". However, "reliability is only a prerequisite to validity. It cannot guarantee it" (Krippendorff, 2011). Raters' consistency is the most relevant factor when studying and analyzing reliability. Through reliability, we try to figure out if raters are consistent in their judgments or assessments, without taking into account the level of agreement they reach; "The consistency of a marker is more important than whether he or she disagrees with another marker" (Brown, Bull, & Pendlebury, 1997, p. 235).

Hayes and Krippendorff (2007) claim that "choosing an index of reliability is complicated by the number of indexes that have been proposed". For starters, we should reject measuring IRR by means of percentages of agreement (Hallgren, 2012) because it ignores the level of agreement, in favor of a 'correct' or 'incorrect' evaluation. Information loss is therefore severe unless the analysis is limited to dichotomic, or even nominal, variables.

Pearson's Correlation Coefficient (PCC), also known as the "Product Moment Correlation Coefficient" (PMCC) has been used in several studies as an interrater reliability estimator (Ashenafi, 2017; Cho, Schunn, & Wilson, 2006; Jones & Wheadon, 2015)). Particularly, it has been applied to the analysis of quantitative variables in peer assessments. However, this coefficient, besides assuming a state of normality, can only be applied if the raters are only two and if they are in charge of assessing all participants. This measure is, therefore, not applicable in our case. Some studies have chosen to overcome the limitation in the number of raters by using Fleiss' kappa (Raman & Joachims, 2014; Schaer, 2012, pp. 124–135). In this way, they have managed to include more raters, but this measure can, once again, only be either dichotomic or nominal. Cohen's kappa (Cohen, 1960), which is a non-parametric test for qualitative variables, or Scott's pi (Scott, 1955), are some of the other statistical methods that have been used for IRR measuring (Antoine, Villaneau, & Lefeuvre, 2014; Zapf, Castell, Morawietz, & Karch, 2016; Lombard, Snyder-Duch, & Bracken, 2004). The most common methodology found when studying reliability in peer evaluations is the Interclass Correlation Coefficient (ICC), or other derived versions from it (Cho et al., 2006; Formanek et al., 2017; Luo et al., 2014; Shieh, 2016; Xiao & Lucking, 2008; Yoon, Park, Myung, Moon, & Park, 2018). Its basic advantage is that it allows high flexibility on the number of raters per test. However, within our data collection, we have 63 distinct peer assessment activities from our platform, that sum up to 27,745 submitted tasks, with three or

more raters in each task distributed across different courses. Furthermore, we find differences in the number of raters within each activity due to how the peer assessment is operationalized in the MOOC platform. For all this, ICC requirements do not match the properties of our sample.

Anyhow, Shrout and Fleiss (1979) presented a statistical method similar to ICC which has already been used within the MOOC context by Luo et al. (2014). The variability in the number of raters made the authors limit their ICC study to only those tests that had five raters. We consider that subsetting the data for an ICC statistical analysis based on the number of raters, clearly undermines the robustness and trustworthiness of the reliability analysis we want to conduct.

Krippendorff's alpha statistic (Krippendorff, 1970; 2011; 2018) provides a reliability measure based on the expected and the observed disagreement. This method comes along with a very high data flexibility: it works with two or more raters, and it does not require that every rater has evaluated every test (the statistic can handle missing values). Besides, it is applicable to all sorts of data types, like ordinal, interval or binary variables. Attending to the measurement scale in our case study, the requisites that the statistic must meet are any number of raters and the existence of missing data. Therefore, we decide to use in this article Krippendorff's alpha statistic to analyze peer assessment reliability in MOOCs for the reasons already given: i) we require a statistic that can handle more than two raters, ii) we require flexibility in the number of raters for each subset, iii) we require to handle missing values, and finally iv) we require a statistic able to deal with ratio variables.

3. Methodology

3.1. Context

UNED-COMA was developed under the open platform OpenMOOC (https://github.com/OpenMOOC) and integrated within the framework of OpenupED (https://www.openuped.eu). By the date when this study was conducted, there were 23 courses, from

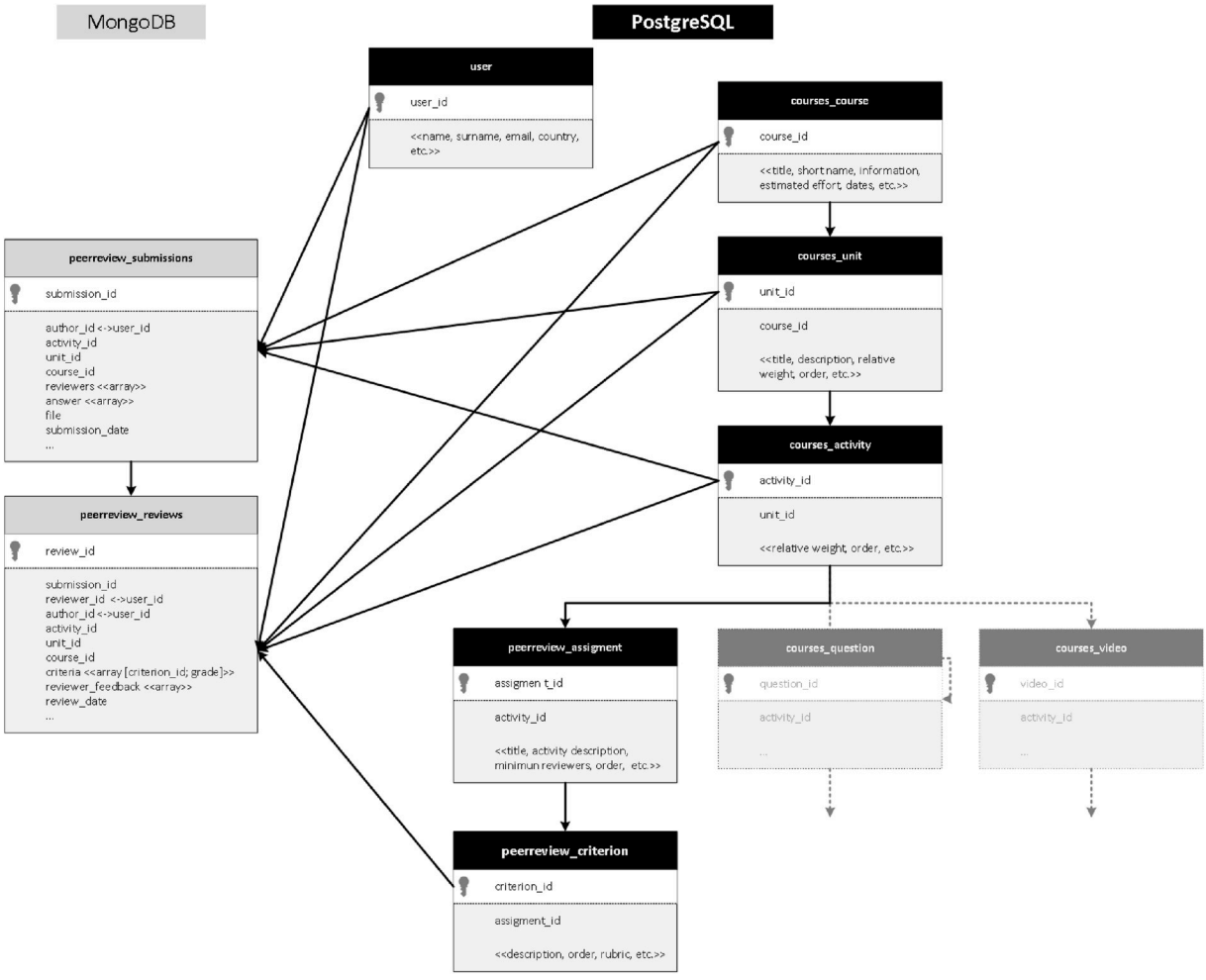

Fig. 1. MongoDB and PostgreSQL joint data schema.

technical topics such as basic analytical chemistry or practice-based electrical/electronics circuits, to second language learning or focused on continuous training (Capdevila & Aranzadi, 2014; Garcia-Loro et al., 2014). The platform also hosts Small Private Online Courses (SPOCs) targeting teachers. The platform has around 140k unique students and 220k enrolments in courses that have triggered more than 25k certification badges.

The structures and the activities designed by the Educational Boards (EBs) —to-do activities, questions, answers and evaluation criteria— can be found in PostgreSQL. Answers and student activities are recorded in MongoDB. Students' data are stored in a different DDBB tables, separated from the rest of the structure. Fig. 1 depicts the structure we have just described. The different Postgres tables are nested through the fields shown in the arrows in Fig. 1, except for the table of users, which is independent. Each activity provided by the platform is nested in the activity table. Fig. 1 exclusively presents peer evaluation activities.

3.2. Peer assessment implementation

Peer evaluation activities on the platform are organized in the following two steps, which are also a requisite in order to consider the peer assessment activity as completed:

1. The student needs to upload the task developed to the platform. Strict deadlines are optional in this step.
2. The student needs to assess a minimum number of tasks from other peers. This number is fixed by the EB, and most of the times is around 3 reviews. However, they have no control on which tasks are assigned to which student since this process is automatically run by the matchmaking system of the platform.

Once the student has completed both steps, the platform marks the task as completed by the student. Nevertheless, before the grading process can be finished, the students' assignment needs to be evaluated by a minimum number of students (fixed by the EB). Even if the student already completed both steps, they will need to wait until other students complete the evaluation of their own assignment.

The assessment of each task implies both a summative and a formative component. They both respond to the criteria previously set by the EB. The assessments provided to students can be classified into two types:

- Quantitative evaluation (summative assessment): The assignment is graded based on whole numbers from 1 to 5 (min and max respectively), according to evaluation criteria or rubrics, provided by the EB.
- Qualitative evaluation (formative assessment): the author of the task receives feedback written by the reviewer. It is implemented in an optional way on the platform.

The full process for a peer assessment activity is shown in Fig. 2. Fig. 2A shows the creation of a peer assessment task with the different settings that EBs may use: (A1) here the EB's may add additional contents for the activity, like a video or documents; (A2) this selection box is used to establish the minimum number of reviewers required; (A3) short description of the activity; and (A4) the definition of the criterion (title and short description) for each of the criteria to be assessed. Fig. 2B shows the student interface to complete a peer assessment task: (B1) provides the short description provided by the EBs in (A3); meanwhile (B2) shows the criteria information provided by the EBs in (A4); (B3) and (B4) are the options provided by the platform to submit the answer, either as plain text (B3) or attaching a document (B4). Fig. 2 (C) shows the interface that a student sees when acting as a reviewer in a peer assessment. (C1) provides the description provided by the EB in the section (A3); (C2) is the answer provided by the student (plain text, no file attached); (C3) and (C3′) are the criteria to be graded by the student, which was set up by the EB in (A4); (C4) and (C4′) are the scale

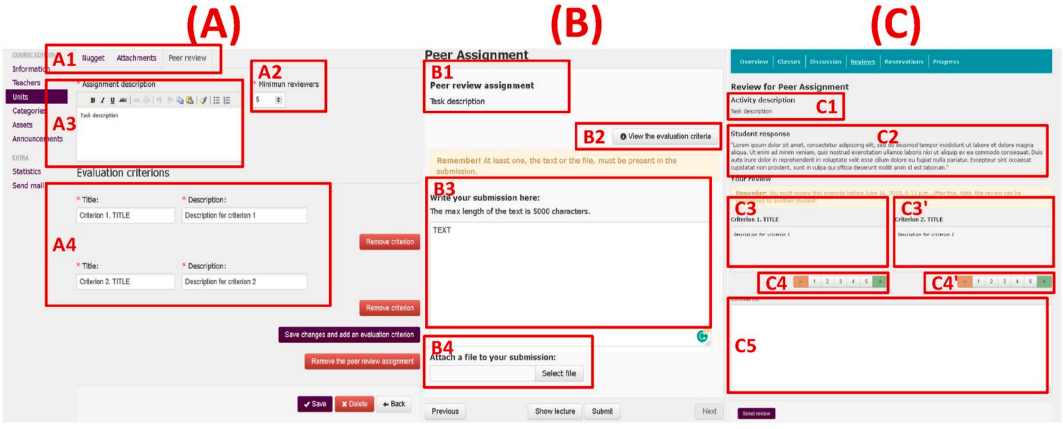

Fig. 2. Implementation of a peer activity and different stages of a peer assessment task in the platform. From left to right: (A) teacher's design of the activity; (B) student's answer; (C) peer's review.

(1–5) to grade each criterion (in this example we have two criteria); (C5) is intended for the reviewer's written feedback.

Analyzing Fig. 2 you might have deducted that all criteria have the same weigh in the grade of the task: the grade of each individual rater will be the unweighted average of the scores of each criteria proposed for the peer assessment task. The final grade will be the average of all peer raters' grades. The assessment of a certain peer activity is based, or should be based, on criteria established by the EB. The summative evaluation on the platform is mandatory, in other words, no review can be submitted unless it includes the grade. However, formative feedback is optional and raters can submit the review to the system without introducing one. Additionally, the feedback box (C5) is not particular for each criterion, but it is a global feedback, yet some EBs may choose to promote it given the bidirectional benefits we have talked about in the previous section. Since the platform does not include a detailed control of this aspect of the evaluation, we do not focus on it.

3.3. Krippendorff's alpha

The study described in this paper has extracted the data from all the summative evaluations from UNED-COMA platform. As we analyzed in Section 3.2., Krippendorff's alpha effectively works with the data we have collected, since the number of raters is independent, it works with different data types and it can handle missing values. It also takes into account the coincidences derived from randomized answers. According to Krippendorff (2011, 2004), Krippendorff's alpha is formulated as follows:

$$\alpha = 1 - \frac{D_o}{D_e} = 1 - (n-1)\frac{\sum_c \sum_{k>c} o_{ck}\delta_{ck}^2}{\sum_c n_c \sum_{k>c} n_k \delta_{ck}^2}$$

$$\delta_{ck} = \left(\frac{c-k}{c+k}\right)$$

where:

α Krippendorff's alpha
D_o the observed disagreement
D_e the expected disagreement
o_{ck}, n_c, n_k and n frequencies of values in coincidence matrix
δ_{ck}^2difference function
c, k elements in the difference function for the weights (row & columns)

The resulting statistical measure is a coefficient ranged from 0 to 1, where 0 is perfect disagreement and 1 is perfect agreement. The coincidence matrix is constructed from the ratings given by the reviewers. It is a square and symmetrical matrix which columns and rows are tagged with the grades assigned by raters. The coincidence matrix assigns a tabulation of the number of coincidences between values, "it visualizes the reliability of the data it tabulates" Krippendorff (2018:408). The difference function is defined according to the metric of the data in order to "weight the observed and expected coincidences of c-k pairs of values", Krippendorff (2004:232).

3.4. Data collection

Our data include a total number of 89 peer evaluation activities, of which 63 have been considered valid for this study. The main rationale behind this selection has been the validity of the activity, given that, in many cases, EBs have rejected or redesigned some activities, which have consequently become obsolete. We have determined validity based on those contents that were ratified by EBs.

Table 1
Extracted and post-processed information.

author_id	activity_id	reviewer_ids	N. reviewers	Reviewers assessment
84613	1170	[80610, 89931, 52632]	3	[4.0, 4.0, 5.0]
53370	1170	[89931, 52632, 49306]	3	[2.75, 3.75, 4.75]
7534	1171	[40684, 89931, 67346]	3	[3.75, 4.25, 4.25]
44385	1237	[89399, 60279, 90426]	3	[4.0, 4.0, 5.0]
875428	1168	[66530, 41933, 60878]	3	[3.0, 4.0, 4.0]
87985	1237	[89277, 65993, 60593]	3	[3.0, 5.0, 5.0]
99445	1168	[72232, 72332, 89931]	3	[3.0, 3.5, 3.5]
78769	1237	[89399, 60279, 58740]	3	[4.0, 5.0, 5.0]
65257	1237	[89399, 38090, 26724]	3	[3.0, 4.0, 5.0]
33956	1171	[89931, 49306, 52632]	3	[1.0, 2.5, 3.25]
89452	1172	[80610, 49306, 67346]	3	[3.25, 4.0, 4.75]
103407	1174	[80610, 49306, 67346]	3	[3.25, 4.0, 5.0]
28732	1170	[49306, 67346, 54142]	3	[3.0, 3.25, 3.5]
73482	1171	[67346, 52632, 64663]	3	[2.75, 3.5, 5.0]
29452	1174	[89931, 67346, 40684]	3	[3.0, 4.75, 4.75]

Another reason has been based on the size of the sample of tasks submitted; if it was too small the peer activity has not been considered.

Table 1 shows one example of the, already, pre-processed raw information extracted from our DDBB, according to the methodology we have specified above, from which we have post-processed and analyzed the data.

4. Results

4.1. Distribution of the peer review assessments

We have collected globally a total number of 37,506 submitted tasks that belonged to peer evaluation activities. 9761 tasks were discarded due to they belonged to the not validated peer activities aforementioned in section 3.4 or because they were not reviewed by at least three raters. We have thus included 27,745 valid tasks.

Regarding to the final grades, most of them span from 3.5 to 4.5 (55.81%). The most common final grade (mode) has been 4 (6.33%). 5.32% peer tasks obtained the highest grade (5); while the lowest grade (1) was only given to 43 tasks (0.155%). The average grade has been 3.859 out of 5; meanwhile, the median is 3.917. Therefore, given that the mean is lower than the median, and that they are both lower than the mode, the distribution of grades is slightly biased to the right as Fig. 3 shows. Regarding to the peer reviews, we have a sample of 93,334 reviews, most of them were scored between 4 and 5 (56.74%), the mode has been 5 (24.46%), while only 2.33% of the reviews were marked with the minimum grade.

Each validated task of this study involves, at least, three reviews. Taking into account that each review task has several evaluation criteria, we had to consider almost 334,000 assessed criteria to come up with the summative evaluations of each revision. All this information is contained in Fig. 4 for each activity where it represents the number of submitted tasks on the *x*-axis, the average number of raters per activity on the *y*-axis, and the number of evaluation criteria for each activity.

4.2. Results of reliability based on Krippendorff's alpha

Krippendorff's alpha considers observers interchangeable with the number of pairs used. Consequently, the results are based on all the data provided by all observers, and it is not affected by their number (Hayes & Krippendorff, 2007).

The value of Krippendorff's alpha (see equation) must be found between '1', when the observed disagreement (D_o) is null, and '0' when the observed disagreement (D_o) matches the expected disagreement (D_e). According to Krippendorff (2011), as a general rule of thumb, we assume that the relevant values, or the statistically significant values for Krippendorff's alpha, should be over 0.80. However, some positive conclusions or trends can be drawn from 0.67 onwards. To this respect, Hallgren (2012) points out that these values can vary depending on research methodology and goals.

Table 2 presents the Krippendorff's alpha results for the considered peer activities based on the aforementioned equation and the macro provided by Hayes and Krippendorff (2007). The box-plot representation for Krippendorff's alpha of the 63 analyzed activities in the different courses is shown in Fig. 5. The mean for all 63 activities is of 0.2327; while the first and the third quartiles are on 0.1573 and 0.3092 respectively. In other words, most of the activities have a very low Krippendorff's alpha.

4.3. Factors influencing reliability

Considering all tasks, the average standard deviation (SD) and the Pearson's Coefficient of Variation (PCV) of the Krippendorff's alpha are 0.12 and 0.5 respectively. The mean of Krippendorff's alpha for all peer review activities is 0.2327 (Fig. 6). By analyzing the peer assessment tasks by course, we can draw some conclusions, e.g., in Fig. 6 the reliability of the peer assessment tasks is grouped by

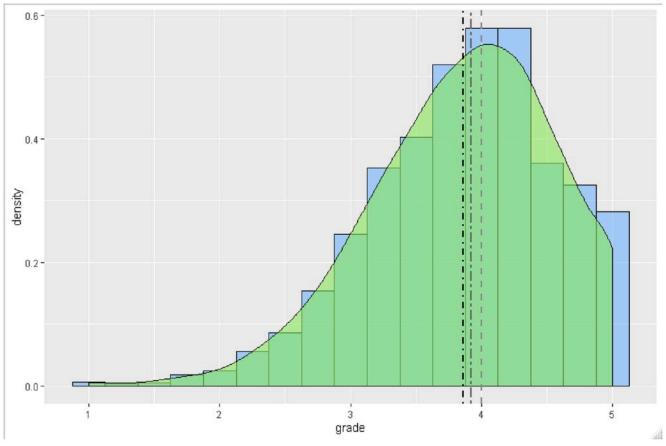

Fig. 3. Distribution of the final grade over the 27,745 tasks validated; dark vertical line indicates the mean of all final grades, grey vertical line denotes the median while light-grey vertical line the mode.

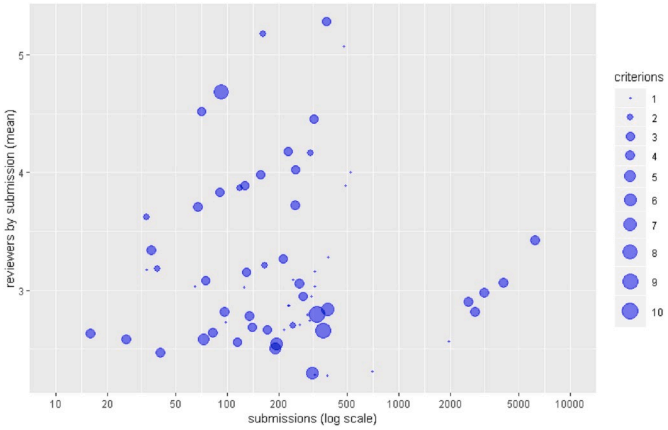

Fig. 4. Scatterplot representing the available data. Each dot represents an activity with the average number of raters on the *y*-axis and the number of submitted tasks (log scale) in *x*-axis. The size of the dot codifies the number of criterions in the task.

course and arranged by its sequence order within the course. The dispersion of the reliability by course is, in general terms, much better than the global one. Considering those courses containing at least two peer assessments tasks, averaging the reliability of the tasks by course provides a better result in terms of dispersion: Only one course (C23 in Fig. 6) presents worse dispersion values (SD ~ 0.16, PCV ~ 0.7), and two courses (C24 and C25) present similar dispersion values (C24: SD ~ 0.11, PCV ~ 0.5; C25: SD ~ 0.14, PCV ~ 0.48). Most courses (8 courses) present dispersion values for the reliability around half of the global one, both for the SD and PCV. It should be noted the case of course C20, which, with five peer assessment tasks, presents the lowest dispersion values (0.01 and 0.07 for SD and PCV respectively).

Table 3 presents the percental distribution of disagreement in the subset of raters of each task assessed. To generate this distribution, we compute the maximal distance between the grades given by each group of raters in each task and classify them in their disagreement range. Table 3 shows the dispersion between maximal and minimal grades in the subset of raters for each peer evaluation task on the platform, which is calculated without considering the number of raters in each subset of raters. Obviously the more raters, the higher the chance of disagreeing evaluations as the probability of getting larger maximal distances increases.

We believe that additional explanation regarding the Krippendorff's alpha reliability peer assessment will be helpful to avoid misinterpreting some data points. For the results in Table 3, grades vary from 1 (lowest) to 5 (highest) in a 1 by 1 scale of whole numbers. For those tests that contain only one evaluation criterion, which is the case in over 20 activities, the lowest level of disagreement would be a distance of 1. Therefore, this is the reason why we consider the maximal distance of 1 as acceptable for an agreement percentage. Fig. 7a and Fig. 7b scatterplots show the relationship between Krippendorff's alpha and the percentage of tests in which the evaluation of the raters has shown a strong agreement (distance between grades below or equal to 1) and the percentage of tests in which the evaluation provided by the subset of raters has shown a strong disagreement (distance between grades bigger or equal to 3). The PCC coefficient for the Krippendorff's alpha and the percentage of peer assessment tasks with a strong agreement between the raters of each subset is low, 0.311 (p-value = 0.013). In the case of the correlation between the disagreement and the reliability, the correlation is stronger, −0.395 (p-value = 0.001).

Fig. 4c and d show the relationship between the number of criteria of the activity and the average number of raters, respectively, with the Krippendorff's alpha. In both cases, the PCC coefficient is not significant (p-values = 0.7901 and 0.2845 respectively), thus we accept the hypothesis that true correlation is equal to 0. Furthermore, the correlation is low in both cases (0.034 and −0.137 respectively).

5. Discussion

Attending to the values obtained for Krippendorff's alpha statistic in the 63 assessed activities, and considering the recommendations offered in Krippendorff and Bock (2009: 354) and Krippendorff (2004: 241) to "rely on Krippendorff's alpha above 0.80", we find that in our peer review activity dataset there are no significant values in terms of agreement between reviewers. Therefore, none of the peer evaluation activities carried out in the different courses on the platform can be considered reliable when talking about the evaluations performed by the students.

The maximum value of Krippendorff's alpha was obtained in activity PAT#007 (0.5718). However, not even this value is enough to be used for "drawing tentative conclusions", because the value remains under the threshold value (0.667) (Krippendorff & Bock, 2009:354; Krippendorff, 2004:241).

Under the assumption that reliability is, although not sufficient, a necessary condition to guarantee the validity of the established evaluation methodology, with the obtained results in hand we can conclude that grades obtained by means of peer assessment in this study are not trustworthy. Jonsson and Svingby (2007) highlight that reliability is not always required for validity, because there are certain scenarios where "the basis of the assessment can be easily changed" (for example, in-classroom assessments). These scenarios

Table 2
Krippendorff's alpha results.

	Alpha	Units	Observers	Pairs
PAT#001	0.16	91	98	832
PAT#002	0.21	160	178	1787
PAT#003	0.23	518	560	3207
PAT#004	0.23	245	282	1300
PAT#005	0.22	374	399	4332
PAT#006	0.14	318	333	2612
PAT#007	0.57	59	86	213
PAT#008	0.21	305	317	2055
PAT#009	0.40	50	59	196
PAT#010	0.19	249	259	1717
PAT#011	0.16	226	227	1688
PAT#012	0.31	155	157	930
PAT#013	0.17	6206	6615	27299
PAT#014	0.17	3867	4324	13455
PAT#015	0.17	2878	3284	9324
PAT#016	0.15	2138	2699	7123
PAT#017	0.15	2161	2500	6931
PAT#018	0.11	1049	1548	3496
PAT#019	0.23	201	398	679
PAT#020	0.17	103	187	315
PAT#021	0.36	82	159	280
PAT#022	0.04	476	529	2923
PAT#023	0.36	124	131	724
PAT#024	0.24	69	75	577
PAT#025	0.20	115	119	727
PAT#026	0.32	89	95	538
PAT#027	0.21	314	355	976
PAT#028	0.29	219	294	762
PAT#029	0.22	66	70	354
PAT#030	0.24	242	303	806
PAT#031	0.17	163	218	777
PAT#032	0.14	211	240	802
PAT#033	0.17	187	215	591
PAT#034	0.42	148	178	613
PAT#035	0.19	30	35	171
PAT#036	0.09	81	133	378
PAT#037	0.38	92	122	317
PAT#038	0.53	37	58	157
PAT#039	0.13	93	125	279
PAT#040	0.14	62	95	189
PAT#041	0.23	288	347	1010
PAT#042	0.47	30	35	143
PAT#043	0.46	36	38	135
PAT#044	0.34	200	258	674
PAT#045	0.27	213	262	690
PAT#046	0.15	466	532	5324
PAT#047	0.21	293	336	1133
PAT#048	0.24	218	287	946
PAT#049	0.22	350	383	1536
PAT#050	0.23	162	199	492
PAT#051	0.44	139	180	417
PAT#052	0.29	92	139	342
PAT#053	0.31	90	146	279
PAT#054	0.27	118	119	378
PAT#055	0.14	89	153	267
PAT#056	0.32	171	175	804
PAT#057	0.39	102	115	453
PAT#058	0.20	53	51	204
PAT#059	0.22	229	261	792
PAT#060	0.15	78	98	234
PAT#061	0.01	31	37	114
PAT#062	0.05	15	20	54
PAT#063	0.50	29	52	87

are nowhere close to our case study.

Despite we cannot perform a direct comparison between our results and the ones reported in other studies due to the use of different statistics, the differences and conclusions from each separate study suggest that our study presents much lower reliability than the rest of studies that performed similar analysis in other contexts and using different metrics. The results obtained in classical learning

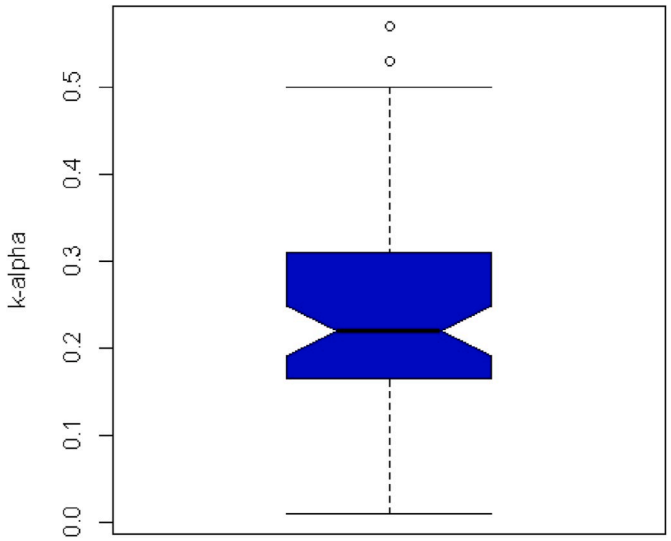

Fig. 5. Boxplot of the Krippendorff's alpha values of all peer review activities in all courses.

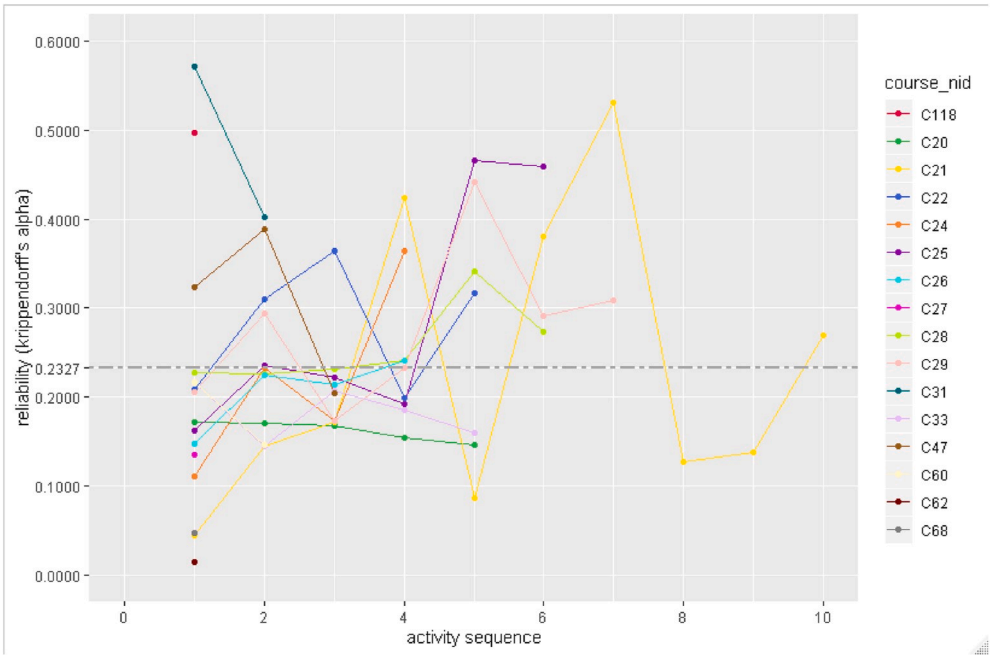

Fig. 6. Evolution of the Krippendorff alpha value among the different courses and through the different tasks.

scenarios tend to provide a solid reliability. For example the ones provided by Yoon et al. (2018) —with ICCs values obtained from 0.390 to 0.863; being the overall average 0.659, from 141 students, who were divided into 18 groups in 11 team-based learning classes— or the ones obtained by Salehi and Masoule (2017) —Cronbach's alpha values from 0.709 to 0.900 for peer assessing oral production in three groups. Moreover, in other studies using MOOCs as learning scenario; for example the ICC averages measures obtained by Formanek et al. (2017) and Luo et al. (2014) −0.591 for the ICC and 0.579 respectively.

Anyway, and according to our results, the fact that we did not find any peer assessment activities with Krippendorff's alpha values even close to the recommended threshold values, drives us to think that the reason might be a systematic problem and not particularly associated with specific peer assessment activities in our case study. However, analyzing Fig. 6 we can see how the mean of the Krippendorff's alpha between courses is quite different. We do not find substantial differences after grouping courses by topic and, according to the data obtained, it does not look as if there is a significant relationship between the topic of the course and the reliability achieved. Conversely, even if they are focused on similar topics, such as C21, C31 and C32, all of them focused on TICs and its

Table 3
Maximum distance among the subset of raters (percentages).

	n	reviews (mean)	0	(0, 0.5]	(0.5, 1]	(1, 2]	(2, 3]	(3, 4]
PAT#001	91	4,70	0,00%	4,40%	9,89%	36,26%	28,57%	20,88%
PAT#002	160	5,21	0,00%	3,25%	9,74%	37,66%	32,47%	16,88%
PAT#003	518	4,01	10,16%	0,00%	32,42%	32,42%	14,84%	10,16%
PAT#004	245	3,76	3,35%	5,86%	22,59%	36,40%	23,01%	8,79%
PAT#005	374	5,31	0,27°%	0,82%	10,87%	50,27%	29,62%	8,15%
PAT#006	318	4,48	0,32%	2,88%	20,19%	51,60%	21,47%	3,53%
PAT#007	59	3,20	41,51%	0,00%	20,75%	20,75%	13,21%	3,77%
PAT#008	305	4,17	2,01%	7,36%	26,42%	46,15%	14,72%	3,34%
PAT#009	50	3,30	52,27%	0,00%	18,18%	15,91%	9,09%	4,55%
PAT#010	249	4,04	0,82%	2,47%	28,81%	48,97%	16,46%	2,47%
PAT#011	226	4,18	1,36%	3,18%	27,73%	49,09%	14,55%	4,09%
PAT#012	155	3,99	1,34%	3,36%	22,82%	46,31%	22,82%	3,36%
PAT#013	6206	3,43	0,66%	9,65%	20,85%	45,82%	19,11%	3,90%
PAT#014	3867	3,12	1,68%	13,21%	26,91%	42,19%	13,52%	2,49%
PAT#015	2878	3,06	1,78%	14,38%	26,78%	42,76%	12,36%	1,95%
PAT#016	2138	3,06	2,44%	13,13%	28,80%	42,68%	10,79%	2,16%
PAT#017	2161	3,06	2,97%	14,52%	30,58%	39,44%	10,39%	2,09%
PAT#018	1049	3,05	21,00%	0,00%	41,32%	24,07%	8,53%	5,08%
PAT#019	201	3,07	16,41%	1,54%	39,49%	29,23%	9,74%	3,59%
PAT#020	103	3,02	10,31%	0,00%	46,39%	27,84%	9,28%	6,19%
PAT#021	82	3,10	19,74%	0,00%	39,47%	17,11%	13,16%	10,53%
PAT#022	476	3,93	13,83%	0,00%	20,00%	24,47%	14,47%	27,23%
PAT#023	124	3,92	3,39%	0,85%	28,81%	39,83%	24,58%	2,54%
PAT#024	69	4,55	1,59%	0,00%	15,87%	47,62%	28,57%	6,35%
PAT#025	115	3,91	2,75%	3,67%	23,85%	40,37%	22,02%	7,34%
PAT#026	89	3,84	2,41%	3,61%	13,25%	40,96%	33,73%	6,02%
PAT#027	314	3,03	1,95%	9,42%	25,97%	44,81%	12,34%	5,52%
PAT#028	219	3,08	3,29%	15,02%	29,11%	39,44%	7,04%	6,10%
PAT#029	66	3,74	0,00%	3,33%	23,33%	38,33%	23,33%	11,67%
PAT#030	242	3,08	2,12%	15,68%	28,39%	34,32%	11,44%	8,05%
PAT#031	163	3,59	7,01%	10,19%	26,75%	33,12%	13,38%	9,55%
PAT#032	211	3,25	34,63%	0,00%	28,29%	19,02%	9,27%	8,78%
	N	reviews (mean)	0	(0, 0.5]	(0.5, 1]	(1, 2]	(2, 3]	(3, 4]
PAT#033	187	3,05	50,28%	0,00%	22,65%	15,47%	6,63%	4,97%
PAT#034	148	3,34	20,42%	9,86%	33,80%	21,13%	5,63%	9,15%
PAT#035	30	3,80	4,17%	8,33%	25,00%	33,33%	16,67%	12,50%
PAT#036	81	3,40	10,67%	2,67%	17,33%	24,00%	18,67%	26,67%
PAT#037	92	3,13	12,79%	13,95%	23,26%	31,40%	12,79%	5,81%
PAT#038	37	3,41	19,35%	6,45%	22,58%	35,48%	9,68%	6,45%
PAT#039	93	3,00	20,69%	5,75%	22,99%	32,18%	6,90%	11,49%
PAT#040	62	3,02	26,79%	0,00%	14,29%	23,21%	19,64%	16,07%
PAT#041	288	3,15	25,89%	0,00%	37,94%	22,70%	10,64%	2,84%
PAT#042	30	3,57	4,17%	12,50%	16,67%	41,67%	20,83%	4,17%
PAT#043	36	3,25	0,00%	6,67%	40,00%	40,00%	13,33%	0,00%
PAT#044	200	3,12	62,37%	0,00%	21,13%	14,95%	1,03%	0,52%
PAT#045	213	3,08	70,53%	0,00%	17,87%	7,73%	0,97%	2,90%
PAT#046	466	5,16	2,39%	0,00%	18,48%	37,17%	27,39%	14,57%
PAT#047	293	3,28	14,63%	0,00%	49,83%	27,18%	6,62%	1,74%
PAT#048	218	3,31	8,96%	0,00%	54,25%	29,25%	6,60%	0,94%
PAT#049	350	3,41	10,17%	0,00%	42,15%	36,63%	9,30%	1,74%
PAT#050	162	3,01	54,49%	12,18%	8,33%	12,82%	5,13%	7,05%
PAT#051	139	3,00	69,17%	0,00%	16,54%	12,03%	1,50%	0,75%
PAT#052	92	3,11	12,79%	16,28%	17,44%	36,05%	15,12%	2,33%
PAT#053	90	3,03	28,57%	11,90%	20,24%	26,19%	7,14%	5,95%
PAT#054	118	3,07	29,46%	0,00%	29,46%	21,43%	4,46%	15,18%
PAT#055	89	3,00	0,00%	10,84%	30,12%	48,19%	10,84%	0,00%
PAT#056	171	3,56	1,21%	7,27%	35,15%	39,39%	16,97%	0,00%
PAT#057	102	3,46	6,25%	4,17%	34,38%	30,21%	21,88%	3,13%
PAT#058	53	3,28	0,00%	4,26%	27,66%	44,68%	14,89%	8,51%
PAT#059	229	3,15	1,35%	11,21%	27,35%	40,36%	16,59%	3,14%
PAT#060	78	3,00	1,39%	8,33%	25,00%	52,78%	12,50%	0,00%
PAT#061	31	3,23	7,69%	0,00%	0,00%	38,46%	23,08%	30,77%
PAT#062	15	3,20	38,46%	0,00%	30,77%	7,69%	23,08%	0,00%
PAT#063	29	3,00	4,35%	13,04%	34,78%	43,48%	4,35%	0,00%

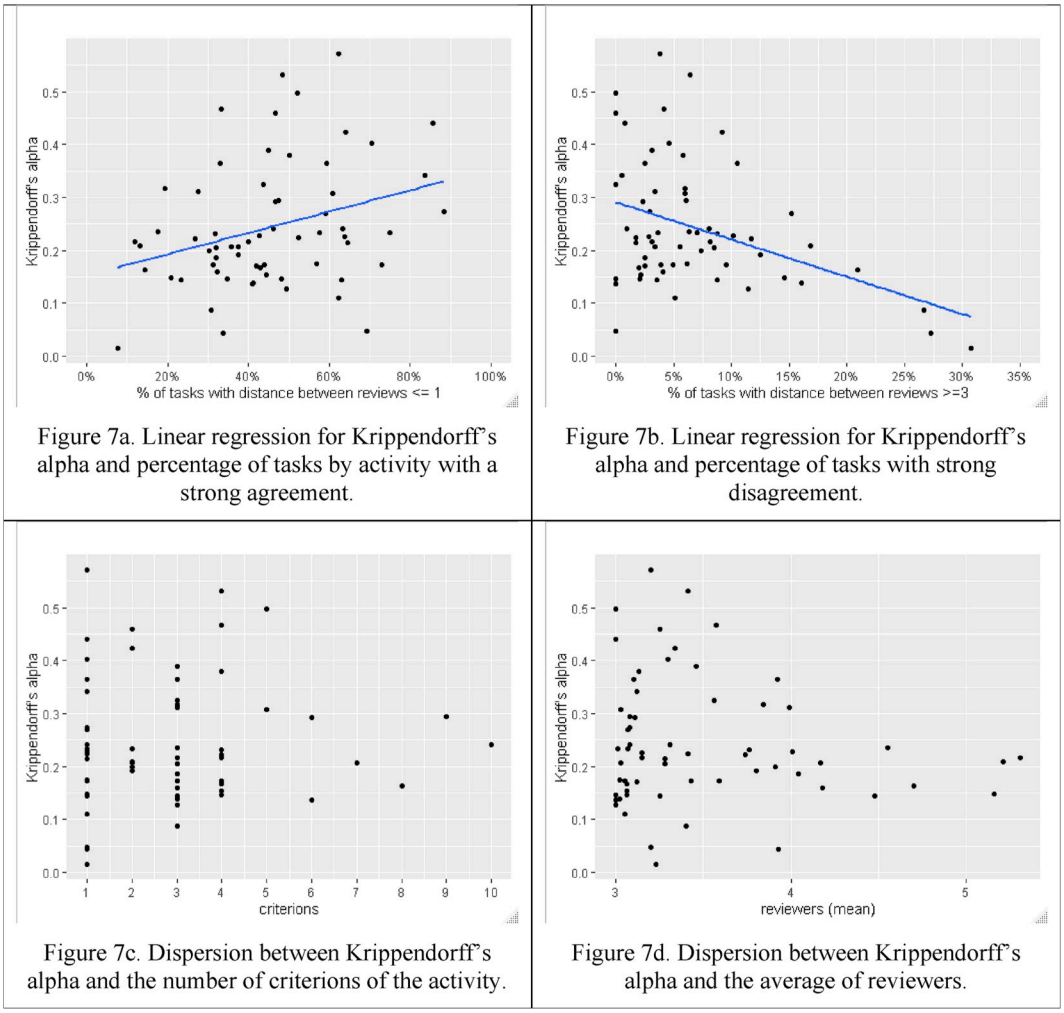

Fig. 7. Scatterplots showing the reliability dispersion based on different factors.

applications, which have completely different results (C21: mean ~ 0.23, SD ~ 0.16, PCV ~ 0.7; C31: mean ~ 0.049, SD ~ 0.12, PCV ~ 0.24; C32: mean ~ 0.027, SD ~ 0.07, PCV ~ 0.26). Another example is C20 and C24, both dedicated to the study of foreign languages, which have relatively similar Krippendorff's alpha value (C24–0.22; C20–0.16), but with dispersion rates quite far from each other (C24: SD ~ 0.11, PCV ~ 0.5; C20: SD ~ 0.01, PCV ~ 0.07). Therefore, in our case study we do not find the topic of the course as a relevant factor affecting reliability, in accordance with the conclusions obtained by Falchikov and Goldfinch (2000).

It is noteworthy the high grades obtained in the peer assessment activities within the platform. One potential explanation regarding this aspect may be related to the involved social factors. While in MOOCs certain social aspects, described in section 2.2, are avoided due to the physical distance and anonymity, some others might still be playing a role, such as the "perception of criticism as socially uncomfortable" (Topping, 2009). Students may be more generous when grading a fellow peer, if we compare grades with instructors' ones (Marks & Jackson, 2013). Hanrahan and Isaacs (2001) pinpoint that students experience empathy with lecturers/tutors because of the large numbers of assignments, however they do not feel the same way towards their peers. In this direction, the results obtained by Formanek et al. (2017) do not show a global trend: "Peer graders tend to underestimate the top-scoring submissions while overestimating the lowest scoring ones". In the meta-analysis conducted by Falchikov and Goldfinch (2000), from 22 studies (not considering atypical ones), 11 studies resulted in over-grading while 7 in under-grading, turning out a weighted mean very slightly under-grading (effect size −0.02).

Training and practicing peer assessment tasks are highlighted as requirements for students before an actual implementation in a real educational scenario (Topping, 2009). However, this training is sometimes focused on how to conduct the grading side following the recommendations of the EB, instead of on the educational component, reliability and/or validity (Kulkarni et al., 2013). In any case, Sluijsmans et al. (2002a) indicate that training promotes a more critical attitude when assessing, but that long training periods are required in order to provide tangible improvements (Sluijsmans et al., 2002b). Formanek et al. (2017) found that the performing a previous training stage in how to assess, helped to improve reliability: an average ICC of 0.591 for graders without previous training

against an average ICC of 0.682 for those trained graders. If we look at the reliability of our students as they progress in each course, we hypothesize that it should improve as they are getting more experienced in conducting peer assessment. When comparing in each course the average reliability of the first half of peer assessment tasks with the average reliability of the final half of peer assessment tasks (e.g., course 25 comprises six tasks: we have compared the average reliability of tasks 1, 2 and 3 with the average reliability of tasks 4, 5 and 6; while course 29, which comprises seven tasks: we have compared the average reliability of tasks 1, 2 and 3 with the average reliability of tasks 5, 6 and 7), the next conclusions, which are in concordance with the aforementioned studies, are envisaged: Courses with more than six tasks present an improvement in the reliability. An average improvement of 54.63% when comparing the reliability of initial tasks with final tasks.

For those courses with four or five peer assessment tasks, the results present a clear difference between the reliability of the first and final halves. Perhaps new research in this direction can experiment on the impact of having an initial peer-review training as a MOOC activity in the reliability of the rest of peer-review assignments. If we recall the technical implementation of the evaluation model based on Krippendorff's alpha values, one of the underlying assumptions was the idea of equity among peer raters. As aforementioned, whereas traditional learning contexts can assume a high similarity degree in the background of their learners, the 'Open' nature of MOOCs highly increments the diversity in learners' profiles, hence potentially breaking the equity among learners' condition. In MOOCs we find that learners have multiple backgrounds in content knowledge (especially those regarding STEM), diverse sets of skills related to writing, text comprehension, synthesis and very different intentions when enrolling in a MOOC (Alario-Hoyos et al., 2014; Watson et al., 2017).

Two factors traditionally analyzed in the reliability have been the number of criterions and the number of reviewers. Fig. 4c and d show the null relationship between the number of criteria of the activity and the average number of raters, respectively, with the Krippendorff's alpha. In both cases, the Pearson's product-moment correlation was not significant, thus in our case study we do not find a relationship between these factors and reliability.

Regarding to the number of criteria or categories to be assessed by peer raters, and in contrast to what Sadler and Good (2006) and Meletiadou and Tsagari (2014) found, or the conclusions obtained by Falchikov and Goldfinch (2000), we do not find any trend in this sense. In our scenario, we found an absence of a significant correlation between the number of criteria and the reliability obtained (Fig. 7c). The number of criteria for each task does not imply any correlation with the Krippendorff's alpha coefficient. In our analysis, the value of Krippendorff's alpha ranges from 0.225 to 0.275 (Fig. 4c). The highest average, 0.275 is obtained with tasks requiring two criterions to be assessed, followed by 0.267 for five or more criterions. On the opposite side, the lowest average value is obtained with three criterions, 0.225.

In the case of the effect of the number of peer raters in the reliability of the assessment process, we do not find any correlation neither (Fig. 7d). In our case study, we have not found any trend as the ones described in the literature review.

In Fig. 7a and b scatterplots with the relationship between Krippendorff's alpha and the percentage of tests in which the evaluation of the raters has shown a strong agreement —distance between grades below or equal to 1— and the percentage of tests in which the evaluation provided by the subset of raters has shown a strong disagreement—distance between grades bigger or equal to 3. Both scatterplots show a correlation between the percentage of agreement and reliability. However, we can see how a strong agreement or the absence of disagreement does not necessarily imply high reliability. The observable dispersion confirms that agreement among raters is mainly irrelevant from the reliability as Krippendorff (2011) predicts.

6. Conclusions and future lines of research

In the particular scenario of UNED-COMA that we have analyzed, we find that the reliability of peer evaluation activities in MOOCs is untrustworthy. Therefore, under the assumption that reliability is a necessary condition to guarantee the validity of the evaluation, peer rating might not be a very trustworthy assessment method in MOOCs, especially if implemented as a summative assessment that counts towards the certification grade. However, our analyses do not take into account the learning benefits of these kind of activities, which have been presented in our introduction. Peer-assessments have been extensively analyzed in the educational literature, finding that students engage more easily in the learning process, they develop critical thinking, etc. Therefore, beyond their reliability and validity as an evaluation method, peer assessments can still provide multiple benefits for students such as a more complex cognitive learning process or personalized feedback; for example, strategies as the one described in (Staubitz, Petrick, Bauer, Renz, & Meinel, 2016) can be applied in order to motivate reviewers to enhance their feedbacks. However, for students to rigorously and fully engage in a learning activity, they often need an incentive towards the final grade. Under this case scenario, one potential pedagogical approach is to mitigate this effect by assigning a relatively low weight to these evaluations in final grades, while maintaining the rest of side transversal advantages. Based on the results obtained, we perceive the need to adapt peer assessment activities, which are traditionally carried out in (relatively) homogeneous and "quasi-controlled" environments, to massive and highly heterogeneous environments.

Future work might lead us to explore if the results of this case study replicate in the peer-assessment systems of other MOOC environments, a comparison of the Krippendorff's alpha statistic with others inter-reliability statistics, experimentation around the effect on reliability of conducting peer-review training before the actual peer-review activities, to analyze the existence and significance of any correlation between the weighting of peer assessments and the reliabilities, or a more in-depth analysis of which qualitative factors moderate the disagreement between raters, such as type of course, background of raters or if it might be more specific to the implementation of the peer evaluation activity.

Declaration of competing interest

The authors declare no conflict of interest in this article.

CRediT authorship contribution statement

Felix Garcia-Loro: Methodology, Formal analysis, Data curation, Investigation, Visualization, Writing - original draft. **Sergio Martin:** Supervision, Conceptualization, Writing - review & editing. **José A. Ruipérez-Valiente:** Writing - original draft, Writing - review & editing, Supervision, Methodology. **Elio Sancristobal:** Writing - review & editing, Resources. **Manuel Castro:** Conceptualization, Funding acquisition, Project administration.

Acknowledgement

This work has been co-funded by the Madrid Regional Government, through the project e-Madrid-CM (S2018/TCS-4307). The e-Madrid-CM project is also co-financed by the Structural Funds (FSE and FEDER). Authors also acknowledge the support of the e-LIVES. e-Learning InnoVative Engineering Solutions- Erasmus + Capacity Building in Higher Education 2017 - 585938- EPP-12017-1-FR-EPPKA2-CBHE-J, IoE-EQ. Internet of Energy - Education and Qualification, Erasmus + - Cooperation for Innovation and the Exchange of Good Practices nº 2017-1-IT01-KA202-006251 and I4EU - Key Competences for an European Model of Industry 4.0, Erasmus + Strategic Partnership nº 2019-1-FR01-KA202-06296. As well as to the projects 2020-IEQ15, 2020-IEQ14 and 2020-IEQ13 from the Escuela Superior de Ingenieros Industriales of UNED. Ruipérez-Valiente acknowledges support from the Spanish Ministry of Economy and Competitiveness through the Juan de la Cierva Formación program (FJCI-2017-34926)

References

Alario-Hoyos, C., Pérez-Sanagustín, M., Delgado-Kloos, C., Parada, H. A., & Muñoz-Organero, M. (2014). Delving into participants' profiles and use of social tools in MOOCs. *IEEE Transactions on Learning Technologies, 7*(3), 260–266.

Alcarria, R., Bordel, B., Andres, D. M.d., & Robles, T. (2018). Enhanced peer assessment in MOOC evaluation through assignment and review analysis. *International Journal of Emerging Technologies in Learning (iJET), 13*(1), 206–219.

Aleven, V., Sewall, J., Popescu, O., Xhakaj, F., Chand, D., Baker, R., et al. (2015). The beginning of a beautiful friendship? *Intelligent Tutoring Systems and MOOCs*, 525–528.

AlFallay, I. (2004). The role of some selected psychological and personality traits of the rater in the accuracy of self- and peer-assessment. *System, 32*(3), 407–425.

Ambekar, D., & Phatak, D. B. (2014). *Evaluation of essays using incremental training for Maximizing Human-Machine agreement*. Doctoral dissertation. Bombay: Indian Institute of Technology.

Antoine, J., Villaneau, J., & Lefeuvre, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation. In *Proceedings of EACL 2014*.

Ashenafi, M. (2017). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward. *Assessment & Evaluation in Higher Education, 42*(2), 226–251. https://doi.org/10.1080/02602938.2015.1100711.

Beldarrain, Y. (2006). Distance education trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education, 27*(2), 139–153.

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349–377.

Benlloch-Dualde, J. V., & Blanc-Clavero, S. (2007). Adapting teaching and assessment strategies to enhance competence-based learning in the framework of the European convergence process. In *Proceedings - frontiers in education conference* (p. S3B6). FIE.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability(formerly. Journal of Personnel Evaluation in Education, 21*(1), 5.

Boekaerts, M., & Corno, L. (2005). Self-regulation in the classroom: A perspective on assessment and intervention. *Applied Psychology, 54*(2), 199–231.

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

Brown, G. A., Bull, J., Pendlebury, M., Bull, J., & Pendlebury, M. (1997). *Assessing student learning in higher education*. London: Routledge.

Capdevila, R., & Aranzadi, P. (2014). Los cursos online masivos y abiertos: ¿Oportunidad o amenaza para las universidades iberoamericanas?= massive open online courses: Opportunity or threat for Iberoamerican universities? *RIED. Revista Iberoamericana de Educación a Distancia, 17*, 69–82. n.1), 2014.

Cartney, P. (2010). Exploring the use of peer assessment as a vehicle for closing the gap between feedback given and feedback used. *Assessment & Evaluation in Higher Education, 35*(5), 551–564.

Chang, C., Liang, C., & Chen, Y. (2013). Is learner self-assessment reliable and valid in a Web-based portfolio environment for high school students? *Computers & Education, 60*(1), 325–334.

Chang, C., Tseng, K., & Lou, S. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education, 58*(1), 303–320.

Chen, C.-. (2008). Intelligent web-based learning system with personalized learning path guidance. *Computers and Education, 51*(2), 787–814.

Cho, K., Schunn, C. D., & Wilson, R. W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891–901.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*, 37–46.

Cornford, I. R. (2002). Learning-to-learn strategies as a basis for effective lifelong learning. *International Journal of Lifelong Education, 21*(4), 357–368.

De Grez, L., Valcke, M., & Roozen, I. (2012a). How effective are self- and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129–142.

De Grez, L., Valcke, M., & Roozen, I. (2012b). How effective are self-and peer assessment of oral presentation skills compared with teachers' assessments? *Active Learning in Higher Education, 13*(2), 129–142.

De Miguel Díaz, M., Alfaro Rocher, I. J., Apodaca Urquijo, P., Arias Blanco, J. M., García Jiménez, E., & Lobato Fraile, C. (2005). *Modalidades de enseñanza centradas en el desarrollo de competencias: Orientaciones para promover el cambio metodológico en el espacio europeo de educación superior*. Servicio de Publicaciones. Universidad de Oviedo.

Delgado García, A. M., Borge Bravo, R., García Albero, J., Oliver Cuello, R., & Salomón Sancho, L. (2005). Competencias y diseño de la evaluación continua y final en el Espacio Europeo de Educación Superior. *Retirado em Março, 8*, 2012.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: A review. *Studies in Higher Education, 24*(3), 331–350.

Dochy, F. J. R. C., & McDowell, L. (1997). Assessment as a tool for learning. *Studies In Educational Evaluation, 23*(4), 279–298.

Duran, D. (2017). Learning-by-teaching. Evidence and implications as a pedagogical mechanism. *Innovations in Education & Teaching International, 54*(5), 476–484.

Earle, S. (2014). Formative and summative assessment of science in English primary schools: Evidence from the primary science quality mark. *Research in Science & Technological Education, 32*(2), 216–228.
Epstein, M. L., Lazarus, A. D., Calvano, T. B., Matthews, K. A., Hendel, R. A., Epstein, B. B., et al. (2002). Immediate feedback assessment technique promotes learning and corrects inaccurate first responses. *Psychological Record, 52*(2), 187–201.
Falchikov, N. (2013). *Improving assessment through student involvement: Practical solutions for aiding learning in higher and further education*. London: Taylor and Francis.
Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.
Formanek, M., Wenger, M. C., Buxner, S. R., Impey, C. D., & Sonam, T. (2017). Insights about large-scale online peer assessment from an analysis of an astronomy MOOC. *Computers & Education, 113*, 243–262.
Friedman, B. A., Cox, P. L., & Maher, L. E. (2008). An expectancy theory motivation approach to peer assessment. *Journal of Management Education, 32*(5), 580–612.
Garcia-Loro, F., Diaz, G., Tawfik, M., Martin, S., Sancristobal, E., & Castro, M. (2014). A practice-based MOOC for learning electronics. In *2014 IEEE global engineering education conference (EDUCON)* (pp. 969–974).
Garcia-Loro, F., San Cristobal, E., Diaz, G., Macho, A., Baizan, P., Blazquez, M., et al. (2018). Pilar: A federation of VISIR remote laboratory systems for educational open activities. In *2018 IEEE international conference on teaching, assessment, and learning for engineering (TALE)* (pp. 134–141).
Garcia-Loro, F., Sancristobal, E., Gil, R., Diaz, G., Castro, M., Albert-Gómez, M., et al. (2016). In *Electronics remote lab integration into a MOOC-achieving practical competences into MOOCs. EADTU 2016, the online, open and flexible higher education conference* (pp. 367–379).
Gauci, S. A., Dantas, A. M., Williams, D. A., & Kemm, R. E. (2009). Promoting student-centered active learning in lectures with a personal response system. *American Journal of Physiology - Advances in Physiology Education, 33*(1), 60–71.
Gipps, C. V. (2005). What is the role for ICT-based assessment in universities? *Studies in Higher Education, 30*(2), 171–180.
Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in quantitative methods for psychology, 8*(1), 23–34.
Hanrahan, S. J., & Isaacs, G. (2001). Assessing self- and peer-assessment: The students' views. *Higher Education Research and Development, 20*(1), 53–70.
Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies In Educational Evaluation, 38*(1), 21–27.
Hayes, A. F., & Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures, 1*(1), 77–89.
Hood, N., Littlejohn, A., & Milligan, C. (2015). Context counts: How learners' contexts influence learning in a MOOC. *Computers & Education, 91*, 83–91.
Hovardas, T., Tsivitanidou, O. E., & Zacharia, Z. C. (2014). Peer versus expert feedback: An investigation of the quality of peer feedback among secondary school students. *Computers & Education, 71*, 133–152.
Hsia, L.-H., Huang, I., & Hwang, G.-J. (2016). Effects of different online peer-feedback approaches on students' performance skills, motivation and self-efficacy in a dance course. *Computers & Education, 96*, 55–71.
Jackson, L. (2014). *Validity and rater reliability of peer and self assessments for urban middle school students*.
Jones, I., & Alcock, L. (2014). Peer assessment without assessment criteria. *Studies in Higher Education, 39*(10), 1774–1787.
Jones, I., & Wheadon, C. (2015). Peer assessment using comparative and absolute judgement. *Studies In Educational Evaluation, 47*, 93–101.
Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*(2), 130–144.
Kehrer, P., Kelly, K., & Heffernan, N. (2013). Does immediate feedback while doing homework improve learning?. In *Paper presented at the FLAIRS 2013 - proceedings of the 26th international Florida artificial intelligence research society conference* (pp. 542–545).
Kilic, D. (2016). An examination of using self-, peer-, and teacher-assessment in higher education: A case study in teacher education. *Higher Education Studies, 6*, 136.
Kilic, G. B., & Cakan, M. (2007). Peer assessment of elementary science teaching skills. *Journal of Science Teacher Education, 18*(1), 91–107.
Kizilcec, R. F., Davis, G. M., & Cohen, G. L. (2017a). Towards equal opportunities in MOOCs: Affirmation reduces gender & social-class achievement gaps in China. *ACM conference on learning@ scale*, 121–130.
Kizilcec, R. F., Saltarelli, A. J., Reich, J., & Cohen, G. L. (2017b). Closing global achievement gaps in MOOCs. *Science, 355*(6322), 251–252.
Krippendorff, K. (1970). Estimating the reliability, systematic error and random error of interval data. *Educational and Psychological Measurement, 30*(1), 61–70.
Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication Methods and Measures, 5*(2), 93–112.
Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). Thousand Oaks, CA: Sage Publications, Inc.
Krippendorff, K., & Bock, M. A. (2009). *The content analysis reader*. Sage.
Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*(1), 79–97.
Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., et al. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction, 20*(6), 31, 33.
Lange, R. T. (2011). Inter-rater reliability. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (p. 1348). New York, NY: Springer New York.
Lavrakas, P. (2008). *Encyclopedia of survey research methods*.
Lawrence-Brown, D. (2004). Differentiated instruction: Inclusive strategies for standards-based learning that benefit the whole class. *American Secondary Education, 32*(3), 34–62.
Leenknecht, M., & Prins, F. (2018). Formative peer assessment in primary school: The effects of involving pupils in setting assessment criteria on their appraisal and feedback style. *European Journal of Psychology of Education, 33*(1), 101–116.
Lehmann, M., Christensen, P., Du, X., & Thrane, M. (2008). Problem-oriented and project-based learning (POPBL) as an innovative learning strategy for sustainable development in engineering education. *European Journal of Engineering Education, 33*(3), 283–295.
Li, H., Xiong, Y., Zang, X., L Kornhaber, M., Lyu, Y., Chung, K. S., et al. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264.
Lin, G.-Y. (2018). Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education; Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system, 116*, 81–92.
Liu, N., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Wesley.
Lüftenegger, M., Schober, B., van de Schoot, R., Wagner, P., Finsterwald, M., & Spiel, C. (2012). Lifelong learning as a goal – do autonomy and self-regulation in school result in well prepared pupils? *Learning and Instruction, 22*(1), 27–36.
Lukassen, N. B., Pedersen, A., Nielsen, A., Wahl, C., & Sorensen, E. K. (2014). Digital education with IT: How to create motivational and inclusive education in blended learning environments using flipped learning - a study in nurse education. *Proceedings of the European Conference on e-Learning, ECEL, 2014-January*, 305–312.
Luo, H., Robinson, A., & Park, J. (2014). Peer grading in a MOOC: Reliability, validity, and perceived effects. *Online Learning Journal, 18*(2).
Marks, L., & Jackson, M. (2013). Student experience of peer assessment on an MSc programme. *Bioscience Education, 21*(1), 20–28.
Marton, F., & Bowden, J. (1999). The university of learning:: Beyond quality and competence. *Education + Training, 41*(5) (ii).
McKeachie, W. J., Pintrich, P. R., Lin, Y., & Smith, D. (1986). *Teaching and learning in the college classroom*. Ann Arbor, MI: University of Michigan.
Meek, S. E. M., Blakemore, L., & Marks, L. (2017). Is peer review an appropriate form of assessment in a MOOC? Student participation and performance in formative peer review. *Assessment & Evaluation in Higher Education, 42*(6), 1000–1013. https://doi.org/10.1080/02602938.2016.1221052.
Meletiadou, E., & Tsagari, D. (2014). An exploration of the reliability and validity of peer assessment of writing in secondary education. *Major Trends in Theoretical and Applied Linguistics, 3*, 235–250 (Sciendo Migration).
Miller, T. (2009). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education, 34*(2), 181–192.
Moerkerke, G. (1996). *Assessment for flexible learning*. Utrecht: Lemma.

Mullen, J., Byun, C., Gadepally, V., Samsi, S., Reuther, A., & Kepner, J. (2017). Learning by doing, high performance computing education in the MOOC era. *Journal of Parallel and Distributed Computing, 105*, 105–115.
Ng, E. (2014). Using a mixed research method to evaluate the effectiveness of formative assessment in supporting student teachers' wiki authoring. *Computers & Education, 73*, 141–148.
Ng, E. (2016). Fostering pre-service teachers' self-regulated learning through self- and peer assessment of wiki projects. *Computers & Education, 98*, 180–191.
Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199–218.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology, 3*(1), 1–18. https://doi.org/10.1016/0022-2496(66)90002-2.
Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education, 21*(3), 239–250.
Orsmond, P., Merry, S., & Reiling, K. (2000). The use of student derived marking criteria in peer and self-assessment. *Assessment & Evaluation in Higher Education, 25*(1), 23–38.
Piech, C., Huang, J., Chen, Z., Do, C., Ng, A., & Koller, D. (2013). *Tuned models of peer assessment in MOOCs*. arXiv preprint arXiv:1307.2579.
Quintana, C., Reiser, B. J., Davis, E. A., Krajcik, J., Fretz, E., Duncan, R. G., et al. (2004). A scaffolding design framework for software to support science inquiry. *The Journal of the Learning Sciences, 13*(3), 337–386.
Raman, K., & Joachims, T. (2014). Methods for ordinal peer grading. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1037–1046.
Reinholz, D. (2016). The assessment cycle: A model for learning through peer assessment. *Assessment & Evaluation in Higher Education, 41*(2), 301–315.
Richmond, S., Shaw, W. C., O'brien, K. D., Buchanan, I. B., Jones, R., Stephens, C. D., et al. (1992). The development of the PAR index (peer assessment rating): Reliability and validity. *The European Journal of Orthodontics, 14*(2), 125–139.
Sadler, P. M., & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1–31. https://doi.org/10.1207/s15326977ea1101_1.
Salehi, M., & Masoule, Z. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. *Southern African Linguistics and Applied Language Studies, 35*(1), 1–15.
Schaer, P. (2012). *Better than their reputation? On the reliability of relevance assessments with students*.
Scott, W. A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly, 19*, 321–325.
Shieh, G. (2016). Exact power and sample size calculations for the two one-sided tests of equivalence. *PloS One, 11*(9).
Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.
Sluijsmans, D., Brand-Gruwel, S., & van Merriënboer, J. (2002a). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454.
Sluijsmans, D., Brand-Gruwel, S., van Merriënboer, J., & Bastiaens, T. (2002b). The training of peer assessment skills to promote the development of reflection skills in teacher education. *Studies In Educational Evaluation, 29*(1), 23–42.
Staubitz, T., Petrick, D., Bauer, M., Renz, J., & Meinel, C. (2016). Improving the peer assessment experience on MOOC platforms. In *Proceedings of the third (2016) ACM conference on learning @ scale* (pp. 389–398).
Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75.
Suen, H. K. (2014). Peer assessment for massive open online courses (MOOCs). *The International Review of Research in Open and Distributed Learning, 15*(3).
Sung, Y., Chang, K., Chang, T., & Yu, W. (2010). How many heads are better than one? The reliability and validity of teenagers' self-and peer assessments. *Journal of Adolescence, 33*(1), 135–145.
Susskind, R. E., & Susskind, D. (2015). *The future of the professions: How technology will transform the work of human experts*. USA: Oxford University Press.
Thelwall, M. (2000). Computer-based assessment: A versatile educational tool. *Computers & Education, 34*(1), 37–49.
Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273–286.
Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.
Topping, K. J. (2009). Peer assessment. *Theory Into Practice, 48*(1), 20–27.
Topping, K. J. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology, 1*(1), 1–17. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.00 [7].
Tsai, C., Lin, S. S. J., & Yuan, S. (2002). Developing science activities through a networked peer assessment system. *Computers & Education, 38*(1), 241–252.
Van Den Bossche, P., Gijselaers, W. H., Segers, M., & Kirschner, P. A. (2006). Social and cognitive factors driving teamwork in collaborative learning environments: Team learning beliefs and behaviors. *Small Group Research, 37*(5), 490–521.
Van der Pol, J., Van den Berg, B., Admiraal, W. F., & Simons, P. R. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education, 51*(4), 1804–1817.
Van Hattum-Janssen, N., & Lourenço, J. (2008). Peer and Self-Assessment for First-Year Students as a Tool to Improve Learning. *Journal of Professional Issues in Engineering Education and Practice, 134*(4), 346–352. https://doi.org/10.1061/(ASCE)1052-3928(2008)134:4(346).
Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279.
Watson, S. L., Watson, W. R., Yu, J. H., Alamri, H., & Mueller, C. (2017). Learner profiles of attitudinal learning in a MOOC: An explanatory sequential mixed methods study. *Computers & Education, 114*, 274–285.
Webb, J. M., Stock, W. A., & McCarthy, M. T. (1994). The effects of feedback timing on learning facts: The role of response confidence. *Contemporary Educational Psychology, 19*(3), 251–265.
Weinstein, C. E., Acee, T. W., & Jung, J. (2011). Self-regulation and learning strategies. *New Directions for Teaching and Learning, 2011*(126), 45–53.
Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37.
Xiao, Y., & Lucking, R. (2008). The impact of two types of peer assessment on students' performance and satisfaction within a Wiki environment. *The Internet and Higher Education, 11*(3), 186–193.
Yoon, H. B., Park, W. B., Myung, S., Moon, S. H., & Park, J. (2018). Validity and reliability assessment of a peer evaluation method in team-based learning classes. *Korean journal of medical education, 30*(1), 23–29.
Zapf, A., Castell, S., Morawietz, L., & Karch, A. (2016). Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate? *BMC Medical Research Methodology, 16*(1), 93.
Zimmerman, D. W., Williams, R. H., Zumbo, B. D., & Ross, D. (2005). Louis Guttman's Contributions to Classical Test Theory. *International Journal of Testing, 5*(1), 81–95. https://doi.org/10.1207/s15327574ijt0501_7.
European Higher Education Area. (2009). In *Communiqué of the conference of European ministers responsible for higher education, leuven and louvain-la-neuve, 28-29 April 2009*. European Higher Education Area (EHEA).
Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2004). *Practical resources for assessing and reporting intercoder reliability in content analysis research projects.*. Retrieved September, 2019. http://matthewlombard.com/reliability/index_print.html