



A systematic review of educational online peer-review and assessment systems: charting the landscape

Dmytro Babik¹ · Edward Gehringer² · Jennifer Kidd³ · Kristine Sunday³ · David Tinapple⁴ · Steven Gilbert⁵

Accepted: 4 February 2024
© The Author(s) 2024

Abstract

Over the past two decades, there has been an explosion of innovation in software tools that encapsulate and expand the capabilities of the widely used student peer assessment. While the affordances and pedagogical impacts of traditional in-person, “paper-and-pencil” peer assessment have been studied extensively and are relatively well understood, computerized (online) peer assessment introduced not only shifts in scalability and efficiency, but also entirely new capabilities and forms of social learning interactions, instructor leverage, and distributed cognition, that still need to be researched and systematized. Despite the ample research on traditional peer assessment and evidence of its efficacy, common vocabulary and shared understanding of online peer-assessment system design, including the variety of methods, techniques, and implementations, is still missing. We present key findings of a comprehensive survey based on a systematic research framework for examining and generalizing affordances and constraints of online peer-assessment systems. This framework (a) provides a foundation of a design-science metatheory of online peer assessment, (b) helps structure the discussion of user needs and design options, and (c) informs educators and system design practitioners. We identified two major themes in existing and potential research—orientation towards scaffolded learning vs. exploratory learning and system maturity. We also outlined an agenda for future studies.

Keywords Architectures for educational technology system · Peer assessment · Peer review · Online systems · Systematic review · Framework

Introduction

Peer assessment has been widely used in pedagogical practice and intensively studied by education researchers since the 1970s. The seminal work by Topping (1998) that provided a comprehensive review of 31 studies and offered a typology of peer assessment has been cited by nearly every paper on the topic published since then (over 3270 citations shown by Google Scholar as of March 2024). Educational peer assessment (also called student peer review) was defined as “an arrangement in which individuals consider the amount, level, value, worth, quality, or success of the products or outcomes of learning of peers of similar

Extended author information available on the last page of the article

status” (Topping, 1998, p. 250) and typically entails “the quantitative evaluation of and qualitative feedback to a learners’ performance by another learner” (Patchan et al., 2017, p. 2263). Also widely studied and closely associated with peer assessment practices is self-assessment, or self-evaluation, i.e., the evaluation of an artifact or a contribution by its own creator (Sargeant et al., 2008; Topping, 2003). The tandem of peer and self-assessment, used in conjunction with conventional instructor assessment, promises the benefits of high-level learning (Bostock, 2000; Boud & Falchikov, 1989; Falchikov & Boud, 1989; Sadler & Good, 2006; Sargeant et al., 2008).

Multiple studies and meta-reviews present extensive evidence of positive pedagogical outcomes of peer and self-assessment in various contexts (Chang et al., 2021; Double et al., 2020; Li et al., 2020; Misiejuk & Wasson, 2021). By introducing self- and peer assessment in face-to-face or online classrooms, instructors and course designers attempt to promote constructivist learning by (a) engaging learners in high-level cognitive activities of solving complex, open-ended, ill-structured problems and reviewing solutions (Wooley et al., 2008); (b) quantifying and understanding interactions among learners (Berg et al., 2006; Rotsaert et al., 2018; Willey & Gardner, 2009); (c) providing scalable, timely, and targeted feedback (Taraborelli, 2008); (d) creating sustainable, self-regulating, self-curating learning environments in which students are motivated to produce high-quality solutions and provide extensive, professional, and developmental feedback to each other (Baikadi et al., 2016; Rotsaert et al., 2018; Steffens, 2006).

With increasing use of computer information technologies (CITs) in education, computer-aided peer assessment has been praised as an enabler of pedagogies for developing both higher-level competencies (Bull & McCalla, 2002; Topping, 2005; Verma, 2015), and scalable methods for reliable assessment in large and online classes (Kulkarni et al., 2013; Raman & Joachims, 2014; Shah et al., 2013). CITs not only make peer assessment more *efficient* (i.e., simpler, faster, and cheaper), but also more *effective* and *versatile* (i.e., enabling many different types of interactions among peer learners as well as between learners and instructors that were not possible with face-to-face, paper-and-pencil peer- and self-assessment techniques) (Søndergaard & Mulder, 2012). In over 25 years that web-based CITs have been used in education, instructors, course designers, and software developers alike have made many attempts to computerize peer assessment by employing existing generic tools (e.g., Google Forms and Spreadsheets) or building specialized educational peer-review and assessment applications. These systems began emerging in the late 1990s and seem to have peaked between 2009 and 2018; by now some of the systems have reached maturity and are widely used, while others became defunct. The better known systems, built professionally and used in at least several universities, include Aropä (Hamer, 2006; Hamer et al., 2007), Calibrated Peer Review (Russell, 2001), CritViz (Tinapple et al., 2013), CrowdGrader (de Alfaro & Shavlovsky, 2014), Expertiza (Gehring et al., 2007), Mobius SLIP (Babik et al., 2012, 2017a, 2017b), Peerceptiv/SWoRD (Cho & Schunn, 2007), peerScholar (Joordens et al., 2009), and SPARKplus (Wu et al., 2010). These systems are used by educators in a wide variety of courses in STEM, liberal arts, and business disciplines, such as English and Writing (Cho & Schunn, 2007), Sciences (Russell, 2001), Computer Science (Søndergaard & Mulder, 2012), Business (Babik et al., 2017a, 2017b), Visual Art and Design (Tinapple et al., 2013), among others (Alqassab et al., 2023).

Evidently, most of these peer assessment systems have been designed and built without prior exploring other systems to see what methods, techniques, and implementations already exist, have been tested and are appropriate for a particular pedagogical need, and how they could be improved. This likely happened because, despite the abundance of literature on peer assessment, there is little literature specifically focusing on the design of

CIT-supported peer-assessment systems. Several attempts have been made to systematize CIT-supported peer assessment by reviewing research literature (Alqassab et al., 2023; Fu et al., 2019; Topping, 2023) or existing applications (Luxton-Reilly, 2009), developing inventories of peer assessment diversity (Gielen et al., 2011), and proposing classifications of peer assessment emphases (Søndergaard & Mulder, 2012) (see the Literature Review section for further details). However, the need for a *comprehensive, systematic survey* of educational CIT-enabled peer-assessment systems that explores and generalizes affordances and constraints of these systems based on a structured research framework still has not been addressed (Alqassab et al., 2023). Such a framework would inform new system designs, guide the improvement of existing ones, and help various categories of users navigate the much-needed innovations in this domain. The current study aimed to fill this gap by conducting such a survey and developing the said framework.

The purpose of this paper is to present our framework and summarize the key findings of our study. This framework serves as a foundation of a design-science metatheory of CIT-enabled peer assessment. It helps structure the discussion of user needs and design options that address these needs, and informs educators and system design practitioners.¹ For this study, we adopted the term *online peer-review and assessment (OPRA) system* and defined it as a web-based software application purposefully designed and developed to facilitate and automate student self- and peer-review and assessment process. Specifically, OPRA systems support collecting submission artifacts, allocating artifacts to peer reviewers for critiquing and/or evaluating, setting deadlines, guiding and scaffolding reviewers' qualitative and quantitative feedback, aggregating quantitative evaluations, and conducting other components of the peer-review process. This definition covers a broad range of applications for educational peer review and assessment, also referred to in the literature as computer²-supported³ (or digital) peer assessment⁴ systems. Many synonymous, albeit competing, terms found in the literature, such as “*computer-mediated peer review (CMPR)*” (Carlson & Smith, 2017), when applied to the current generation of web 2.0- and cloud-based technologies, appear outdated and limited in spectrum. Therefore, we propose the term “OPRA” as more general and current. Educational OPRA systems are a subset of a broader class of social computing systems that explicitly or implicitly involve peer review (including social networking and social media applications, such as wikis, blogs, discussion forums) and collaborative editing and annotating (e.g., Google Docs, Hypothesis). OPRA systems, however, are distinct by being designed specifically for educational peer-review practice. The aim of our systematic review and the proposed framework is to guide the future design of OPRA systems by addressing these important research questions: *What is the current landscape of educational online peer-review and assessment technology? What essential common characteristics of the OPRA systems address user needs? How are these characteristics defined by pedagogical objectives? How does this technology support, advance, and transform the pedagogical process? How does the diversity of contexts, in which OPRA systems are applied, define the diversity of features these systems have?* We constructed this research framework for the systematic exploration of the current state of educational OPRA systems based on the rigorous literature review, examination

¹ Due to space limitations, detailed discussions of specific user needs, designs, and learning implications found in this study will be presented in follow-up publications.

² Other known terms: technology-, IT-, CIT-, ICT-, network-, internet-, web-, cloud-.

³ Other known terms: aided, assisted, automated, based, enabled, mediated.

⁴ Other known terms: review, evaluation.

of individual OPRA systems, and analysis of data collected through focus-group discussions, questionnaires and interviews. We used this framework to categorize functionality of OPRA systems developed since 2005, analyze design choices made by their originators, identify affordances and limitations of these choices, inform new development efforts, and suggest future research agenda. We consciously excluded from our study the systems developed prior to the advent of the web 2.0, since they were limited compared to modern systems and have largely disappeared. In this paper, we did not aim to analyze and report all possible *variations of the peer-review process*; instead, we focused on the high-level view of *affordances* enabled by these variations. We sought to aggregate and systematize knowledge of the core characteristics of many existing OPRA systems and to use them as illustrative examples to help users, designers, and researchers make informed decisions, rather than to present in detail any specific individual OPRA systems.

The authors of this paper are researchers and instructors from several universities, who have created their own peer assessment systems, and subsequently worked together under the umbrella of the NSF-sponsored PeerLogic Project (Gehringer, 2019) to pursue several goals: (1) to systematically explore the domain of CIT-enabled peer assessment systems; (2) to develop an arsenal of web services for a wide range of applications in such systems; and (3) to develop a meta-language and a data repository for in-depth research of student peer review. We address this study to researchers and practitioners interested in OPRA and motivated to advance its use. We invite educational assessment and learning analytics researchers, system designers, educational technologists, instructional designers, and instructors, who enter the field of OPRA, to use this review as a guide to various functionalities and design options. Researchers in learning analytics will discover what data can be extracted from OPRA systems and mined to demonstrate learning outcomes. Educational software designers will learn from what has been developed and implemented in the past and incorporate this knowledge in their future projects. Instructors applying peer-review pedagogy in their classes will find what systems and functionalities exist to make informed choices about what approaches would best meet their needs. Oftentimes, instructors turn with these questions to ed-tech specialists and instructional designers; thus, the latter may also find this work useful. Conversely, marketers of OPRA systems may identify the unique and differentiating features of their products and better inform their target users.

This paper is organized as follows. Sect. “[Literature review](#)” presents a literature review. In Sect. “[Methodology](#)”, we define key terms, describe our methodology, the framework, and its application to systematic survey of multiple existing educational OPRA systems. In Sect. “[Results and discussion](#)”, we discuss the findings of our analysis, propose a general research agenda for future studies of OPRA, and summarize contributions and limitations of the paper.

Literature review

The focus of this study is information technology that enables the pedagogy of peer assessment. We explored existing literature to conduct a systematic analysis of the multitude, variety, and complexity of such implementations, functionalities, and design choices in OPRA. Several attempts have been made to survey computerized peer-assessment practices (Bouzidi & Jaillet, 2009; Chang et al., 2021; Davies, 2000; Doiron, 2003; Gikandi et al., 2011; Luxton-Reilly, 2009; Tenório et al., 2016; Topping, 2005), or some specific aspects of peer assessment, such as approaches to reliability and

Table 1 Layout of the research framework for surveying educational OPRA systems. (With the example of user need, use case, features, and design options)

Problem domain layers (implementation-independent, user needs)		Solution domain layers (implementation-dependent, system affordances)	
User needs	Use cases	Features	Design options
1. Eliciting evaluations and critiques	1.1 Reviewer provides quantitative evaluation	1.1.1 Rubric	Holistic rubric Specific (analytic) rubric
		1.1.2 Scale	Cardinal (rating) scale Ordinal (ranking) scale Hybrid scale “Exotic” scale
	1.2 Reviewer provides qualitative critiques	1.2.1
			...
		1.2.2
			...
2. ...	2.1 ...	2.1.1
		2.1.2
			...

validity of peer evaluations (Gehring, 2014; Misiejuk & Wasson, 2021; Patchan et al., 2017). However, meta-analysis of OPRA systems is complicated because their design space has high dimensionality; OPRA practices and designs vary across many disciplines in many different ways (Søndergaard & Mulder, 2012).

With the aim of describing the common, as well as unique, features of the OPRA systems, Luxton-Reilly (2009) conducted a systematic review of literature and identified 18 OPRA systems:

- Six *generic* systems: Peer Grader, Web-SPA, OPAS, CeLS, PRAISE, and Aropä (Luxton-Reilly, 2009, Table 1, p. 213),
- Seven *domain-specific* systems: Calibrated Peer Review (CPR), CAP, Praktomat, SWoRD, PeerWise, peerScholar, and an unnamed system by Sitthiworachart and Joy (2004) (Luxton-Reilly, 2009, Table 2, p. 217), and
- Five *context-specific* systems: Peers, NetPeas, OASYS, PEARS, and an unnamed system by Wolfe (2004) (Luxton-Reilly, 2009, Table 1, p. 213).

[At the time of this writing, five systems from the Luxton-Reilly (2009) review, namely, PeerGrader, Web-SPA, OPAS, CeLS, and PRAISE, appeared to be defunct, not maintained or not extensively used (Purchase & Hamer, 2017)]. Luxton-Reilly’s study identified the following common elements of the OPRA systems: *anonymity*, *allocation and distribution*, *grading/marking criteria* (rubrics), *calculating peer grade/mark* (aggregation), *controls for quality of reviews*, and *workflow*. The author noted a significant tradeoff between *flexibility of an OPRA system* (its ability to accommodate a variety of workflows and use cases) and its *ease of use*; that is, a more flexible and effective system

Table 2 Compendium of Systematic Analysis of OPRA Systems

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
1	Eliciting sub-missions	What types and formats of sub-missions can be subjected to review and assessment?	1.1	Author creates and posts submission for assessment	1.1.1 1.1.2	Assessed submission type Submission artifact format	Artifact Observed behavior Plain text Rich text format (WYSIWYG RTF) or embedded HTML document External source URL Single file attachment Multiple file attachments As an integral part of the submission artifact (embedded in the primary submission document) As a separate entry	All systems in the sample CATME, Mobius SLIP All systems in the sample Any system with the Plain text option All systems in the sample (except PeerWise) unknown Any system
			1.2	Author provides instructions/directions/requests for specific actions/help from reviewers	1.2.1	Author's request for specific reviewer actions (input)		

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
2	Structuring peer assessment process (work-flow)	What are the common components of the online peer review process? What variations of this process do exist?	NN	2.1 Instructor or system allocates reviewers to submissions (submissions to reviewers)—Selecting from the participant pool	NNn	2.1.1 Automatic allocation by system	Automatic random idiosyncratic allocation	All systems in the sample
							Automatic random patterned allocation	Mobius SLIP, OSBLE, Peerceptiv, SPARKPlus
							Automatic non-random allocation	CrowdGrader, OSBLE, peerScholar's "tournament" allocation, PRAZE
					2.1.2	Manual allocation by Instructor	Random	Most systems
					2.1.3	Self-allocation of reviewers	Intentional/deliberate Pure self-selection Hybrid = self-selection of topic + automatic allocation of submission	Most systems PeerWise, SPARKPlus Aropá, Expertiza, PRAZE

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
	2.2	Instructor or system allocates reviewers to submissions (submitters to reviewers)—Peer network mapping	2.2.1	Clustering	Peer review clusters/groups	Dispersed peer review network	Eli Review, Mechanical TA, Mobius SLIP	Most systems
			2.2.2	Directionality (reciprocity)	Fully symmetric/reciprocal/mutual review	Asymmetric/non-reciprocal review	Eli Review, Mechanical TA, Mobius SLIP	Most systems
			2.2.3	Cardinality/Multiplicity of submission and review authorship	Individual-to-individual	Team-to-team	All systems	Aropä, Eli Review, Mobius SLIP, OSBLE, peerScholar, PRAZE, SPARKPlus
					Individual-to-team		Aropä, CritViz, CrowdGrader, Expertiza, Mobius SLIP, OSBLE, Peerceptiv, Praze, SPARKPlus	
					Team-to-individual		Aropä, OSBLE, and SPARKPlus	

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Design option	Example of the design option implementation *
N	User need	Descriptive question	NN		
	2.3	Instructor or system allocates reviewers to submissions (submissions to reviewers)—Equalizing allocation	NNn	Allocation of submissions per reviewer	Aropä, CritViz, Expertiza, Mobius SLIP, SPARKPlus, CrowdGrader, Mobius SLIP, Expertiza
			2.3.1	Even Uneven	
			2.3.2	Allocation of reviews per submission	
			2.3.3	Timing of allocation equaling	Most systems Aropä, CPR, CritViz, Emarking, Mechanical TA
				Re-allocation (adjusted allocation) Dynamic allocation	Eli Review, Mobius SLIP, OSBLE, SPARKPlus, Expertiza, CrowdGrader, Peerceptiv, peerScholar, PRAISE, SPARKPlus

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	Design option
	2.4	Instructor determines how participants assess artifacts (Review and assessment workflow)	NN	Use case	NNn	Assessing actors	Peer Self	All systems Aropä, CPR, Mobius SLIP, peerScholar, SPARKKPlus
			2.4.1				Instructor Teaching assistant External assessors	All systems Mechanical TA Expertiza, Caesar
			2.4.2			Assignment modular-ity	Assignment-based	CPR, CrowdGrader, Expertiza, Mechanical TA, Mobius SLIP, Peerceptiv, peerScholar, PeerWise, PRAZE, PRAISE
							Task-based	Aropä, CritViz, eMarking, Eli Review, OSBLE, SPARKKPlus
			2.4.3			Review (artifact-feedback exchange) interactions	Single loop Double loop	All systems Aropä, CrowdGrader, Eli Review, eMarking, Expertiza, Mobius SLIP, peerScholar, Peerceptiv, Praktomat, PRAZE
							Continuous feedback	Mobius SLIP

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *		
N	User need	Descriptive question	NN	System feature	Design option	
	2.5	System conceals participants' identities	NNn	Review process anonymity	Double-blind	All systems
			2.5.1		Single-blind (only reviewer knows)	eMarking, OSBLE, SPARKPlus
					Single-blind (only author knows)	eMarking, OSBLE, SPARKPlus
					Fully open	Reported in Eli Review, eMarking, Expertiza, Mobius SLIP, OSBLE, peerScholar, SPARK-Plus; de-facto can be found in any system
			2.5.2	Identity ambiguity	Partial ambiguity	Peerceptiv, Expertiza
					Complete ambiguity	Mobius SLIP
			2.5.3	Post-review confidentiality (privacy) and anonymity	No disclosure Anonymous public disclosure	All systems
					Full public disclosure	CritViz, Expertiza

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
3	Eliciting evaluations and critiques	How do reviewers input assessment data (quantitative and qualitative, structured and semi-structured)? What input controls are used to elicit responses?	3.1	Reviewer provides quantitative evaluation	3.1.1	Rubrics/evaluation criteria	Holistic	Mobius SLIP, CritViz
					3.1.2	Scale	Specific/analytic Cardinal/rating Ordinal/ranking Hybrid Exotic	Expertiza, CPR, Peerceptiv Most systems CritViz, Eli Review, CeLS, Mobius SLIP, peerScholar, SPARKPlus Mobius SLIP OSBLE

Table 2 (continued)

User needs—problem domain (implementation-independent) layers			System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *
N	User need	Descriptive question	NN	System feature	
			NNn	Design option	
	3.2	Reviewer provides qualitative critiques/comments	3.2.1	Critique artifact format	All systems
				Plain text	
				Rich text format/hyper-text/embedded HTML/URL	
				Inline annotation of plain or rich text	Praktomat
				Inline file annotation of attached files	Mobius SLIP, Canvas through APIs; no systems use native functionality
				Upload annotated submission file	
				Multimedia attachments	
				Discrete choice of predefined options	
			3.2.2	Non-contextualized, holistic	All systems
				Contextualization of critique	
				Contextualized (“many comments per submission in specific locations” or inline file annotation or comments associated with scores (e.g. justifying))	Expertiza, CritViz, Peereceptiv, CPR (via rubric); Mobius SLIP, Canvas (via inline file annotation)

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
4	Aggregating results	How are multi-peer assessment results aggregated? What assessment metrics are used?	4.1	System computes attainment metric	4.1.1	Aggregation of multiple peer evaluations	Average Median	Most systems Aropä, Eli Review, CrowdGrader, Mechanical TA, PRAZE peerScholar unknown
			4.2	System detects inaccurate/unreliable/inconsistent peer evaluations	4.2.1	Measuring deviations (spread, controversy) in evaluations received by an artifact	Deviation from mean (DFM) Deviation from co-evaluators (DFC)	All systems Mobius SLIP
			4.2.2			Measuring deviations (bias) of evaluations given by a reviewer	Deviation from mean (DFM) Deviation from co-evaluators (DFC)	Mobius SLIP Mobius SLIP

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
5	Communicating/presenting results	How are peer assessment results presented to participants, instructors or third-party users? What representations are used to convey the aggregated results of multi-peer assessment?	5.1	System presents peer evaluations received and given	5.1.1	Level of detail	Individual peer-to-peer evaluations Summarized/aggregated evaluations Tables (structured numeric presentation) Visualizations (structured visual presentation)	Most systems Most systems
			5.2	System presents peer critiques received and given	5.2.1	Level of detail	Individual peer-to-peer critiques Summarized/aggregated critiques	All systems CritViz, Eli Review, Expertiza, Mobius SLIP, Peereptiv, SPARKPlus
			5.3	System generates transferable tokens of attainment	5.3.1	Micro-credentialing	Digital badges	Large number of diverse implementations Mechanical TA, PeerWise

Table 2 (continued)

User needs—problem domain (implementation-independent) layers		System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *				
N	User need	Descriptive question	NN	Use case	NNn	System feature	Design option	
6	Improving review value/quality (Evaluation accuracy and critique quality)	How can evaluation accuracy and critique value be controlled and improved?	6.1	System corrects for inaccurate/unreliable/inconsistent peer evaluations	6.1.1	Adjusting received evaluations	Attainment as weighted average of peer evaluations; less reliable evaluations receive lesser weight Ignoring outliers (special case of weighting) Reputation (weights based on historic performance)	CrowdGrader, Expertiza, Peerceptiv Mobius SLIP
			6.2	System trains reviewers to evaluate more accurately	6.2.1	Comparing reviewer's evaluation of sample artifact to expert evaluations	Calibration	CPR, Peerceptiv
			6.1.2		6.1.2	Adjusting reviewer's performance metric based on inaccuracy of provided evaluations	Grade adjustment	

Table 2 (continued)

User needs—problem domain (implementation-independent) layers			System affordances—solution domain (implementation-dependent) layers		Example of the design option implementation *
N	User need	Descriptive question	NN	Use case	
	6.3	System holds reviewers accountable for critique quality	NN	Use case	
	6.3.1		NNn	Human evaluation of critiques	Expertiza, Mobius SLIP, Peereptiv, PRAZE
	6.3.2		NNn	Automated text analysis of critiques	Expertiza, Atropä
	6.3.3		NNn	Involving participants' in creating rubric	All systems
				Rejoinders	Expertiza
				Peer meta-reviewing	Expertiza, Atropä
				Instructor meta-reviewing	All systems
				Measuring comment length and encouraging longer comments	Expertiza
				Automated meta-reviewing (sentiment text analysis) with immediate feedback	

may be too complex to use for a user with weaker computer skills or a lack of understanding of the processes. This tradeoff highlights the need for a comprehensive analysis of various affordances of the OPRA systems. Luxton-Reilly (2009) also called for more usability studies and further evaluation studies of differences among the OPRA systems.

Gielen et al. (2011) updated Topping's (1998) typology of peer assessment by reviewing studies on educational peer assessment published between 1997 and 2006. Specifically, they refined Topping's variables, identified new variables, dimensions and values, and extended variable clustering proposed by Berg et al. (2006). They developed a classification framework called an *inventory of peer assessment diversity* that focused on the organizational aspects of the peer-assessment processes rather than use-case implementations in OPRA applications. While discussing the *contact*, *time*, and *place* characteristics of peer assessment, they concluded that the "internet-based learning environments are now often the preferred location for peer assessment" (Gielen et al., 2011, p. 146).

Goldin et al. (2012) highlighted advantages of computer tools specially designed for peer review and assessment (which we define as OPRA systems) over general-purpose applications (including file sharing systems, online discussion boards, etc.), such as the ability to track the interactions of peers in greater depth and to manipulate specific components of peer interactions. The authors also emphasized many variations in the OPRA process (even within a narrow context, such as academic writing) and distinctions between peer review and similarly sounding activities, such as peer editing and peer evaluation.

Søndergaard and Mulder (2012) explored peer reviewing as a source of formative feedback in the more general context of collaborative learning, specifically in the context of collaborative learning in STEM disciplines. They identified the essential attributes for OPRA systems: *automation* (including anonymization and distribution of artifacts); *simplicity* (including easy-to-use, intuitive, and attractive user interface; integration with LMS; technical support); *customizability* (including handling any file format and creating individualized review rubrics); and *accessibility* (free, web-based, globally available, mobile). In addition, they discussed other interesting desirable attributes, such as *rule-based review allocation* (distribution); *reviewer training/calibration*; *similarity checking*; *reporting tools* (for review comparisons and instructor monitoring). Based on these attributes, Søndergaard and Mulder (2012) identified four *approaches to formative peer review*, namely *training-oriented*, *similarity-checking-oriented*, *customization-oriented*, and *writing-skills-oriented*, and illustrated implementations of these attributes using four OPRA systems, respectively Calibrated Peer Review, PeerMark, PRAZE, and Peerceptiv/SWoRD. To the best of our knowledge, to date, this is the only attempt to offer a taxonomy of OPRA systems based on a systematic analysis. Its limitation, however, is that the taxonomy framework was developed from only four systems, two of which are currently defunct.

Based on the analysis of five OPRA systems (Peerceptiv/SWoRD, peerScholar, PRAZE, OASIS, and Aropä), Purchase and Hamer (2017) identified the following important features of an effective OPRA system:

- *Anonymity*,
- *Peer allocation method*,
- *Submission method*,
- *Grading/marking criteria* (specifying criteria or rubric),
- *Grade/mark calculation* (aggregating peer evaluations into an attainment measure, i.e., a "grade", and a metric for evaluation discrepancies, inconsistencies, or the lack of reliability),
- *Backward feedback* (author's responses to peer reviews).

Purchase and Hamer (2017) concluded that an increasing number of instructors are willing to try peer assessment of complex, open-ended assignments in order to quickly provide more feedback to students and help them develop higher-level transferable skills, such as critical thinking, creativity, communication, and collaboration. They noted that, despite the common basic peer-review process, specifics of peer-review and assessment activities vary across different instructors; therefore, often instructors ask for specific unique features and design choices.

Wahid et al. (2016) attempted to provide a systematic analysis of the domain and form a general understanding by applying a cognitive mapping approach to find criteria for categorizing OPRA systems. They analyzed a sample of 17 systems, of which only 13 match our definition of OPRA; about half of the sample was represented by various OPRA research projects in Europe. The authors identified three dimensions for categorizing OPRA systems, namely *system design*, *efficiency*, and *effectiveness*. Within the system design dimensions, they identified six features (*anonymity*, *delivery*, *grading weightage*, *channel*, *review loop*, *collaboration*). The dimensions of efficiency and effectiveness were not well defined, but efficiency included the sole feature of *feedback timing*, and effectiveness included *rubrics*, *validation*, *reviewer calibration*, and *reverse reviews*. Despite analyzing a fairly large sample of systems, Wahid et al. (2016) concluded that the majority of systems were designed similarly, differing only in small number of features or the ways the features were implemented.

Carlson and Smith (2017) conducted in-depth comparison of two OPRA systems—Calibrated Peer Review (CPR) and Moodle’s Workshop—based on their set of four criteria for an effective OPRA system:

1. Does the system include a *cohesive mental model* that deconstructs the process and demonstrates staged problem solving?
2. Does the system include *scaffolding* to move students forward, both in task accomplishment and in enhanced independent learning?
3. Does the system *encourage students to learn* from peer feedback?
4. Does the system provide *data/outcomes* for instructors to assess both assignment-specific and programmatic gains for individuals and for larger aggregates?

Albeit comparing only two OPRA systems, this study was markedly different from other studies (typically focused either on pedagogical or technological aspects of peer review) in that it integrated these two views into a coherent analysis of how technological affordances and constraints translate into pedagogical effects. Carlson and Smith (2017) pointed out that, although the OPRA systems provide advantages over the “old-school, paper-and-pencil” process, using them is still “labor-intensive” and involves a steep learning curve for instructors because of the variety of available options. In addition, they suggested several ways to help students see the value of peer assessment. They cautioned against inflated expectations for digital applications and suggested that the true value of OPRA, as both a learning and an assessment tool, is in “*informating*” the pedagogies dealing with complex problem-solving competencies, rather than simply *automating* them.

Gehringer (2014) examined six OPRA systems (Calibrated Peer Review, CrowdGrader, Expertiza, Mobius SLIP, Peerceptiv/SWoRD, PeerWise) in order to catalog methods for improving peer-review quality (both qualitative critique content and quantitative evaluation accuracy). The identified methods are *calibration*, *reputation*, human (“manual”) and machine (“automatic”) *meta-reviewing*, *rejoinders* (feedback from the author to the reviewer), and different *scales* for evaluating critiques (cardinal/rating-based and ordinal/

ranking-based). While this study contrasted quality-control strategies for reviewing, it did not attempt to differentiate methods for improving qualitative critique content and quantitative evaluation accuracy.

Patchan et al. (2017) also studied quality-control mechanisms for assessment accuracy and critique quality. They examined literature on about 13 OPRA systems and identified the following approaches to encourage participants to provide more valuable reviews:

- For controlling *accuracy of evaluations*:
 - Reviewer weight/reputation systems/accuracy grades;
 - Calibration/training within the application;
- For controlling *quality of critiques* (reviewer comments):
 - Minimum word count;
 - Non-anonymous reviewing;
 - Instructor oversight;
 - Rejoinders (aka back-review, reverse review, double-loop feedback, meta-reviewing);
 - Automated meta-review/feedback;
 - Training outside the application.

In addition, Patchan et al. (2017) conducted an experiment in Peerceptiv/SWoRD to test two hypotheses:

- (a) *Direct accountability hypothesis*: positive effects of holding participants accountable for the accuracy of evaluations;
- (b) *Depth-of-processing hypothesis*: positive effects of holding participants accountable for the quality of critiques.

In this experiment, they conceptualized holding participants accountable in three ways: (i) being “graded” only on the quality of critiques they give, (ii) being “graded” only on the accuracy of evaluations they give, and (iii) being “graded” on both the quality of critiques and the accuracy of evaluations. The experiment demonstrated that:

- (a) Both types of participants’ perceptions about being held accountable (i) and (iii) positively affect evaluation accuracy; at the same time, participants’ perceptions of (ii) does not significantly affect evaluation accuracy;
- (b) Similarly, both types of participants’ perceptions (i) and (iii) positively affect critique quality (measured as approximated critique length and the number of longer comments); at the same time, participants’ perceptions of (ii) does not significantly affect critique quality.

Thus, overall, this study did not support the *direct accountability hypothesis* but did support the *depth-of-processing hypothesis*. This study examined the quality-control mechanisms and offered a basic classification of this very important aspect of OPRA. In our framework, we extended and refined this classification.

Misiejuk and Wasson (2021) conducted scoping review of the studies exploring *backward evaluation* (or ‘the feedback that an author provides to a reviewer about the quality of the review’ per Luxton-Reilly (2009)) published during 2000–2021. In this earliest literature review addressing backward evaluations in PA, the authors focused specifically on the characteristics of the empirical studies rather than user needs and designs of this feature in various OPRA systems. They also pointed out variety and diversity of terminology and suggested a need to establish common vocabulary to describe various aspects of OPRA processes and systems.

Attempts have also been made to create *generalized models* of peer review and assessment that could guide the design of OPRA systems. For example, Millard et al. (2007) and Millard et al. (2008) analyzed various peer-review processes (and in particular, reviewer *allocation patterns*) and proposed a *canonical model* integrating a set of *peer-review cycles*, each of which is defined by a set of *peer-review transforms*. Based on this model, they created a prototype of generalist web-based OPRA system called PeerPigeon and a Domain Specific Language (DSL). The project, however, at this point appears to be discontinued. Pramudianto et al. (2016), Song et al. (2016), and Babik et al. (2018) described a *generalized domain model* of peer review and assessment intended for integrating data from multiple OPRA systems into large-scale research data sets. At the time of writing, this work had been largely in progress, and we found no other attempts to present generalized OPRA domain models in the literature.

In summary, the large amount of research and development on OPRA systems has created a need for a systematic, comprehensive, framework-based analysis of this domain. Previous studies of the OPRA systems have been limited in *scope* (the number of considered characteristics, factors, attributes, or variables), *scale* (the number of examined systems), and *depth* of analysis (the level of considered detail, structuration, conclusions, and generalizations drawn). This study is an effort to fill the gap.

Methodology

Definitions

The current peer-assessment literature lacks standard terminology. Different authors use diverse terms for the same concepts or the same terms for different concepts. For the purposes of this study, we used the following definitions:

- A *user* is any person who interacts with an OPRA system.
- An *instructor* is a user who sets up a peer-reviewed assignment.
- A *participant* is a user who completes activities in the peer-review assignment. Typically, participants are students in a course.
- An *artifact* is any kind of digital object that represents a solution to a problem or signifies completion of a task; for example, a document posted/submitted by a student to fulfill the requirements of an assignment.
- A *submission* is an artifact or an outcome subjected to peer review.
- An *author* is a participant or a team of participants which creates and posts a submission.

- A *review* is a process and an artifact of completing a peer-review task; it includes a quantitative evaluation (assessment) and/or qualitative feedback (comments and/or critiques).
- A *reviewer* is a participant or a team who reviews and assesses submissions authored by other peers.
- An *evaluation* is a process and a result of a participant's assigning some *quantitative measure of attainment* ("score", "grade", or "mark") to an artifact.
- A *critique* is a set of *qualitative, textual, or verbal comments* on an artifact; comments provided by a given reviewer to a given submission are referred to as a *critique artifact*.
- A *rejoinder* is an author's response to a received critique or evaluation; this response may take the form of a critique, evaluation, or both; in the peer assessment literature and in OPRA systems, rejoinder is also referred to as *appeal* (Wright et al., 2015), *backward evaluation* (Misiejuk & Wasson, 2021), *backward feedback* (Purchase & Hamer, 2017), *back-review* (Goldin, 2011), *concordance* or *double-loop feedback* (Babik et al., 2017a, 2017b), or *reaction* (Babik, 2015).
- *Attainment* is the degree to which a participant succeeded in solving a particular problem or in performing a specific complex task; attainment reflects the degree to which an artifact possesses some desired properties or values, such as efficacy, verity, accuracy, utility, or style.

In a typical peer assessment process, participants, as authors, create an artifact and make a submission (individually or as a team); the submission artifact is distributed (based on predetermined settings) to several participants, who now act as peer reviewers. Reviewers complete evaluations of submissions they received to review (typically using a preset scale and/or criteria/rubric), as well as provide critiques of the submissions; this step may also involve some form of self-assessment. The evaluation and critique data are then processed and distributed back to the authors. In many OPRA systems this basic process is followed by additional steps, such as rejoinder, or complemented by various treatments, such as training, calibration, or instructor feedback.

Note that we decompose reviewing into two separate activities: *critiquing* (e.g., providing *qualitative*, typically textual, feedback regarding attainment and possible improvements of the artifact), and *evaluating* (i.e., expressing judgment by assigning a *quantitative measure of attainment* to an artifact).

Data collection and analysis

We applied a grounded theory approach to systematically construct our framework and develop a design-science meta-theory of OPRA systems through the analysis of data about existing systems (Babik et al., 2012; Denzin & Lincoln, 2011; Hevner et al., 2004; Martin & Turner, 1986; Strauss & Corbin, 1994). Since our intent was not to test a set of specific hypotheses, but rather to build a framework for exploring designs of a class of systems, we operated inductively. In summary, we began with our research questions, collected and examined qualitative data, identified and coded apparent repeated ideas, concepts or elements. As more data were collected and re-examined, codes were grouped into concepts, and then into categories. These categories became the basis for our framework.

Initial formal data collection was conducted through keyword search for journal publications and conference proceedings describing various educational OPRA systems.

Keywords such as “peer assessment”, “peer evaluations”, “student peer review”, “student self-assessment” and “computer-based student peer assessment” were used to identify academic publications through Google Scholar. We examined abstracts to identify articles dealing specifically with educational online peer-review and assessment systems, rather than peer review in general. We paid particular attention to articles comparing different OPRA systems. We also examined the reference lists in each of the found articles to identify additional sources. Overall, over 50 publications for the period 2005–2019 were identified and reviewed. We systematically reviewed this literature and discussed it during the online conference calls of the Online Peer Assessment PI Forum. We also reviewed and discussed our experiences of examining and experimenting with multiple available OPRA systems, as well as designing and implementing our own systems. In addition, we networked and collaborated with many originators and users of the OPRA systems; we organized and invited originators of well-known OPRA systems to participate in the PI Forum online meetings to demonstrate their applications to interested academics and practitioners. These demos and discussions were documented (as typed notes and video recordings, including screencasts) and shared online for further review and analysis.

Based on these research activities, we compiled a list of 57 systems developed between 1995 and 2015 that conform to our definition of an educational OPRA system (Appendix A; the complete detailed metadata in a publicly shared Google spreadsheet can be accessed at <https://shorturl.at/cjvB4>). Peer-assessment systems prior to 1995 typically were desktop applications limited to the use in local area networks (LANs), which we label “computer-assisted PA systems.” Advances in the CITs, such as the Web (1995), Web 2.0 (2002), and HTML 5 (2008), permitted the creation of truly online and interactive PA applications, which we defined as OPRA systems.

As we explored the identified OPRA systems, the recurring ideas, common patterns, and themes about user-system interactions, functionality, and design choices emerging through these demos and discussions, were coded as *use cases*. *Use case* refers to a given interaction between a user and a system needed to achieve a goal or satisfy a need (Jacobson, 1992). For example, the use case “Provide quantitative peer evaluation” requires a participant to enter a quantitative value for a reviewed submission with the goal of assessing it. We noted use cases common across multiple OPRA systems, as well as some unique use cases, pertinent only to some individual systems. To document objects, relationships, and use cases we identified as *essential* for the OPRA systems, we first created a concept map (see, for example, Fig. 1), and then applied systems analysis techniques to create use-case and class diagrams that were iteratively reviewed and refined by co-authors (see, for example, Babik et al., 2018).

Next, we refined and validated the preliminary list of use cases through an informal focus-group discussion, in which instructors who practice peer assessment, described various user needs, situations, and scenarios that had occurred in their OPRA practice, such as collecting student work and assigning reviewers. In addition, we revisited academic papers describing various OPRA systems to examine how these user needs have been addressed by their designers. We applied concept mapping to visualize discovered use cases and to further group them into categories of *user needs* that the OPRA system must accommodate.

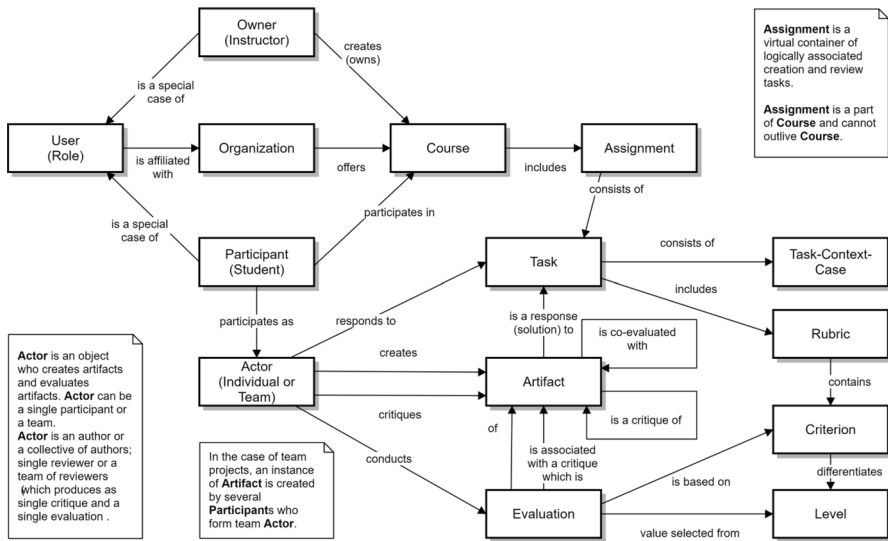


Fig. 1 Example of a concept map of the online peer review and assessment domain

Classification framework and systematic survey of OPRA systems

We constructed a preliminary framework for analysis and classification of OPRA systems by organizing identified, formalized, and categorized user needs, use cases, features, and design options in four layers of abstraction from more specific to more general (Table 1). Categorized user needs and essential use cases (i.e., the use cases that describe only the minimum essential issues necessary to understand the required functionality) form the two *layers of the problem domain* (or *implementation-independent layers*) of the framework, because they are determined by the needs of the users and are independent of any specific implementation or technology. These user needs and use cases apply to any OPRA system; in other words, a system is not an OPRA system unless it accommodates these needs and use cases. A given use case may be implemented in various systems differently, with varying design options. Therefore, functionality features and design options implemented in specific OPRA systems form the two *layers of the solution domain* (or *implementation-dependent layers*). In organizing our framework as hierarchical layers, we follow generally accepted principles of systems analysis and design (Dennis et al., 2015). We focused on features relevant to peer review and assessment and left outside the scope of this study any features pertinent to any learning, content-management, or communication system (e.g., learning-object content management).

Visually, our framework is structured as hierarchically organized layers, where the top, most general, layer defines *user needs*, the second layer includes *use cases*, the third layer contains *features*, and the bottom, the most specific layer consists of specific *design options* (Table 1). (Note that such hierarchically structured frameworks are also used in other domains of information systems; see, for example, the National Institute of Standards and Technology's Framework for Improving Critical Infrastructure Cybersecurity).

To illustrate the use of this framework, consider the following example (Table 1). Every OPRA system must accommodate the user need of "Eliciting evaluations and critiques", i.e., for reviewers to input assessment data (quantitative or qualitative, structured

or semi-structured). This user need creates two essential use cases—“Provide quantitative evaluation” and “Provide qualitative critiques”. To support the former use case, typically two features are required—a rubric (a set of evaluation criteria) and a scale. A rubric may be implemented as two design choices—as either holistic or specific (analytic) rubric. Similarly, the scale feature may be implemented as either cardinal (rating), ordinal (ranking), hybrid (combining rating and ranking), or some other “exotic” scale, such as “dividing a pie.”

We validated our framework by applying the multi-case method (Stake, 2013). We designed a questionnaire combining closed-ended and open-ended items and distributed it via a Google Form to the originators of existing OPRA systems to collect structured and semi-structured data and to verify whether our framework fits their responses. We contacted originators of 23 currently used systems (40% of the total number of identified OPRA systems); 19 responses were received, of which 15 usable responses were selected (a response rate of 65%). The collected data on specific OPRA systems’ functionality and design choices were mapped to our framework and any inconsistencies or divergences in the framework were addressed. In addition, whenever possible, we conducted short teleconference interviews with the originators to obtain additional comments and suggestions. The authors of this paper also contributed data for the systems they originated. The assembled framework was also compared to previously published surveys of peer assessment to ensure commonality of terms and definitions (see the Literature Review section). After detailed analysis of 16 systems, we reached saturation in classification of user needs, use cases, features, and design options.

Our framework allowed us to systematically survey, compare, and analyze the ways in which various user needs are addressed in multiple OPRA systems through implementations of various features and design options. Based on this framework and using data obtained through our multiple case study, we identified six primary categories of user needs accommodated in OPRA systems. The compendium of our classification framework is presented in Table 2.⁵ This framework is not only complete in terms of well-defined layers of abstraction, but also extensible in the sense that additional use cases can be added as OPRA systems evolve over time. However, by no means is our list of features and design options exhaustive. OPRA systems (just as any other type of information systems) continuously evolve—new technologies emerge, enabling new implementations, while some old technologies and implementations become obsolete. Therefore, we consider our framework to be extensible and invite users and designers of OPRA systems to contribute to its evolution.

Results and discussion

Major contributions and findings

Contributions

The major contribution of this study is construction of the classification framework and its application for the systematic review and analysis of existing and emerging

⁵ Detailed reviews and analyses of OPRA user needs, use cases, and system functionality are left outside the scope of this paper due to space constraints and are presented in separate publications.

educational OPRA systems. This helps researchers and practitioners understand the current landscape of technologies supporting and transforming student peer assessment. Importantly, it helps educators, as well as course and system designers, make informed decisions about the available choices of applications to fulfill teaching and learning needs. Analysis based on this framework can help identify the major gaps in existing designs and suggest directions for improving existing and developing new OPRA systems that fit better in the broader and ever-evolving educational technology landscape.

Summary of findings

The diversity of functionalities in existing OPRA systems can satisfy a broad variety of pedagogical and administrative needs. However, this diversity also leads to a multitude of painful tradeoffs between flexibility, comprehensibility, and ease of use. Educational peer-review practices vary greatly across courses, institutions, and countries, so no single OPRA application can comprehensively satisfy all their needs. Thus, every stakeholder involved in design, implementation, and use of such systems should carefully consider the fit between users' needs and a system's affordances and constraints.

In summary, the current landscape of educational OPRA systems can be characterized by the following generalizations:

- (a) User needs are met by a wide variety of solutions;
- (b) Most solutions dictate a peer-review process that is more amenable to certain disciplines, pedagogies, and types of assignments than others; based on underlying design choices, OPRA systems can be generally described as leaning toward either scaffolded or exploratory peer learning;
- (c) Some systems offer a single model of peer-review process while others enable multiple models and may require instructors to assert their own pedagogy in determining the design of the peer-review process. Instructors new to peer assessment may appreciate the former, whereas instructors with greater expertise or established pedagogical practices may prefer the latter;
- (d) There is a broad spectrum of the system maturity (i.e., the degree to which an OPRA system can satisfy diverse and conflicting user needs) as technology advances, new design ideas and new systems constantly emerge;
- (e) Although empirical studies provide some insight into the merits of various solutions, there are still many research opportunities to discover more effective models; in addition, there may be no single most effective model for all user needs;
- (f) No existing solutions address all user needs.

Computerized systems are good at automating very standardized and uniform processes and making them more efficient. The need for variety and flexibility is an enemy of unification, automation, and, consequently, efficiency, but it creates richness, adaptability, and effectiveness. This is the dilemma OPRA system designers have yet to resolve, and our framework and review are meant to aid in this effort. Beyond the benefit of efficiency and scalability, OPRA systems give educators a new kind of leverage, enabling them to develop in their students a certain new kind of cognition and mentality—away from “studying for the test” and toward “building competency for life”. OPRA benefits arise not only from being reviewed and receiving prompt and rich feedback

but also from constantly immersing in the practice of peer-reviewing and assessing at all education levels. The practice of peer-reviewing is not a means to an end of producing more feedback or arriving at a more accurate final score, but rather a learning outcome in itself. When a classroom activity is designed around peer review, students begin approaching their own work differently, less as a series of obstacles, and more as an exploration of possibility. In this way, peer review becomes less of a tool and more of an environment that supports and stimulates exploratory learning.

Our analysis revealed two fundamental factors that determine the diversity of OPRA systems' functionalities and how they may fit in a particular educational context. The first factor can be broadly characterized as *orientation towards scaffolded learning versus exploratory learning*. The second factor can be described as the *degree of system maturity*.

Scaffolded learning versus exploratory learning

With regard to the first factor, generally speaking, all OPRA systems are used to encourage learning and develop higher-level competencies for dealing with complex, open-ended problems. However, the systems oriented toward *scaffolded learning* tend to favor better-defined, more-structured assignments aimed at assessing proficiency in a specific skill set, such as critical writing and programming, or solving a particular type of problem, usually with a preconceived proficient solution. This orientation leans towards using peer assessment to aid summative assessment and seeks to elicit primarily quantitative evaluations that are deemed to "accurately" assess student performance rather than to generate a volume of qualitative critiques. Therefore, systems with this orientation (and empirical research based on them) are concerned with issues of reliability and validity of peer evaluations. To improve reliability and validity, they tend to rely on purposeful allocation, analytic rubrics, rating scales, calibration (based on preconceived sample "correct" answers), as well as instructor grading, feedback, intervention, and censoring of "inaccurate" peer feedback.

In contrast, the OPRA systems with orientation towards *exploratory learning* tend to promote less-defined and less-structured assignments and projects with no expected "ideal proficient" solution. Such systems aim at exposing learners to the ambiguity and uncertainty of the problem, stimulating their exploration and holistic understanding of the problem domain, encouraging social-learning interactions and discussions among peers. Therefore, OPRA systems based on this orientation typically use random allocation, focus on qualitative critiques, holistic rubrics, and ranking scales. While providing capabilities for quantitative evaluations that may enable "peer grading" if desired, these systems focus on providing learning analytics highlighting weaknesses and gaps in students' shared understanding of a problem and highlighting emerging differences in opinions, perceptions, and approaches. These systems tend to lack features for "improving the accuracy" of quantitative evaluations and curation of peer critiques, but instead emphasize features that encourage self-curated and self-regulated social learning through imitation of successful examples (Bandura, 1986; Coleman et al., 1957; Rogers, 2005), proximal development (Vygotsky, 1980), critical dialogue, cultural consensus, and intersubjectivity (Matusov, 1996; Matusov & Marjanovic-Shane, 2017). Skeptics of this orientation often cite the "blind leading the blind" adage, suggesting that without a certain basic level of disciplinary knowledge and competency, participants' ability to self-regulate and self-curate cannot be trusted. The proponents of this orientation argue that although features, such as calibration and analytic rubrics, reduce evaluation inconsistency (or "inaccuracy") and foster authors' confidence

in the reviewers' competence, when overused, they discourage divergent thinking, intellectual exploration, experimentation with provocative ideas, student agency and creativity, while encouraging authors' and reviewers' conformity and the quest for the "right answer" and "better grade".

While we do not suggest that any OPRA system strictly adheres to any one of these orientations, the combination of features in systems, such as CPR, Peerceptiv, Expertiza, and SPARKPlus (Willey & Gardner, 2010), indicates their originators' orientation towards *scaffolded learning*, while the design of the systems, such as CritViz and Mobius SLIP, appears to be oriented towards *exploratory learning*.

Our investigation shows that online peer review and assessment, as any other assessment process, is inherently value-laden and intersubjective. The means by which OPRA is conducted, and therefore, the design of the system to support it, inevitably reflect the pedagogical values of the system designer. Moreover, the choice of the system should be primarily driven by the values and aims of the instructor. While we maintain that both system orientations are grounded in constructivist learning theories, they do so differently, in accordance with different epistemological orientations regarding knowledge. This warrants further examination of their ontological significance. A deep explanation of constructivist learning theories in the OPRA context is beyond the scope of this paper, however, we posit that both system orientations lean towards constructivism that proffers "all cognitive activity takes place within the experiential world of a goal-directed consciousness" (Von Glasersfeld, 1984, p. 10). Put another way, constructivism assumes that cognition organizes its experiential world by organizing itself. Consistent with constructivist pedagogy, both OPRA system orientations avail themselves as a *more knowledgeable other (MKO)*, albeit a non-human one, suggesting a pedagogical model that emphasizes the gradual release of responsibility between the MKO and a learner.

In the case of the *scaffolded learning orientation*, *scaffolding* means that, while the "true" or "correct" problem solution is socially and culturally constructed and interpreted, it can ultimately be known. An OPRA system then enacts a type of epistemological determinism, in which the submissions and reviews are assessed against the "correct" solution determined by the educator through tools such as calibration, analytical rubrics, and quantitative summative assessment. Subsequently, while students engage in the "peer" component of learning, they do so in accordance with the knowledge claims of said instructor. Thus, what is valued as knowledge becomes strikingly visible within the system itself. By limiting what is possible or acceptable, the *scaffolded-learning-oriented* OPRA defines knowledge by what it is not, by simply ignoring or discouraging solutions and critiques which do not meet the given criteria. Such systems align with traditional assessment tools, such as tests, as they determine what knowledge is worth knowing.

The OPRA systems with *exploratory learning orientation*, conversely, lean towards an epistemological stance that knowledge is in a state of constant flux, generated and evolving through particular intersubjective interactions among learners and between learners and instructors in the peer-review process. One might align such an orientation with *radical constructivism* (Von Glasersfeld, 1984, 1995) and, in doing so, consider that "we can check our perceptions only by means of other perceptions" (Von Glasersfeld, 1984, p. 6). While it may be true that some might be troubled by the potential for a relativist pedagogy to emerge out of such an open and exploratory approach to knowledge, it also suggests an alternative view in which troubling a representational view of reality might effectuate a "search for fitting ways of behaving and thinking" (Von Glasersfeld, 1984, p. 14). This interpretation suggests that instructors skeptical of prescriptive scaffolding practices, like

grading rubrics or calibration exercises, may find benefit in OPRA as a process scaffold for socially constructed solutions to ill-structured problems.

System maturity

With regard to the second factor, we found that the *degree of system maturity* reflects diversity and flexibility of an OPRA system's functionality (i.e., its ability to satisfy diverse and conflicting user needs), and generally correlates with the age of the system. Older and more mature systems (e.g., Peerceptiv, Expertiza, peerScholar, SPARKPlus) tend to have a greater variety of well-tested features, supported by extensive experimental research, and they cater to a broader institutional audience. Less mature systems (e.g., CritViz, Eli Review, eMarking, Mechanical TA, Moodle's Workshop) usually focus on a specific pedagogical approach to peer review and assessment, with somewhat restrictive workflow, fewer options, and tend to provide their services to niche users. Importantly, the relation between system maturity and feature diversity is not perfectly deterministic. Systems such as SPARKPlus, Peerceptiv, and PeerWise are fairly old and have a wide variety of features; in contrast, while CPR could be considered one of the most mature and widely used OPRA systems, it lacks in diversity and flexibility of features in comparison to many younger systems. Also, in this age, systems tend to evolve very quickly, thanks to agile development practices. Therefore, by the time this paper is published, it is very likely that newer systems listed here will reach a higher degree of maturity. Equally likely, however, is that development of some of them may be discontinued.

Future research opportunities and system design recommendations

Our analysis also serves to highlight exciting future research and system design opportunities. Peer review and assessment have been extensively studied for several decades, but research on how technology-enabled processes affect pedagogical and administrative outcomes remains an emerging stream. One such opportunity is investigating the effects of participant anonymity on the OPRA process and outcomes. Anonymity has been generally considered a good remedy against various social and personal biases in evaluating and critiquing, as well as adverse reviewers' behaviors driven by these biases (e.g., retaliation against negatively toned critiques or favorable evaluation given to friends' submissions). Therefore, double-blind or single-blind reviews are prevalent modes in practically all OPRA systems. At the same time, appropriate uses of identity disambiguation (i.e., revealing authors' or reviewers' true identities) may be used as motivating factors for both authors and reviewers (Lin, 2018; Lu, 2011; Lu & Bol, 2007; Yu & Sung, 2016; Yu & Wu, 2011). Moreover, anonymity has been interpreted very narrowly, usually as authors' and reviewers' knowledge of each other's identity at the time of review. However, we found it to be a multifaceted aspect of peer review. For example, anonymity also specifies whether identities remain hidden or revealed *after* the peer-review process is complete; and whether identities are revealed only to the author and reviewers of a particular artifact or to the larger pool of other participants. Ambiguity could be eschewed in multiple ways and could have a variety of effects. Research questions about the effects of these various aspects of anonymity and privacy in OPRA should be explored in the future.

Future research may also explore the effects of various aspects of allocation, peer-review workflow, and quantitative evaluations on various variables measuring the peer-review process and outcomes. One particularly important issue is how to motivate reviewers to

provide deeper, richer, and more professional feedback. It would also be interesting to review and compare algorithms for computing attainment, accuracy, reputation, and other metrics used in different OPRA systems.

As the OPRA systems evolve to include various web-based tools for *data visualization*, it is worthwhile to investigate the effects of visualizations on participants' behavior. New types of digital visual representations of peer-review processes and outcomes can better depict cognitive work, relationships, trends, and activities of the learners. Furthermore, these representations can be interactive and real-time, thus influencing individual and group learning behavior dynamics (Babik et al., 2017a, 2017b).

Another emerging opportunity is incorporating transferable *micro-credentials* ("digital badges") in the peer-review process. Micro-credentialing is a new trend in education technology that offers certain transformative changes in education (Abramovich et al., 2013; Carey, 2012; Casilli & Hickey, 2016). There is a synergy between OPRA and micro-credentialing: OPRA makes micro-credentialing more scalable, trustworthy, and versatile, whereas micro-credentialing allows peer-assessed learning outcomes to be conveyed beyond a single course. Credentials from peer assessment can be shared with anyone, providing credible documentation of a student's skills, learning experiences, and achievements to interested third parties, such as prospective employers. For these reasons, integrating OPRA and micro-credentialing is a logical next step for both of these technologies. Our study, however, found that only a couple of systems (Mechanical TA and PeerWise) provide digital badges, and these are used only internally and are not transferable across multiple platforms. A couple of other systems claimed to have micro-credentialing features in development (e.g., Expertiza and Mobius SLIP). It would be exciting to explore the effects of integrating micro-credentialing and OPRA on individual and group learning, as well as on institutional competitiveness.

Our study would be of little value if it did not present any practical recommendations for future OPRA design. While the detailed analyses and justifications had to be left out due to space constraints (and are presented in related papers), we would like to give a sampler of the following observations and recommendations. *Dynamic allocation* is more flexible than *en-masse allocation* and *reallocation*, but the implications require careful study, as they may lead to undesired side-effects, such as procrastination. *Dispersed unidirectional allocation* allows for a more balanced workload (even number of submissions) per reviewer and, therefore, should be preferred as a default setup (unless there are other mechanisms to compensate for extra work, such as extra credit). When advantages of *clustered reciprocal allocation* are important (e.g., when giving peer-review-group-specific variations of assignments), this allocation may be created as a special case of dispersed unidirectional allocation. The desirable number of submissions per reviewer is between three and five; the preferred number of reviews per submission is five to six. It is desirable to have an interface for implementing flexible analytic rubrics, with the ability of creating holistic rubric as a special case. Combining ranking and rating evaluation scales in a single activity with a single interface control (as implemented in Mobius SLIP with the SLIP Slider) offers an interesting opportunity for deeper learning analytics than using solely rating or ranking data. Rejoinders (back evaluations) appears to be a popular approach for holding reviewers accountable for review value, and we recommend using it, with the single-loop reduced workflow as a special-case option.

Limitations

Our analysis has the following limitations. First, our assessment of OPRA systems' capabilities is accurate only as of the time of data collection. These systems are constantly being developed and upgraded with new features based on the originators' vision and users' functional and non-functional requirements. As of time of publication, some systems may have been decommissioned while some new systems may have emerged. The software market is very dynamic, and the fact that most OPRA systems are provided as software-as-a-service (SaaS) over the internet makes provision of new features even more agile. Thus, another survey, preferably based on the proposed framework, may be due in a few years.

Second, our survey that provided data for framework validation and illustrative examples, due to resource constraints, covered only a subpopulation of OPRA systems. Therefore, some innovative use cases and features may have been unintentionally omitted. In addition, the OPRA systems designed specifically to assess observed behaviors rather than artifacts (e.g., a contribution to a team project) are underrepresented in our sample. Engaging more closely with the system originators and collecting richer information through system demonstrations may help update this survey in the future.

Perhaps the most significant limitation of the proposed framework is that it imposes a hierarchical one-to-many relationship between use cases and system features. In other words, it treats every use case as being addressed by one or several features, but every feature as addressing one specific use case. We found that, while typically several features are implemented as an ensemble to address a particular user need through a specific use case, oftentimes one feature serves more than one use case. For example, double-loop peer assessment with rejoinders can be treated as both a means to motivate participants to do a better job as reviewers and a way to assess attainment of critiques. In addition, some features serving the same use case, that we consider to be distinct features, may, in the view of some users and readers, be indistinguishable or considered to be a single feature. For example, in our discussions, some instructors treated scales and rubrics as the same feature. Thus, while mapping user needs, use cases, features, and design choices is always based on certain assumptions and simplifications made by the system analyst, this limitation offers an opportunity for further refinement of the framework to incorporate this cross-functional aspect. Importantly, the proposed research framework provides a foundation for exploring a dynamic socio-technological phenomenon. We invite other researchers to apply, update, and augment our framework on changing technologies and practices.

Conclusion

Online peer review and assessment enhances conventional and virtual classrooms by offering scalable and efficient “grading” and large volumes of prompt feedback. However, its greatest advantage is that it transforms a classroom into a self-curating, self-regulating learning environment as education undergoes technology-driven transformation. OPRA systems might be the first step toward new types of “instructorless classrooms”, such as the “42” computer-programming training program (42.us.org) (think “driverless cars”), but they require informed and careful development, implementation, and execution to maximize advantages and remedy pitfalls (think “driverless cars” again) (Beach, 1974; Fu et al., 2019; Morrison, 2014).

Charting the abstruse landscape of OPRA technologies to enable this type of pedagogy is the primary objective of this study. We sought not only to provide a structured and comprehensive overview of the state of things, but also attempted to offer a framework for ongoing analyzing the ever-changing landscape. Our framework and systematic survey can be expected to inform audiences of decision-makers involved in designing, researching, promoting, adopting, and applying online peer review and assessment. This paper contributes to the emerging stream of literature promoting educational design research (Akker et al., 2006, 2012; Søndergaard & Mulder, 2012). This work may also have implications for other domains relying on peer review, such as academic publishing and grant application processing.

Funding The study reported in this manuscript was supported by the National Science Foundation (Directorate for Education and Human Services) under grants DUE-1432347, 1431856, 1432580, 1432690, and 1431975.

Declarations

Conflict of interest There is no potential conflicts of interest to disclose related to this study.

Research involving human and animal rights No human participants and/or animals were involved in this study. The unit of analysis is a technical system. Survey administered to collect data did not include any information on human subjects.

Informed consent No informed consent related to collect data about human subjects was used in this study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abramovich, S., Schunn, C., & Higashi, R. M. (2013). Are badges useful in education?: It depends upon the type of badge and expertise of learner. *Educational Technology Research and Development*, 61(2), 217–232. <https://doi.org/10.1007/s11423-013-9289-2>
- Alqassab, M., Strijbos, J.-W., Panadero, E., Ruiz, J. F., Warrens, M., & To, J. (2023). A systematic review of peer assessment design elements. *Educational Psychology Review*, 35(1), 18. <https://doi.org/10.1007/s10648-023-09723-7>
- Babik, D., Iyer, L., & Ford, E. (2012). Towards a comprehensive online peer assessment system: Design outline. *Lect. Notes Comput. Sci. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7286 LNCS, pp. 1–8
- Babik, D. (2015). *Investigating intersubjectivity in peer-review-based, technology-enabled knowledge creation and refinement social systems*. The University of North Carolina at Greensboro.
- Babik, D., Gehringer, E. F., Tinapple, D., Pramudianto, F., & Song, Y. (2018). Domain model and meta-language for peer review and assessment. *Proceedings of Western DS, I*, 7.
- Babik, D., Singh, R., Zhao, X., & Ford, E. (2017a). What you think and what I think: Studying intersubjectivity in knowledge artifacts evaluation. *Information Systems Frontiers*, 19(1), 31–56. <https://doi.org/10.1007/s10796-015-9586-x>

- Babik, D., Tinapple, D., Gehringer, E. F., & Pramudianto, F. (2017b). The effect of visualization on students' miscalibration in the context of online peer assessment. *Proceedings of Western DS, I*, 7.
- Baikadi, A., Schunn, C. D., & Ashley, K. D. (2016). *Impact of revision planning on peer-reviewed writing*. Educational Data Mining.
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory* (1st ed.). Prentice Hall.
- Beach, L. R. (1974). Self-directed student groups and college learning. *Higher Education*, 3(2), 187–200. <https://doi.org/10.1007/BF00143791>
- Bostock, S. (2000). *Student peer assessment*. Keele University.
- Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, 18(5), 529–549. <https://doi.org/10.1007/BF00138746>
- Bouzidi, L., & Jailliet, A. (2009). Can online peer assessment be trusted? *Educational Technology & Society*, 12(4), 257–268.
- Bull, S., & McCalla, G. (2002). Modelling cognitive style in a peer help network. *Instructional Science*, 30(6), 497–528. <https://doi.org/10.1023/A:1020570928993>
- Carey, K. (2012). A future full of badges. *The Chronicle of Higher Education*, A60. <https://www.chronicle.com/article/A-Future-Full-of-Badges/131455>
- Carlson, P., & Smith, R. (2017). Computer-mediated peer review: A comparison of calibrated peer review and Moodle's workshop. *Faculty Publications—English & Literature*. <https://peer.asee.org/28064>
- Casilli, C., & Hickey, D. (2016). Transcending conventional credentialing and assessment paradigms with information-rich digital badges. *The Information Society*, 32(2), 117–129. <https://doi.org/10.1080/01972243.2016.1130500>
- Chang, C.-Y., Lee, D.-C., Tang, K.-Y., & Hwang, G.-J. (2021). Effect sizes and research directions of peer assessments: From an integrated perspective of meta-analysis and co-citation network. *Computers & Education*, 164, 104123. <https://doi.org/10.1016/j.compedu.2020.104123>
- Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education*, 48(3), 409–426. <https://doi.org/10.1016/j.compedu.2005.02.004>
- Coleman, J., Katz, E., & Menzel, H. (1957). The diffusion of an innovation among physicians. *Sociometry*, 20(4), 253–270. <https://doi.org/10.2307/2785979>
- Davies, P. (2000). Computerized peer assessment. *Innovations in Education and Teaching International*, 37(4), 346–355.
- de Alfaro, L., & Shavlovsky, M. (2014). CrowdGrader: A tool for crowdsourcing the evaluation of homework assignments. *Proceedings of the 45th ACM Technical Symposium on Computer Science Education*. <https://doi.org/10.1145/2538862.2538900>
- Dennis, A., Wixom, B. H., & Tegarden, D. (2015). *Systems analysis and design: An object-oriented approach with UML* (5th ed.). John Wiley & Sons.
- Denzin, N. K., & Lincoln, Y. S. (2011). *The SAGE handbook of qualitative research* (4th ed.). SAGE Publications, Inc.
- Doiron, G. (2003). The value of online student peer review, evaluation and feedback in higher education. *CDTL Brief*, 6, 1–2.
- Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review*, 32(2), 481–509. <https://doi.org/10.1007/s10648-019-09510-3>
- Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research*, 59(4), 395–430. <https://doi.org/10.3102/00346543059004395>
- Fu, Q.-K., Lin, C.-J., & Hwang, G.-J. (2019). Research trends and applications of technology-supported peer assessment: A review of selected journal publications from 2007 to 2016. *Journal of Computers in Education*, 6(2), 191–213. <https://doi.org/10.1007/s40692-019-00131-x>
- Gehringer, E. F. (2014). A survey of methods for improving review quality. *New horizons in web based learning* (pp. 92–97). Springer.
- Gehringer, E. F. (2019). Board 60: PeerLogic: Web services for peer assessment. In *2019 ASEE annual conference & exposition*.
- Gehringer, E. F., Ehresman, L., Conger, S. G., & Wagle, P. (2007). Reusable learning objects through peer review: The Expertiza approach. *Innovate: Journal of Online Education*, 3(5), 4.
- Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation in Higher Education*, 36(2), 137–155. <https://doi.org/10.1080/02602930903221444>
- Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, 57(4), 2333–2351. <https://doi.org/10.1016/j.compedu.2011.06.004>

- Goldin, I. (2011). *A focus on content: The use of rubrics in peer review to guide students and instructors*. <http://d-scholarship.pitt.edu/8375/1/goldin%2Ddissertation%2D20110805.pdf>
- Goldin, I., Ashley, K. D., & Schunn, C. (2012). Redesigning educational peer review interactions using computer tools. *Journal of Writing Research*, 4(2), 111–119.
- Hamer, J. (2006). Some experiences with the “contributing student approach.” *SIGCSE Bulletin*, 38(3), 68–72. <https://doi.org/10.1145/1140123.1140145>
- Hamer, J., Kell, C., & Spence, F. (2007). Peer assessment using Aropä. *Proceedings of the Ninth Australasian Conference on Computing Education*, 66, 43–54.
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, 28(1), 75–105.
- Jacobson, I. (1992). *Object oriented software engineering: A use case driven approach* (1st ed.). Addison-Wesley Professional.
- Joordens, S., Desa, S., & Paré, D. (2009). The pedagogical anatomy of peer assessment: Dissecting a peerscholar assignment. *Journal of Systemics, Cybernetics & Informatics*, 7(5), 1.
- Kulkarni, C., Wei, K. P., Le, H., Chia, D., Papadopoulos, K., Cheng, J., Koller, D., & Klemmer, S. R. (2013). Peer and self assessment in massive online classes. *ACM Transactions on Computer-Human Interaction*, 20(6), 1–33. <https://doi.org/10.1145/2505057>
- Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniuw, R. (2020). Does peer assessment promote student learning? A meta-analysis. *Assessment & Evaluation in Higher Education*, 45(2), 193–211. <https://doi.org/10.1080/02602938.2019.1620679>
- 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*, 116, 81–92. <https://doi.org/10.1016/j.compedu.2017.08.010>
- Lu, R. (2011). *Anonymity in collaboration: Anonymous vs. identifiable E-peer review in writing instruction*. Trafford Publishing.
- Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable E-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning*, 6(2), 100–115.
- Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education*, 19(4), 209–232. <https://doi.org/10.1080/08993400903384844>
- Martin, P. Y., & Turner, B. A. (1986). Grounded theory and organizational research. *The Journal of Applied Behavioral Science*, 22(2), 141–157. <https://doi.org/10.1177/002188638602200207>
- Matusov, E. (1996). Intersubjectivity without agreement. *Mind, Culture, and Activity*, 3(1), 25–45. https://doi.org/10.1207/s15327884mca0301_4
- Matusov, E., & Marjanovic-Shane, A. (2017). Many faces of the concept of culture (and education). *Culture & Psychology*, 23(3), 309–336. <https://doi.org/10.1177/1354067X166655460>
- Millard, D., Fill, K., Gilbert, L., Howard, Y., Sinclair, P., Senbanjo, D. O., & Wills, G. B. (2007). Towards a canonical view of peer assessment. *Seventh IEEE international conference on advanced learning technologies (ICALT 2007)*, pp. 793–797. <https://doi.org/10.1109/ICALT.2007.260>
- Millard, D., Newman, D., & Sinclair, P. (2008). *PeerPigeon: A web application to support generalised peer review*. pp. 3824–3836. <https://www.learntechlib.org/primary/p/30219/>
- Misiejuk, K., & Wasson, B. (2021). Backward evaluation in peer assessment: A scoping review. *Computers & Education*, 175, 104319. <https://doi.org/10.1016/j.compedu.2021.104319>
- Morrison, N. (2014). *The teacher-less classroom is not as close as you think*. Forbes. <https://www.forbes.com/sites/nickmorrison/2014/08/21/the-teacher-less-classroom-is-not-as-close-as-you-think/>
- Patchan, M. M., Schunn, C. D., & Clark, R. J. (2017). Accountability in peer assessment: Examining the effects of reviewing grades on peer ratings and peer feedback. *Studies in higher education*, pp. 2263–2278. <https://doi.org/10.1080/03075079.2017.1320374>
- Pramudianto, F., Aljeshi, M., Alhoussein, H., Song, Y., Gehringer, E. F., Babik, D., & Tinapple, D. (2016). Peer review data warehouse: Insights from different systems. *CSPRED 2016: Workshop on computer-supported peer review in education*. Educational Data Mining
- Purchase, H., & Hamer, J. (2017). *Peer review in practice: Eight years of experience with Aropä*. University of Glasgow. <http://www.dcs.gla.ac.uk/~hcp/aropa/AropaReportJan2017.pdf>
- Raman, K., & Joachims, T. (2014). *Methods for ordinal peer grading*. pp. 1037–1046. <https://doi.org/10.1145/2623330.2623654>
- Rogers, E. M. (2005). Complex adaptive systems and the diffusion of innovations. *The Innovation Journal: The Public Sector Innovation Journal*, 10, 25.
- Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students’ perceptions about peer

- assessment skills. *European Journal of Psychology of Education*, 33(1), 75–99. <https://doi.org/10.1007/s10212-017-0339-8>
- Russell, A. A. (2001). Calibrated peer review: a writing and critical-thinking instructional tool. *UCLA, Chemistry, 2001*. http://www.unc.edu/opt-ed/eval/bp_stem_ed/russell.pdf
- 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
- Sargeant, J., Mann, K., van der Vleuten, C., & Metsemakers, J. (2008). “Directed” self-assessment: Practice and feedback within a social context. *Journal of Continuing Education in the Health Professions*, 28(1), 47–54. <https://doi.org/10.1002/chp.155>
- Shah, N. B., Bradley, J. K., Parekh, A., Wainwright, M., & Ramchandran, K. (2013). A case for ordinal peer evaluation in MOOCs. *NIPS workshop on data driven education*, pp. 1–8.
- Sitthiworachart, J., & Joy, M. (2004). Effective peer assessment for learning computer programming. *SIGCSE BULLETIN*, 36, 122–126.
- Søndergaard, H., & Mulder, R. A. (2012). Collaborative learning through formative peer review: Pedagogy, programs and potential. *Computer Science Education*, 22(4), 343–367. <https://doi.org/10.1080/08993408.2012.728041>
- Song, Y., Pramudianto, F., & Gehringer, E. F. (2016). A markup language for building a data warehouse for educational peer-assessment research. *IEEE Frontiers in Education Conference (FIE)*, 2016, 1–5. <https://doi.org/10.1109/FIE.2016.7757600>
- Stake, R. E. (2013). *Multiple case study analysis*. Guilford Press.
- Steffens, K. (2006). Self-regulated learning in technology-enhanced learning environments: Lessons of a European peer review. *European Journal of Education*, 41(3–4), 353–379.
- Strauss, A., & Corbin, J. (1994). Grounded theory methodology: An overview. In *Handbook of qualitative research*
- Taraborelli, D. (2008). Soft peer review. Social software and distributed scientific evaluation. *Proceedings of the 8th international conference on the design of cooperative systems, Carry-Le-Rouet, 20–23 May 2008*, pp. 99–110
- Tenório, T., Bittencourt, I. I., Isotani, S., & Silva, A. P. (2016). Does peer assessment in on-line learning environments work? A systematic review of the literature. *Computers in Human Behavior*, 64, 94–107. <https://doi.org/10.1016/j.chb.2016.06.020>
- Tinapple, D., Olson, L., & Sadauskas, J. (2013). CritViz: Web-based software supporting peer critique in large creative classrooms. *Bulletin of the IEEE Technical Committee on Learning Technology*, 15(1), 29.
- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3), 249–276. <https://doi.org/10.3102/00346543068003249>
- Topping, K. J. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising new modes of assessment: In search of qualities and standards* (pp. 55–87). Springer.
- Topping, K. J. (2005). Trends in peer learning. *Educational Psychology*, 25(6), 631–645. <https://doi.org/10.1080/01443410500345172>
- Topping, K. J. (2023). Digital peer assessment in school teacher education and development: A systematic review. *Research Papers in Education*, 38(3), 472–498. <https://doi.org/10.1080/02671522.2021.1961301>
- van den Akker, J., Branch, R. M., Gustafson, K., Nieveen, N., & Plomp, T. (2012). *Design approaches and tools in education and training*. Springer.
- van den Akker, J., Gravemeijer, K., McKenney, S., & Nieveen, N. (2006). *Educational design research*. Routledge.
- van den Berg, I., Admiraal, W., & Pilot, A. (2006). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education*, 31(1), 19–36. <https://doi.org/10.1080/02602930500262346>
- Verma, P. (2015). *5 Tech trends that will transform education by 2025*. Forbes. <https://www.forbes.com/sites/centurylink/2015/08/11/5-tech-trends-that-will-transform-education-by-2025/#3e2910b75890>
- Von Glasersfeld, E. (1984). An introduction to radical constructivism. *The invented reality 1740* (pp. 17–40). Norton.
- Von Glasersfeld, E. (1995). *Radical constructivism*. Routledge.
- Vygotsky, L. S. (1980). Mind in society: The development of higher psychological processes. *Journal of Reading Behavior*, 12, 161–162.
- Wahid, U., Chatti, M. A., & Schroeder, U. (2016). A systematic analysis of peer assessment in the MOOC era and future perspectives. *eLmL*, 75, 6.
- Wiley, K., & Gardner, A. (2009). Improving self- and peer assessment processes with technology. *Campus-Wide Information Systems*, 26(5), 379–399. <https://doi.org/10.1108/10650740911004804>

- Wiley, K., & Gardner, A. (2010). Investigating the capacity of self and peer assessment activities to engage students and promote learning. *European Journal of Engineering Education*, 35(4), 429–443. <https://doi.org/10.1080/03043797.2010.490577>
- Wolfe, W. J. (2004). Online student peer reviews. *Proceedings of the 5th conference on information technology education*, pp. 33–37. <https://doi.org/10.1145/1029533.1029543>
- Wooley, R., Was, C., Schunn, C. D., & Dalton, D. (2008). The effects of feedback elaboration on the giver of feedback. *Annual Meeting of the Cognitive Science Society*, 5, 2375–2380.
- Wright, J. R., Thornton, C., & Leyton-Brown, K. (2015). Mechanical TA: partially automated high-stakes peer grading. *Proceedings of the 46th ACM technical symposium on computer science education*, pp. 96–101. <https://doi.org/10.1145/2676723.2677278>
- Wu, C., Chanda, E., & Willison, J. (2010). *SPARKPLUS for self- and peer assessment on group-based honours' research projects*. <https://digital.library.adelaide.edu.au/dspace/handle/2440/61612>
- Yu, F.-Y., & Sung, S. (2016). A mixed methods approach to the assessor's targeting behavior during online peer assessment: Effects of anonymity and underlying reasons. *Interactive Learning Environments*, 24(7), 1674–1691. <https://doi.org/10.1080/10494820.2015.1041405>
- Yu, F.-Y., & Wu, C.-P. (2011). Different identity revelation modes in an online peer-assessment learning environment: Effects on perceptions toward assessors, classroom climate and learning activities. *Computers & Education*, 57(3), 2167–2177. <https://doi.org/10.1016/j.compedu.2011.05.012>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Dmytro Babik received his Ph.D. degree in information systems from the University of North Carolina at Greensboro in 2015, his MBA degree from Tulane University, New Orleans, LA, in 2004, and MA degree in Economics from CERGE-EI, Charles University, Prague, Czechia, in 2002. He is an Associate Professor of Computer Information Systems at James Madison University, Harrisonburg, VA. Dr. Babik is the originator of Mobius SLIP.

Edward Gehringer received his Ph.D. degree in Computer Science from Purdue University, West Lafayette, IN in 1979. He was a Research Associate and Lecturer at Carnegie Mellon University, and a Fulbright Fellow at Monash University in Australia. Since 1984, he has been a faculty member at North Carolina State University, currently a Professor of Computer Science and ECE. Dr. Gehringer is the originator of Expertiza.


Jennifer Kidd received her Ph.D. in urban services/education/curriculum and instruction in 2006 and her MS in education in 1999 from Old Dominion University, Norfolk, VA. Since 2007, she has been a full-time Lecturer in the Department of Teaching and Learning at Old Dominion University.

Kristine Sunday received the Ph.D. degree in art education from Pennsylvania State University, University Park, PA, in 2011. During 2014–2022, she was an Assistant Professor of Teaching and Learning at Old Dominion University, Norfolk, VA.

David Tinapple received his MFA degree from Carnegie Mellon University in 2007 and BFA degree from the Ohio State University in 2003. He is an Associate Professor at the School of Arts, Media and Engineering, Herberger Institute for Design at Arizona State University. Mr. Tinapple is the originator of CritViz.

Steven Gilbert received an AB in Mathematics from Princeton University, an EdM from Harvard Graduate School of Education, and an MBA from the Wharton School of the University of Pennsylvania. In 1998, he founded the Teaching, Learning, and Technology (TLT) Group, an independent nonprofit organization.

Authors and Affiliations

Dmytro Babik¹  · Edward Gehringer² · Jennifer Kidd³ · Kristine Sunday³ · David Tinapple⁴ · Steven Gilbert⁵

✉ Dmytro Babik
babikdx@jmu.edu

Edward Gehringer
efg@ncsu.edu

Jennifer Kidd
jkidd@odu.edu

Kristine Sunday
ksunday@odu.edu

David Tinapple
david.tinapple@asu.edu

- ¹ Department of Computer Information Systems and Business Analytics, James Madison University, 421 Bluestone Dr, Harrisonburg, VA 22807, USA
- ² Department of Computer Science, North Carolina State University, 890 Oval Drive, Raleigh, NC 27695, USA
- ³ College of Education, Old Dominion University, 1 Old Dominion University, Norfolk, VA 23529, USA
- ⁴ Herberger Institute for Design and the Arts School of Art, Media, and Engineering, Arizona State University, 1151 S Forest Ave, Tempe, AZ 85281, USA
- ⁵ Late President & Founder, The TLT Group, Takoma Park, MD, USA