Educational Psychology Review (2023) 35:18
https://doi.org/10.1007/s10648-023-09723-7

REVIEW ARTICLE

A Systematic Review of Peer Assessment Design Elements

Maryam Alqassab[1,2] · Jan-Willem Strijbos[1,3] · Ernesto Panadero[4,5] · Javier Fernández Ruiz[6,7] · Matthijs Warrens[3] · Jessica To[8]

Accepted: 1 September 2022 / Published online: 9 February 2023

Abstract

The growing number of peer assessment studies in the last decades created diverse design options for researchers and teachers to implement peer assessment. However, it is still unknown if there are more commonly used peer assessment formats and design elements that could be considered when designing peer assessment activities in educational contexts. This systematic review aims to determine the diversity of peer assessment designs and practices in research studies. A literature search was performed in the electronic databases PsycINFO, PsycARTICLES, Web of Science Core Collection, Medline, ERIC, Academic Search Premier, and EconLit. Using data from 449 research studies (derived from 424 peer-reviewed articles), design differences were investigated for subject domains, assessment purposes, objects, outcomes, and moderators/mediators. Arts and humanities was the most frequent subject domain in the reviewed studies, and two-third of the studies had a formative purpose of assessment. The most used object of assessment was written assessment, and beliefs and perceptions were the most investigated outcomes. Gender topped the list of the investigated moderators/mediators of peer assessment. Latent class analysis of 27 peer assessment design elements revealed a five-class solution reflecting latent patterns that best describe the variability in peer assessment designs (i.e. prototypical peer assessment designs). Only ten design elements significantly contributed to these patterns with an associated effect size R^2 ranging from .204 to .880, indicating that peer assessment designs in research studies are not as diverse as they theoretically can be.

Keywords Peer assessment · Instructional design · Peer assessment diversity · Systematic review

With the rise in interactive learning practices at all levels of education, student involvement in assessment expanded in parallel. Empirical evidence supports the

✉ Maryam Alqassab
Maryam.alqassab@ulpgc.es

Extended author information available on the last page of the article

positive impact of peer assessment on performance regardless of the variations in some peer assessment design elements (Double et al., 2020; Li et al., 2020). Topping's (1998) semi-systematic narrative literature review of peer assessment in higher education constitutes a pivotal contribution to the increase in practices that involve students in assessment. Testament to its increased popularity is the accumulation of research studies on peer assessment and most notably more than thirty review studies covering a wide variety of topics. Among the topics investigated by these reviews are as follows: (i) the design and implementation of peer assessment practices (Adachi et al., 2018; Gielen et al., 2011; Luxton-Reilly, 2009; Topping, 1998, 2003, 2013, 2017, 2021a, b; Strijbos et al., 2009; Van den Berg et al., 2006a, b; Van Gennip et al., 2009); (ii) reliability and validity of peer assessment (Falchikov & Goldfinch, 2000; Li et al., 2016; Speyer et al., 2011; Topping, 1998, 2003, 2013, 2017); (iii) quality criteria for peer assessment practices (Ploegh et al., 2010; Tillema et al., 2011); (iv) impact of social and interpersonal processes (Panadero, 2016; Panadero & Alqassab, 2019; Panadero et al., 2018; Strijbos et al., 2009; Topping, 2017; Van Gennip et al., 2009); (v) peer assessment of collaborative learning (Lejk et al., 1996; Dijkstra et al., 2016; Forsell et al., 2020; Meijer et al., 2020; Strijbos, 2016; Strijbos et al., 2017); and (vi) how the instructional conditions relate to peer assessment outcomes (Ashenafi, 2017; Double et al., 2020; Hoogeveen & Van Gelderen, 2013; Huisman et al., 2019; Li et al., 2020; Panadero et al., 2018; Sanchez et al., 2017; Sluijsmans et al., 1999; Topping, 1998, 2003, 2013, 2021a, b; Van Popta et al., 2017; Van Zundert et al., 2010).

Topping (1998) initially defined peer assessment 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 status' (p. 250). However, the expansion of contextual, behavioural, and instructional design considerations in both research and implementation of peer assessment—as well as its close connection to interactive learning practices (e.g. collaborative learning; de Hei et al., 2016; Strijbos, 2016)—indicate the need for a more aligned definition. Hence, we define peer assessment as:

> a learning phenomenon where individuals or social constellations (e.g., pair, group, team, community)—within a physical and/or virtual environment—exchange, react to, interact and/or act upon information about individual performance and/or individual contributions to the process and/or product of a social constellation, with the purpose to accomplish implicit or explicit shared and individual learning goals (e.g., domain-specific knowledge or skills, social skills, etc.).

Peer assessment can be structured by instructional scaffolds (which can be faded if no longer needed). The instructional scaffolds are provided by an agent(s) within or outside of the peer assessment process (e.g. teacher, peer, self, technology) to guide and increase the likelihood that individuals and/or social constellations can accomplish their goals in line with criteria and standards established by an agent(s) within or outside of the peer assessment process.

Peer Assessment Typologies and Design Elements

Our refined definition of peer assessment incorporates what we refer to as *design elements* of peer assessment or in other words the variables that can describe the potential design characteristics of a peer assessment activity (e.g. anonymity). Topping (1998) concluded that peer assessment practices varied widely and first proposed a typology of 17 unordered variables to describe peer assessment designs in higher education. Although the peer assessment research community has taken up Topping's (1998) typology, subsequent refinements have not made the systematic description of peer assessment designs easier. Over time researchers (a) (re)ordered variables in conceptually motivated clusters (Adachi et al., 2018; Gielen et al., 2011; Van den Berg et al., 2006a, b) or removed (Van Gennip et al., 2009) and/or assigned variables to a different cluster compared to a prior refinement (Adachi et al., 2018; Gielen et al., 2011; Van Gennip et al., 2009), (b) revised or changed variable labels and/or descriptions (Adachi et al., 2018; Gielen et al., 2011; Van Gennip et al., 2009), (c) added more variables (Adachi et al., 2018; Gielen et al., 2011), and/or (d) subsumed and revised variables from Topping's (1998) initial typology under a new variable label (Adachi et al., 2018; Gielen et al., 2011) (Online Resource 1 provides an overview of these four refinements to Topping's typology; https://osf.io/z2vju/?view_only=cda003b0a7484d029296e7a8ec3829d9). Moreover, these four refinements distinguish between two separate clusters describing on the one hand the interaction between students and on the other hand composition of assessment groups, as if they were unrelated, which is clearly not the case (Strijbos et al., 2009). Although Topping (2013, 2017, 2021a, b) also made consecutive refinements, these complicate rather than elucidate the identification of design elements for four reasons. First, the list of elements was ever expanded (55 important factors in 2013, 43 variations in 2017, 44 variations in 2021a, and 79 pointers in 2021b). Second, Topping not only avoids dealing with the issue of clustering (e.g. 'The difficulty is that different researchers propose different clusters, so I have left the list un-clustered.', Topping, 2017, p. 6; Topping, 2021a), but is also inconsistent (six sections in 2013; seven categories in the 2021b). Third, some refinements provide the impression of limited or fixed design options as reflected by an 'alternative A' and 'alternative B' juxtaposition (Topping, 2017, 2021a). Fourth, the refinements increasingly include aspects that are not design elements per se but rather possible outcome measures such as 'improvement' and 'transferable skills' (Topping, 2017, 2021a) or possible moderators/mediators like 'gender' and 'previous experience' (Topping, 2017, 2021b). Given the highly problematic nature of consecutive refinements by Topping, we focus on the initial Topping (1998) typology and the four refinements that (near exclusively) stress peer assessment elements that teachers can design for.

The variations in the four refinements of Topping's (1998) initial typology are especially notable with respect to the intended purpose of the typology. Topping (1998) explicitly stressed that 'Since peer assessment practices are so varied,

future reports should include information on all 17 parameters in the typology (…) giving the basis for subsequent meta-analytic blocking. Also included should be information on participant characteristics and research design' (p. 268). Gielen et al. (2011) and Adachi et al. (2018) echoed this goal in repeated calls for systematic reporting of peer assessment design elements to foster meta-analysis and systematic review. However, all four refinements of Topping's (1998) typology gradually drifted away from enabling precisely such a comparison of studies on peer assessment in general, as well as the constitutive design elements in particular.

Gielen et al. (2011) proposed the term 'diversity' of peer assessment to highlight the expansion of variation in peer assessment designs that were no longer sufficiently captured by Topping's (1998) typology. In our study, we define *peer assessment diversity* as the variety of design elements that can be considered, the degree to which these design elements are taken into account, and the degree to which they co-occur. Although Topping's (1998) initial typology has been echoed in the literature through four refinements of his work, we are still missing a clear picture of which design elements are more commonly used in research studies and, thus, might reflect core peer assessment designs. The theoretically assumed diversity of peer assessment design elements needs to be empirically tested. Moreover, there might be differences in peer assessment designs given the subject domain (Double et al., 2020; Li et al., 2016; Sanchez et al., 2017; Van Zundert et al., 2010) and whether a formative or summative purpose is adopted (cf. Sanchez et al., 2017). Differences can also be expected in the object of peer assessment (e.g. written assignment, presentation) (Falchikov & Goldfinch, 2000; Li et al., 2016; Topping, 1998, 2003), the target outcome(s) of peer assessment (Double et al., 2020; Huisman et al., 2019; Li et al., 2016; Sanchez et al., 2017; Speyer et al., 2011; Van Zundert et al., 2010), and moderators and mediators that influence peer assessment processes (Double et al., 2020; Huisman et al., 2019; Panadero & Alqassab, 2019; Sanchez et al., 2017; Van Gennip et al., 2009). Despite the continuous increase in peer assessment design elements, some of them are closely interconnected. For instance, peer assessment involving revisions has a formative purpose. Hence, researchers and teachers might make implicit decisions regarding design elements due to decisions on related design elements. Some design elements are, therefore, more likely to co-occur in peer assessment designs. The more prominent co-occurrence of certain peer assessment design elements (cf. Double et al., 2020; Huisman et al., 2019; Sanchez et al., 2017) might reflect latent peer assessment design patterns. Identifying patterns with a limited set of inter-related design elements can support teachers and simplify the design of peer assessment in practice. Hence, we conducted a systematic review of empirical studies on peer assessment to meet these ends.

Aim and Research Questions

In order to obtain a cross-section of the field to improve our understanding of peer assessment design and practices in research studies, we used the refined typology by Gielen et al. (2011) as the starting point for our systematic review, because (a)

it includes earlier typologies with comparatively few adaptations and (b) sufficient time has passed for the field to notice this typology and assume a potential impact on the reporting of peer assessment designs. Yet, we refined some variables drawing from contemporary reviews (Strijbos et al., 2009; Van Zundert et al., 2010) in service of our research goals and analysis (details will be provided in the 'Method' section).

In sum, the aim of the present systematic review is to determine how diverse peer assessment designs and practices are in research studies and whether they are as diverse as they theoretically can be—and are often claimed to be—or whether the designs and practices are in fact limited in research studies. To this end, we will address the following research questions:

1. To what extent do peer assessment research studies differ according to subject domain, assessment purpose, objects, outcomes, and moderators/mediators?
2. Which peer assessment design elements, within all peer assessment elements combined, co-occur more often and thereby reflect latent peer assessment design patterns?

Method

Search Strategy

We conducted a search in March 2015 in the following electronic databases: PsycINFO, PsycARTICLES, ERIC, Academic Search Premier, and EconLit via EBSCOhost, Medline via Ovid, and Web of Science Core Collection. These databases were searched to ensure that studies in different educational fields (e.g. medicine, management) are covered given the wide range of educational contexts in which the implementation of peer assessment is investigated. The following search terms were used: 'peer assessment' OR 'peer feedback' OR 'peer rating' OR 'peer correction' OR 'peer nomination' OR 'peer review' OR 'peer appraisal' OR 'peer grading' OR 'peer marking' OR 'peer evaluation' OR 'peer ranking' in the title, abstract, or keywords search fields. The reproducible searchers for all databases are available online at https://osf.io/x3zhj/?view_only=effb02162d1045cc93023d3099637177.

The search was limited to peer-reviewed journal articles published in English language with no year restriction at this stage. We opted to exclude grey literature (i.e. dissertations, book chapters, and conference proceedings) over concerns regarding lack of design information to be extracted for this review, unclear explanation of the peer assessment intervention, anecdotal reports of peer assessment without data to validate its impact or usefulness, and the feasibility of analysing a large volume of reports given the detailed nature of data extraction involved in this review. We updated the database search in October 2018 using the same search method, except that we narrowed the searches to articles published After March 2015.

A total of 1535 citations were identified from the seven searched databases in the first database search (in March 2015) and a total of 930 in the updated

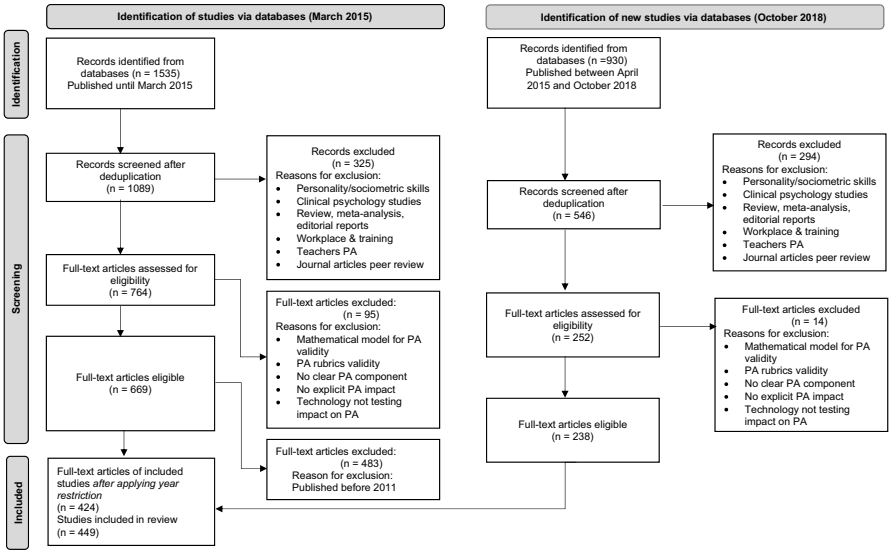

Fig. 1 PRISMA flow diagram for the identification, screening, and inclusion of research studies in this review ($n = 449$ studies in 424 reports)

database search (in October 2018; see Fig. 1). Records were exported to a reference management software, and duplicates were identified and then removed manually. This process rendered a total of 1089 in March 2015 and 546 in October 2018 that were screened against the inclusion and exclusion criteria.

Inclusion and Exclusion Criteria

To be included in this review, studies had to meet the following criteria: (a) the study had to investigate the implementation of peer assessment and/or its impact in any learning context, (b) the peer assessment component was clearly distinguished from other peer-assisted learning activities (e.g. peer tutoring, peer modelling, peer monitoring, peer coaching), (c) the peer assessment was performed on a learning task, (d) the study is empirical and published in a peer-reviewed journal, (e) the study is published in English language. We allowed both qualitative and quantitative studies and studies done at any level of education: primary, secondary, or higher education (including postgraduate). Only articles published from 2011 until October 2018 were included in the analyses for this review. We used Gielen et al. (2011) as the starting point for our systematic review, because (a) it includes earlier typologies with comparatively few adaptations, and (b) sufficient time has passed for the field to notice this typology and assume a potential impact on the reporting of peer assessment designs.

Studies were excluded for the following reasons: (a) studies investigated peer assessment of sociometric skills or personality traits, (b) clinical psychology studies using peer assessment for diagnosis purposes, (c) reviews and meta-analyses of peer assessment research or editorial comments/reports, (d) peer assessment that was

conducted in workplace or other work-related training activities (e.g. workshops), (e) studies on peer assessment between teachers, (f) studies on peer review between researchers (i.e. journal articles peer review), (g) studies that only proposed mathematical models to improve the validity of peer assessment, (h) studies that evaluated the validity of peer assessment rubrics without direct test of peer assessment outcomes, (i) studies on peer tutoring and peer-assisted learning that did not have a clear component of peer assessment, (j) studies that included peer assessment but did not explicitly investigate its implementation or its impact (e.g. teachers' perceptions and beliefs about peer assessment), (k) studies with formative assessment tools involving several components without testing the individual impact of peer assessment (e.g. problem-based learning, collaborative activities), and (l) studies that introduced technology to enhance communication between peers without testing the effect on peer assessment.

Selection Process

We selected the studies following a two-stage process as shown in Fig. 1 (PRISMA flow diagram): (1) screening titles and abstracts of potentially eligible studies followed by (2) screening the full texts of the preliminary selected articles. The assessment of each record was only performed by one reviewer (the first author). In case of unclear records during the full-text screening stage, the first author discussed each case with the second author until a consensus was reached on including or excluding the articles. A total of 907 articles have been identified as eligible, yet only articles published from 2011 onwards were included in the analyses for this review. This process resulted in a total of 424 articles and 449 studies (see Fig. 1 PRISMA flow diagram; Online Resource 2 provides the complete list of articles and is available via https://osf.io/728sc/?view_only=cdc919edc1fe4dbf82775cc307094f62).

Data Extraction

We developed a coding scheme to explore the diversity of peer assessment designs based on the peer assessment typology that was initially proposed by Topping (1998) and later reclassified/extended by other researchers (Adachi et al., 2018; Gielen et al., 2011; Van den Berg et al., 2006a, b; Van Gennip et al., 2009). We started with the typology by Gielen et al. (2011), yet, we refined some of the peer assessment design elements drawing from contemporary reviews (Strijbos et al., 2009; Van Zundert et al., 2010) in service of our research goals and analysis. See Appendix for details on the sources and modifications made to the design elements. Subsequently, the first and second authors regrouped the design elements under four main categories that offer a more meaningful lens in particular for systematic review (meta-analysis more generally), 'context', 'instructional design', 'outcomes', and 'moderators/mediators'. We define each category as follows:

- Context refers to what is given and designers cannot manipulate.
- Instructional design refers to what designers can manipulate.

- Outcomes refer to what designers try to influence by what they manipulate.
- Moderators/mediators refer to what designers typically do not manipulate but need to take into account when determining the effect of what they manipulate.

The resulting coding scheme consisted of (1) five design elements under the *context* category, (2) twenty-two design elements under the *peer assessment instructional design* category, (3) *outcome variables* of the peer assessment activity category, and (4) *moderators/mediators* of the peer assessment effects category (see Table 1 for the list of the design elements as well as Online Resource 3 containing the elaborate coding scheme with definitions and detailed description of each design element is available online via https://osf.io/4jbr3/?view_only=9f5b223115f244ac88ac5b054eb21149).

We classified the subject domain in which the peer assessment was conducted—as one of the context variables—according to the International Standard Classification of Education ISCED-F 2013 (UNESCO Institute for Statistics, 2014). This classification provides a wide range of fields of education and training (e.g. natural sciences, mathematics and statistics, social sciences, information and communication technology) in secondary, post-secondary, and higher education in formal settings. Yet, it can also be applied to other educational levels as well as in informal education (UNESCO Institute for Statistics, 2014).

For each article, the following study details were extracted: names of authors, year of publication, type of study (quantitative, qualitative, mixed-methods), study aims, study design (pre-experimental, quasi-experimental, experimental), study variables (dependent and independent variables), sample type (primary education, secondary education, higher education, vocational education, postgraduate education), sample size, and sample characteristics (e.g. demographics). There were in total 424 articles and 449 studies included in the review (some articles included multiple studies). Table 2 illustrates the main study details of the studies included in this review (the full list of articles and their study detail are available in Online Resource 2 via: https://osf.io/728sc/?view_only=cdc919edc1fe4dbf82775cc307094f62). The coding scheme was used to extract the relevant data from the 424 articles included in this review after establishing a sufficient inter-rater reliability.

Coding Reliability

We established the reliability of coding before applying the year limit to the included articles by randomly selecting articles from the 907 identified articles. This was done to ensure (a) the coding scheme was applicable to peer assessment designs in studies published before Gielen et al. (2011), given that Gielen et al.'s typology we are relying on was built on that literature, and (b) the coders receive sufficient training on a wider range of studies. Over several rounds, two pairs of coders independently coded 10–14% of the total number of the articles. The first author served as a common coder in both coding-pairs, and the fourth and sixth authors were the second coders. Each coding-pair coded a different set of articles each time. In each round, the coders first coded the articles independently, and subsequently disagreements were identified and discussed until resolved. Measures of inter-rater reliability that control for agreement by chance are the golden standard when assessing inter-rater reliability (Krippendorff, 2004; Lombard

Table 1 Summary of coding scheme of peer assessment (PA) design elements

Category/design element	Definition/description
Peer assessment context	*What is given and designers cannot manipulate*
Subject domain	Subject domain the study was done in (e.g., mathematics, English, business, physical education, multiple subjects, etc.)
Place/time	Where was the PA conducted? In class/ class time; out of class/free time
Setting	Formal; informal
Requirement	Was PA compulsory or voluntary for assessor/assesse?
Alignment	Was the PA activity aligned to curriculum, learning goals or teaching?
Peer assessment instructional design	*What designers can manipulate*
Purpose	Summative; formative; both
Object	What was assessed? (e.g., written assignment, presentation)
Product/output of PA	What was the output of PA? (e.g,. scoring, comment, scoring and comment)
Relation to staff assessment	What was the PA relationship to staff assessment? substitutional; supplementary
Official weight	Did participation in the PA activity or the grade given by peer(s) contribute to learners' final grades?
Directionality	Unidirectional; bi-directional
Degree of interactivity	Based on Strijbos et al. (2009): reactive; reciprocal; negotiated
Frequency	How often was the PA of one task done? Once; iterative
Group constellation	Was PA conducted within a group (intragroup) or between groups (intergroup) or both?
Constellation assessor	'The number of assessors assigned to each unit of assesse' (Gielen et al., 2011, p.148). One; two; three or more; five or more; ten or more; twenty or more
Constellation assessee	'The number of assessees per unit of assessor' (Gielen et al., 2011, p.148). One; two; three or more; five or more; ten or more; twenty or more
Unit of assessment (assessor)	At what level did the assessor perform the PA? Group; individual; both
Unit of assessment (assessee)	At what level did the assessee experienc the PA? Group; individual; both
Privacy	Public; confidential; single-blind (assessor); single-blind (assessee); double blind (anonymous); both
Contact	The nature of contact between the assessor and the assessee; whether PA was done face-to-face or online
Matching	How were assessor and assessee matched for the PA activity? (e.g., random, gender, ability)
Reward	Was there a reward for participation in PA?
Format	How was the PA guided (e.g., information, criteria, rubrics, prompts) or freestyle?
Training	Did learners receive PA training before participating in the actual PA activity?
Revision	Did learners revise their work after receiving or providing PA?

Table 1 (continued)

Category/design element	Definition/description
Scope of involvement	Aspects of involving learners in the PA. Gave and received PA; gave PA only; gave and received peer feedback (PF); gave PF only; received PF only; gave AND/OR received PA/PF (in experimental conditions)
Peer assessment outcomes	*What designers try to influence by what they manipulate.* These variables are directly measured as outcomes of the PA activity (i.e., Why was the PA activity conducted?), e.g., beliefs, performance, reliability, and validity of PA
Peer assessment moderators/mediators	*Variables that are not usually manipulated but taken into account when the effect of PA design on outcomes is investigated.* E.g., gender, ability and skills, psychological factors

For detailed description of each design element, refer to Online Resource 3 (available online only); *PA* peer assessment, *PF* peer feedback

Table 2 Main study details of studies included in the current review

Study details	Number of studies	Percentage
Study type		
Quantitative	191	42.54
Qualitative	123	27.39
Mixed-methods	132	29.39
Missing	3	0.67
Total	449	100
Study design		
Pre-experimental	338	75.28
Experimental	37	8.24
Quasi-experimental	72	16.00
Missing	2	0.44
Total	449	100
Level of education		
Primary	19	4.23
Secondary	42	9.35
Vocational	8	1.78
Higher	331	73.72
Postgraduate	33	7.35
Higher and postgraduate	11	2.45
Missing	5	1.11
Total	449	100

et al., 2004; Warrens, 2015). Examples of chance-corrected measures are Cohen's kappa and Krippendorff's α. We used the latter measure in the current study. Values of chance-corrected measures will be deflated (i.e. close to zero) for certain design elements if a

single peer assessment practice is almost always used. Examples in the current study are setting = formal, and alignment to curriculum = yes. In this case of extreme unbalanced marginal distributions, reliability is generally difficult to assess (see, e.g. Feinstein and Cicchetti, 1990; Krippendorff, 2004; Weinberger & Fischer, 2006), and there is no agreement in the literature on what measure should be preferred in this case (Lombard et al., 2004). In the current study, both Krippendorff's α and percentage agreement are reported. The latter measure is a linear transformation of the more robust coefficient proposed in Brennan and Prediger (1981) and is used in this study to assess inter-rater reliability in the case of extreme unbalanced marginals (Lombard et al., 2004; Warrens, 2008, 2010). Furthermore, an inflation of Krippendorff's α was also observed in our data where $\alpha = 1$ when the percentage agreement was merely 80%. An explanation to this observation is that Krippendorff's α relies on list-wise deletion of cases with a missing value by one rater (in the case of two raters).

Due to the vagueness of the description of peer assessment designs in most of the published articles—an issue that was also previously highlighted by Topping (1998)—reaching an agreement of above 80% for all the peer assessment design elements was not feasible. Many studies were also unclear on their study details or inaccurately reported them (e.g. experimental vs. quasi-experimental design). A minimum of 70% agreement was considered adequate for each category of the study details and the peer assessment design elements. As shown in Table 3, coding-pair 1 reached the minimum threshold of agreement on 100% of the study details categories, and most of the percentage agreements were well above 80%. For the peer assessment design elements, an acceptable level of agreement was reached for 26 out of the 28 elements. The only two elements that were below the 70% threshold were place/time and degree of interactivity. For coding-pair 2, the agreement on study details was also reached for all the categories. Agreements on the peer assessment instructional design were reached for 25 out of the 28 categories. The minimum threshold of 70% was not obtained for the following categories: relation to staff assessment, individual constellation assessor, and individual constellation assessee. After establishing an acceptable inter-rater reliability for the majority of variables, the first author (212 articles), fourth author (109 articles), and the sixth author (103 articles) coded the 424 articles included in this review (the dataset can be accessed via https://osf.io/728sc/?view_only=cdc919edc1fe4dbf82775cc307094f62).

Statistical Analysis Plan

We used descriptives and frequency tables to describe the sample of the articles included and to study the first and second research question. Various studies either did not consider certain peer assessment design elements or did not report on them. Both cases were coded as missing values. Furthermore, some of the included articles consisted of multiple studies, and some of these studies considered multiple categories of the peer assessment design elements. For example, some studies covered multiple subject domains or considered multiple outcome variables. Therefore, either the total number of studies, if applicable, or

Table 3 Inter-rater reliability in percentage agreements and Krippendorff's alpha (α) for study details and peer assessment instructional design categories between coding-pair 1 and coding-pair 2

Coding category	Coding-pair 1		Coding-pair 2	
	% agreement	Krippendorff's α	% agreement	Krippendorff's α
Study details				
Study type	77%	*0.604*	70%	*0.613*
Design	89%	*0.670*	82%	*0.536*
Aim	98%	0.977	95%	0.977
Independent variable(s)	96%	0.948	77%	0.717
Dependant variables(s)	88%	0.894	77%	0.855
Sample type	94%	0.895	89%	0.867
Sample size	87%	0.878	82%	0.860
PA context				
Subject domain	81%	0.926	96%	1
Place/time	*68%*	*0.488*	70%	*0.684*
Setting	100%	0	98%	0
Requirement	80%	1	80%	1
Alignment to curriculum	96%	−0.004	95%	0.965
PA instructional design				
Purpose	80%	0.730	75%	*0.592*
Object	96%	1	93%	0.952
Product/ output of PA	80%	0.794	84%	0.813
Relation to staff assessment	70%	*0.387*	*64%*	*0.331*
Official weight	86%	0.715	72%	*0.171*
Directionality	89%	1	70%	*0.333*
Degree of interactivity	*66%*	*0.521*	79%	*0.492*
Frequency	84%	*0.653*	88%	*0.616*
Group constellation	85%	0.791	77%	*0.563*
Individual constellation (assessor)	72%	0.814	*66%*	0.775
Individual constellation (assesse)	72%	0.808	*66%*	0.789
Unit of assessment (assessor)	87%	0.817	82%	*0.659*
Unit of assessment (assesse)	86%	0.869	86%	0.888
Privacy	74%	0.828	74%	0.800
Contact	83%	0.953	82%	0.862
Matching	81%	0.848	79%	0.913
Reward	88%	0.793	86%	*−0.049*
Format	71%	*0.343*	71%	*0.525*
Training	90%	0.793	89%	*0.653*
Revision	78%	0.826	79%	0.900
Scope of involvement	85%	0.809	84%	0.753
PA outcomes	81%	0.838	88%	0.881
PA moderators/mediators	86%	*0.588*	84%	(-)

The values were reached between coding-pair 1 after three rounds of coding and coding-pair 2 after two rounds of coding; values below the thresholds of 70% or 0.700 (for α) are in italic; (-) not possible to calculate due to all missing values for one coder

the occurrences of the categories of the peer assessment design elements were reported. The total number of occurrences did not necessarily correspond to the total number of articles nor to the total number of studies.

The third research question was approached with latent class analysis (LCA; Vermunt & Magidson, 2005), which is a model-based clustering method that can be used to discover complex patterns in multivariate data. LCA can be applied to nominal, ordinal, and interval measurements and any combination thereof. When applied to categorical (nominal, ordinal) data, LCA relates the peer assessment design elements to a set of latent classes, which are discrete latent variables. A class is characterized by a pattern of conditional probabilities that indicate the chance that peer assessment design elements take on certain values. The LCA also identifies which peer assessment design elements discriminate well between the studies (i.e. which variables are important predictors of the latent classes) and which elements do not. Some studies considered multiple categories of certain peer assessment design elements (e.g. multiple outcome variables), and to include all this information in the LCA, the studies were included multiple times—once for each unique combination of its peer assessment design elements. For example, if a study considered three outcome variables and had no multiple values for any of the other peer assessment design elements, the study was included three times—once for each outcome variable, using the same values for the other peer assessment design elements. The sample size for the LCA consisted of 891 cases.

LCA was performed with Latent GOLD (version 5.0; Vermunt & Magidson, 2005). Since there was no theoretically expected number of classes, LCA models with one to eight classes were estimated, and all 891 cases were included in the estimation. Missing data were not a problem for the maximum likelihood estimation procedures in Latent GOLD: the procedures used all information available for each case. All peer assessment design elements were used as indicators in the estimation of the LCA models, except for the subject domain. All peer assessment design elements were treated as nominal variables, except 'constellation assessor' and 'constellation assessee', which were treated as ordinal variables.

The optimal number of classes was determined using (a) the Bayesian information criterion (BIC; Schwarz, 1978), (b) the consistent Akaike's information criterion (CAIC; Bozdogan, 1987), and (c) the corresponding classification error (Nylund et al., 2007; Schreiber, 2017). Lower values of BIC, CAIC, and classification error indicate a better model fit (Aho et al., 2014). As a best practice for reporting on model fit (Schreiber, 2017), the value of the log-likelihood (LL), the number of parameters (Npar), Akaike's information criterion (AIC; Akaike, 1974), the entropy R^2, and the standard R^2 (coefficient of determination) were also reported. The two R^2 statistics reflect how well the latent classes can be predicted from the scores on the design elements and are based on a multinomial logistic regression model. The Wald statistic was used to test whether peer assessment design elements discriminate between the classes. As a measure of effect size, an R^2 value was reported for each design element. This R^2 statistic reflects how well the latent classes can be predicted from the scores on the design element. For interpretation of the classes, only design elements with $R^2 > 0.20$ were considered. According to Cohen (1988, p. 413–414), R^2 values of 0.13 and 0.26 correspond to medium and large effect sizes, respectively.

Table 4 Occurrences of subject domains

Subject domain	Occurrence	Percentage
Education	81	17.65
Arts and humanities	118	25.71
Social sciences, journalism, and information	32	6.97
Business, administration, and law	25	5.45
Natural sciences, mathematics, and statistics	87	18.95
Information and communication technologies	37	8.06
Engineering, manufacturing, and construction	19	4.14
Health and welfare	54	11.76
Services	3	0.65
Agriculture, forestry fisheries, and veterinary	0	0
Generic programmes and qualifications	3	0.65
Total	459	100

Ten studies reported multiple subject domains

We acknowledge that an $R^2 > 0.20$ is an arbitrary rule (Lakens, 2013; Thompson, 2007). Nevertheless, we use this arbitrary rule instead of interpreting the effect size in comparison to other effects in the literature because to our knowledge no other studies on peer assessment have employed LCA. Finally, the LCA conditional probabilities, which indicate the chance that peer assessment design elements take on certain values, were used to find descriptions of the classes of the model with the optimal number of classes.

Results

RQ1: Diversity of Peer Assessment Research Studies

The diversity of peer assessment research studies was investigated in terms of differences in subject domain, assessment purposes, objects, outcomes, and moderators/mediators of peer assessment. As some studies reported multiple subject domains, objects, outcomes, and/or moderators/mediators, the number of these studies is indicated as a note to each respective table.

Subject Domain

Table 4 presents the occurrences of the subject domains of the studies included. Arts and humanities occurred most frequently ($n = 118$; 25.71%), followed by natural sciences, mathematics and statistics ($n = 87$; 18.95%), education ($n = 81$; 17.65%), and health and welfare ($n = 54$; 11.76%). Furthermore, only a few studies researched services ($n = 3$; 0.65%) and generic programmes and qualifications ($n = 3$; 0.65%), and no study covered agriculture, forestry, fisheries, and veterinary.

Table 5 Percentages of purpose of peer assessment across subject domains

Subject domain	Summative	Formative	Both
Education	37.18	57.69	5.13
Arts and humanities	20.87	75.65	3.48
Social sciences, journalism, and information	45.16	54.84	0
Business, administration, and law	48.00	48.00	4.00
Natural sciences, mathematics, and statistics	34.11	61.18	4.70
Information and communication technologies	54.29	42.86	2.86
Engineering, manufacturing, and construction	52.63	36.84	10.53
Health and welfare	54.72	37.74	7.55
Services	33.33	33.33	33.33
Agriculture, forestry fisheries, and veterinary	0	0	0
Generic programmes and qualifications	0	66.67	33.33

Ten studies reported multiple subject domains resulting in 26 cases coming from the same studies

Assessment Purposes

In most studies, the purpose of assessment was formative ($n=257$; 59.08%). In a substantial number of studies, the purpose of assessment was summative ($n=157$; 36.09%), and only in a small number of studies, the purpose was both summative and formative ($n=21$; 4.83). Table 5 presents a cross-classification of the purpose of assessment and the subject domains. In most subject domains, formative assessment was most often used (formative assessment was also the most abundant category). In contrast, Table 5 shows that summative assessment was used relatively often in peer assessment studies in the subject domains of 'information and communication technologies' (54.29%), 'engineering, manufacturing, and construction' (52.63%), and 'health and welfare' (54.72%).

Objects of Peer Assessment

Table 6 presents the occurrences of the objects of assessment considered in the studies included. Almost half of the studies used a written assignment (45.76%). The least used object of assessment was test performance (0.89%).

Outcomes of Peer Assessment

Outcome variables that were most commonly studied were beliefs and perceptions ($n=275$; 38.46%), performance and skills ($n=154$; 21.54%), and content of peer feedback ($n=126$; 17.62%). Reliability of peer assessment ($n=55$; 7.69%), validity of peer assessment ($n=82$; 11.47%), and processing of peer feedback ($n=23$; 3.22%) were studied less often. Thus, most peer assessment designs do not necessarily focus on outcomes in terms of reliability and validity. Table 7 presents the occurrences of the outcome variables used in the studies included.

Table 6 Occurrences of objects used in studies included

Object	Occurrence	Percentage
Written assignment	205	45.76
Presentation	34	7.59
Test performance	4	0.89
Contribution group work	46	10.27
Instructional design	32	7.14
Professional skills	47	10.49
Artifact design	50	11.16
Problem solving	30	6.69
Total	448	100

Seven studies reported multiple objects

Table 7 Occurrences of peer assessment (PA) outcomes in studies included

Moderators/mediators	Occurrence	Percentage
Beliefs and perceptions	275	38.46
Performance and skills	154	21.54
Reliability of PA	55	7.69
Validity of PA	82	11.47
Peer feedback content	126	17.62
Peer feedback processing	23	3.22
Total	715	100

215 studies reported multiple outcomes

Moderators/Mediators of Peer Assessment

Table 8 presents the occurrences of the moderators and mediators considered in the studies included. Moderators and mediators that were most commonly studied in relation to peer assessment designs were gender ($n=28$; 27.45%) and ability and skills ($n=35$; 34.31%). All other categories occurred only in a relatively small number of studies (see Table 8).

RQ2: Latent Design Patterns: Co-occurrence of Peer Assessment Design Elements

Table 9 presents the model fit statistics corresponding to the LCA models with one to eight classes. The 5-class model had both the lowest BIC value (31,462.1) and the lowest CAIC value (31,864.1). Thus, the 5-class model has the optimal number of classes according to the BIC and CAIC statistics. Furthermore, the classification error of the 5-class model (0.036) was very similar to the classification errors of the 3-class and 4-class models. The 3-class model, 4-class model, and 5-class model were further explored, but only the 5-class model was considered in great detail.

Table 8 Occurrences of peer assessment moderators/mediators in studies included

Moderators/mediators	Occurrence	Percentage
Gender	28	27.45
Ability and skills	35	34.31
Age	3	2.94
Culture, ethnicity and race	4	3.92
Perceptions of assessee and/or assessor	2	1.96
Training and prior experience with peer assessment	5	4.90
Peer feedback characteristics	5	4.90
Number of assessors/assesses	3	2.94
Psychological factor	7	6.86
Domain (e.g. major, specialization)	7	6.86
Contribution to group work	3	2.94
Total	102	100

Number of studies investigating moderators/mediators = 77 studies; 16 of these studies reported multiple moderators/mediators

Table 9 LCA model fit evaluation information

Nclass	LL	Npar	BIC	AIC	CAIC	Classification error	Entropy R^2	R^2
1	− 16,772.0	86	34,128.4	33,716	34,214.4	0	1	1
2	− 15,535.1	165	32,191.3	31,400.2	32,356.3	.025	.906	.921
3	− 15,089.7	244	31,837.3	30,667.4	32,081.3	.035	.915	.920
4	− 14,705.0	323	31,604.7	30,056	31,927.7	.036	.925	.923
5	− 14,365.3	402	31,462.1	29,534.7	31,864.1	.036	.938	.930
6	− 14,141.7	481	31,551.7	29,245.5	32,032.7	.048	.930	.915
7	− 13,963.5	560	31,731.9	29,046.9	32,291.9	.032	.954	.944
8	− 13,794.3	639	31,930.3	28,866.6	32,569.3	.034	.952	.941

Both the entropy R^2 (0.938) and the standard R^2 (0.930) of the 5-class model were quite high.

For each variable of the 5-class model, the Wald statistic and associated *p*-value, together with the R^2 value that was used to assess the size of each effect, are presented in Table 10. Most peer assessment design elements were statistically significant at the 0.001 level. This is in part due to the large sample size ($N = 891$). The ten design elements with the highest effect sizes are in bold and are numbered 1 to 10 in Table 10. Each of these design elements had an associated effect size R^2 that exceeds 0.20 and discriminated well across the five classes. The ten design elements are unit of assessment assessee ($R^2 = 0.880$), group constellation ($R^2 = 0.681$), revision ($R^2 = 0.476$), unit of assessment assessor ($R^2 = 0.444$), purpose ($R^2 = 0.370$), scope of involvement ($R^2 = 0.363$), contact ($R^2 = 0.315$), product/output of peer assessment

Table 10 Fit measures for each variable of the 5-class model

Variable	Wald	*p*-value	R^2
Place/time	98.7	< .001	.172
Setting	5.9	.210	.027
Requirement	14.1	.007	.064
Alignment	0.1	1.000	.003
5. Purpose	225.5	< .001	.370
Object	154.4	< .001	.099
8. Product/output of PA	260.5	< .001	.249
Relation to staff assessment	32.8	< .001	.057
Official weight	146	< .001	.167
Directionality	12.3	.015	.069
Degree of Interactivity	19.8	.011	.106
Frequency	12.5	.014	.030
2. Group constellation	174.1	< .001	.681
Constellation assessors	80.5	< .001	.154
Constellation assessee	84.3	< .001	.148
4. Unit of assessment assessor	193.1	< .001	.444
1. Unit of assessment assessee	91.1	< .001	.880
10. Privacy	69.4	< .001	.204
7. Contact	154	< .001	.315
9. Matching	81.1	< .001	.217
Rewards	12.5	.130	.007
Format	146.4	< .001	.136
Training	15.3	.004	.020
3. Revision	90.6	< .001	.476
6. Scope of involvement	222.5	< .001	.363
PA outcomes	160.4	< .001	.059
PA moderators/mediators	56.7	.042	.159

In bold most relevant variables (1–10) in terms of $R^2 > .20$

($R^2 = 0.249$), matching ($R^2 = 0.217$), and privacy ($R^2 = 0.204$). The effect sizes associated with unit of assessment assessee and group constellation were very large, and the other effect sizes in this list can be considered medium to large (Cohen, 1988). The three peer assessment design elements that were not statistically significant at the 0.05 level and had very small effect sizes are alignment ($p = 1.000$; $R^2 = 0.003$), rewards ($p = 0.130$; $R^2 = 0.007$), and setting ($p = 0.210$; $R^2 = 0.027$). There was no evidence that these elements discriminate between the five classes, given the other design elements in the model. In fact, these design elements could not discriminate between the studies since (in our sample) typically a single peer assessment practice was used (e.g. setting = formal, and alignment to curriculum = yes). Furthermore, note that the design elements for which inter-rater reliability was difficult to assess—due to extreme unbalanced marginal distributions—were identified by the LCA as less relevant.

The ten peer assessment design elements with the highest effect sizes were all instructional design variables. It appears that the context variables, outcome variables, and moderators and mediators were less important for discriminating between studies than the instructional design elements. To check what the most important instructional design elements were in a LCA using the instructional design variables only, an additional LCA was performed. The ten most important peer assessment design elements in Table 10 are also the ten most important contributors of this additional LCA. The five classes solution of the variability in peer assessment designs is described below (see Table 11):

- Class 1: Online asynchronous formative anonymous random-matched bi-directional intragroup (individually assessed) peer assessment via written comments [with revision]
- Class 2: Online asynchronous summative anonymous random-matched bi-directional intragroup (individually assessed) peer assessment via scoring [without revision]
- Class 3: Face-to-face synchronous formative non-anonymous self-selection matched bi-directional intragroup (individually assessed) peer assessment via oral and written comments [with revision]
- Class 4: Online asynchronous formative anonymous random-matched bi-directional intergroup (individuals assessing groups) peer assessment via scoring and written comments [with revision]
- Class 5: Online asynchronous summative non-anonymous domain-matched[1] bi-directional intra- and intergroup (individual and group) peer assessment via scoring [without revision]

For the 5-class model, Table 12 presents the class sizes, as well as the conditional probabilities of the ten design elements that had the highest effect sizes (see Table 10). Classes 1, 2, and 3 were relatively large classes that contained 32.8%, 25.7%, and 21.6% of the sample, respectively. Classes 4 and 5 were relatively small classes that contained, respectively, 13.1% and 6.8% of the sample. The conditional probabilities in Table 12 were used to find abovementioned descriptions of the classes. The purpose of peer assessment in studies in classes 1, 3, and 4 was more likely to be formative, whereas the purpose of peer assessment was more likely to be summative in studies in classes 2 and 5. The output of peer assessment was more likely to be written comments in studies in class 1, scoring in studies in classes 2 and 5, written and oral comments in studies in class 3, and scoring and written comments in studies in class 4. The group constellation was more likely to be intragroup

[1] Matching by domain refers to cases in which students were matched based on their studied subject, academic major, or specialization (depending on level of education) and is thus different from the subject domain that is one of the peer assessment context variables, i.e. the domain that students are matched on (e.g. engineering vs. informatics in higher education) can be different from the subject domain of the peer assessment research study (e.g. scientific writing).

Table 11 Peer assessment design elements significantly contributing to the 5-class model

Class	*Decisions concerning the implementation of peer assessment*				*Assessment groups and peer interaction within them*					
	5. Purpose	6. Scope	8. Product	3. Revision	2. Group constellation	4. Unit of assessment assessor	1. Unit of assessment assessee	9. Matching	7. Contact	10. Privacy
1	Formative	Gave and received PF	Comment (written)	Yes	Intragroup	Individual	Individual	Random	Online asynchronous	Anonymous
2	Summative	Gave and received PA	Scoring	No	Intragroup	Individual	Individual	Random	Online asynchronous	Anonymous
3	Formative	Gave and received PF	Comment (oral + written)	Yes	Intragroup	Individual	Individual	Self-select	Face-to-face	Public
4	Formative	Gave and received PF	Scoring + comment (written)	Yes	Intergroup	Individual	Group	Random	Online asynchronous	Anonymous
5	Summative	Gave and received PA	Scoring	No	Both	Both	Both	Domain	Online asynchronous	Public

Elements numbered based on their effect size; class sizes: class 1 = .328, class 2 = .257, class 3 = .216, class 4 = .131, class 5 = .068; *PF* peer feedback, *PA* peer assessment

Table 12 Class sizes and conditional probabilities of the 10 'most important' design elements of the 5-class model

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
Class size	.328	.257	.216	.131	.068
Purpose					
Summative	.143	.758	.058	.294	.867
Formative	.788	.169	.931	.62	.133
Summative + formative	.069	.073	.011	.086	.000
Product/Output of PA					
Scoring	.018	.632	.013	.057	.734
Comment (O)	.011	.000	.207	.034	.000
Comment (W)	.584	.000	.223	.235	.001
Comment (O + W)	.009	.000	.354	.052	.065
Scoring + comment (W)	.378	.346	.068	.459	.134
Scoring + comment (O)	.000	.012	.049	.060	.066
Scoring + comment (O + W)	.000	.009	.085	.103	.000
Group constellation					
Intragroup	.943	.826	1.00	.001	.102
Intergroup	.045	.138	.000	.978	.001
Both	.012	.035	.000	.021	.897
Unit of assessment assessor					
Group	.022	.006	.037	.314	.000
Individual	.978	.976	.914	.55	.102
Both	.000	.018	.048	.136	.898
Unit of assessment assessee					
Group	.011	.020	.000	.919	.000
Individual	.989	.971	1.00	.013	.003
Both	.000	.008	.000	.069	.997
Privacy					
Public	.348	.216	.972	.298	.481
Confidential	.039	.121	.007	.081	.000
Single blind (assessor)	.011	.014	.007	.000	.000
Single blind (assessee)	.032	.227	.000	.089	.222
Double blind (anonymous)	.546	.403	.001	.476	.297
Both	.024	.020	.013	.056	.000
Contact					
Face-to-face synchronous	.064	.382	.825	.331	.439
Online synchronous	.026	.005	.042	.028	.000
online asynchronous	.910	.613	.133	.641	.561
Matching					
Random	.597	.667	.234	.718	.026
Ability & skills	.218	.096	.137	.042	.007
Self-select	.089	.129	.564	.220	.011
Gender	.022	.033	.000	.000	.001

Table 12 (continued)

Variable	Class 1	Class 2	Class 3	Class 4	Class 5
Age	.022	.000	.000	.000	.000
Domain	.022	.043	.000	.000	.954
Other	.030	.033	.065	.019	.002
Revision					
Yes	.779	.091	.971	.681	.284
No	.221	.909	.029	.319	.716
Scope of involvement					
Gave and received peer assessment	.089	.761	.077	.202	.736
Gave peer assessment only	.000	.035	.000	.000	.000
Received peer assessment only	.007	.004	.000	.000	.000
Gave and received peer feedback	.778	.191	.923	.798	.264
Gave peer feedback only	.058	.000	.000	.000	.000
Received peer feedback only	.048	.000	.000	.000	.000
Gave and/or received peer assessment/feedback	.021	.009	.000	.000	.000

in studies in classes 1, 2, and 3 and intergroup or both (intergroup and intragroup) in studies in classes 4 and 5, respectively. Studies in classes 1, 2, and 3 were very likely to use unit of assessment as individual for both assessor and assessee. In contrast, studies in class 4 were likely to use unit of assessment as individual for assessor but as group for assessee, whereas studies in class 5 were likely to use Unit of assessment as both (group and individual) for both assessee and assessor.

Furthermore, studies in classes 1, 2, and 4 were more likely to use anonymous peer assessment, whereas studies in classes 3 and 5 were more likely to use public peer assessment. Moreover, the contact between assessors and assessees in studies in classes 1, 2, 4, and 5 was likely to be online asynchronous, whereas in studies in class 3 was likely to be face-to-face synchronous. The matching in peer assessment was more likely to be random in studies in classes 1, 2, and 4, self-select in studies in class 3, and based on domain (e.g. subject, specialty, academic major, degree) in studies in class 5. Finally, studies in classes 1, 3, and 4 were quite likely to involve students in revising their work after peer assessment (i.e. revision = yes) with giving and receiving peer feedback as scope of involvement, whereas studies in classes 2 and 5 were quite likely to have no revision after peer assessment with giving and receiving peer assessment as scope of involvement.

We found the following correspondence between the 3-class model, 4-class model, and 5-class model. The classes were not identical, since they were composed of (slightly) different studies, but the interpretations of classes 1, 2, and 3 of the 5-class model, in terms of the design elements, were analogous to the interpretations of the classes of the 3-class model. Furthermore, the interpretations of classes 1, 2, and 3 and 5 of the 5-class model, in terms of the design elements, were analogous to the interpretations of the classes of the 4-class model.

Figure 2 presents the distributions of the subject domains across the five classes. Studies from ‘education’, ‘arts and humanities’ and ‘natural sciences, mathematics,

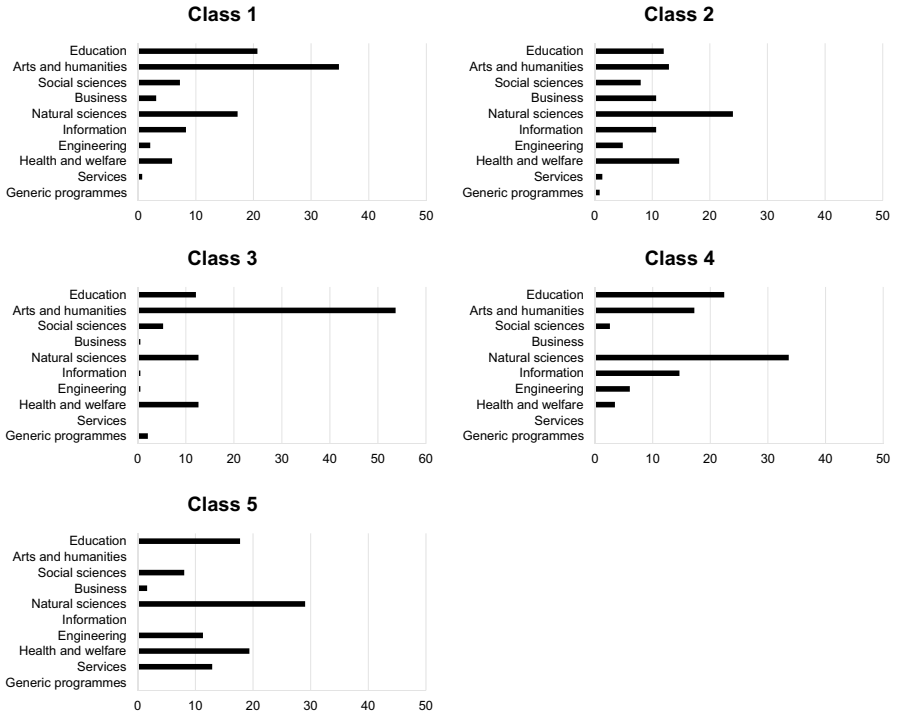

Fig. 2 Distributions of the subject domains across the five classes

and statistics' were major contributors to the compositions of class 1. Class 2 was a quite diverse class, as it contained studies from most subject domains. Class 3 was dominated by studies from 'Arts and humanities'. Major contributors to class 4 were the subject domains 'education' and 'natural sciences, mathematics, and statistics'. Finally, class 5 contained studies from 'education' and 'natural sciences, mathematics, and statistics', as well as 'health and welfare'.

Discussion

Since Topping's (1998) initial typology of 17 variables to describe variations in peer assessment practices and his call that research studies add such information systematically in support of review and meta-analysis, several authors have proposed refinements to the initial typology (Adachi et al., 2018; Gielen et al., 2011; Van den Berg et al., 2006a, b; Van Gennip et al., 2009) and Topping as well (2013, 2017, 2021a, b). Some researchers echoed Topping's observation of a wide variety in peer assessment design and practices (i.e. diversity), and the call for more systematic description of peer assessment designs (Adachi et al., 2018; Gielen et al., 2011; Panadero & Alqassab, 2019). However, the assumed diversity of peer assessment designs and practices has remained theoretically

grounded rather than empirically grounded. Hence, our systematic review aimed to determine how diverse peer assessment designs and practices are in research studies and whether they are as diverse as they theoretically can be or that the designs and practices are in fact limited.

Diversity of Peer Assessment Research Studies According to Subject Domains, Assessment Purposes, Objects, Outcomes, and Moderators/Mediators

In line with the assumptions and analyses in existing review studies, we first explored differences across peer assessment research studies according to subject domains, assessment purposes, objects, outcomes, and moderators/mediators. We found that close to three-quarters of the included studies were conducted in four subject domains: 'education', 'natural sciences, mathematics, and statistics', 'arts and humanities', and 'health and welfare'; whereas few of the included studies were in the domains of 'information and communication technologies', 'social sciences, journalism, and information', 'business, administration, and law', and 'engineering, manufacturing, and construction'.

With respect to the assessment purpose, we found that close to two-third of all studies had a formative purpose, a little over one-third had a summative purpose, and close to 5% had both a summative and formative purpose. When split by subject domains, cross-classification revealed that a formative purpose was most often used in the majority of the domains, except in the domains 'information and communication technologies', 'engineering, manufacturing, and construction', and 'health and welfare' where a summative purpose was used more often compared to a formative purpose.

We observed a large range for the object of peer assessment. Close to half of the studies used a written assignment as the object of assessment and test performance was the least used object of assessment. Furthermore, among recent meta-analyses, only Li et al. (2016) explicitly investigated the object of assessment (referred to as 'task being rated') and concluded that assessment tasks in science/engineering might be more clear-cut and thus it might be easier for students to assess their peers.

Regarding peer assessment outcomes, our results showed that both reliability and validity were less often treated as an outcome measure compared to the impressions that many review studies provide (Ashenafi, 2017; Falchikov & Goldfinch, 2000; Li et al., 2016; Speyer et al., 2011; Topping, 1998, 2003, 2013, 2017; Van Zundert et al., 2010). Although reliability and validity of peer assessment as compared to teacher assessment has been established in general—thus supporting quantitative approaches to peer assessment in educational settings (i.e. peer scores and marks), this does not absolve research studies employing quantitative peer assessment from systematically determining and reporting the reliability and validity of peer assessment for their respective research study and peer assessment design. Both are crucial for determining the quality of the study in general as well as inclusion in future meta-analyses—even when they are not the outcome measures that are under study in a future meta-analysis. Furthermore, even beliefs and perceptions

of peer assessment were treated more often as an outcome variable compared to performance and skills—again, more so than one would expect from existing review studies (Double et al., 2020; Huisman et al., 2019; Sanchez et al., 2017). Considering that peer feedback content and peer feedback processing are comparatively more recent outcome measures in peer assessment research—as compared to scores and marks—the number of studies (18% and 3%, respectively) that included them as outcome measures can be deemed sizeable (See Table 7).

Finally, we found that among the factors that are considered as potential moderators and mediators of both peer assessment processes and outcomes, nearly two-thirds concern 'gender' and 'ability and skills' (Table 8). This clearly contrasts the claim by Ashenafi (2017) that 'Gender effects are the least studied factors in peer assessment in higher education…' (p. 232). Although the remaining studies covered a wide range of factors that can be potential moderators/mediators, it is striking that few studies included 'psychological factors' despite the attention over the past decade for the role of social and interpersonal factors (e.g. psychological safety, trust in self and other as assessor, fairness, friendship, interdependence) as relevant moderators and mediators of both peer assessment processes and outcomes (Panadero, 2016; Panadero & Alqassab, 2019; Panadero et al., 2018; Strijbos et al., 2009; Van Gennip et al., 2009).

Latent Design Patterns: Co-occurrence of Peer Assessment Design Elements

We performed two latent class analyses (LCA) of peer assessment design elements. The first LCA revealed that variables that belong to the context and the categories outcome variables and moderators/mediators had no or a small contribution when discriminating peer assessment designs, whereas the instructional design elements played a major role. We then performed a second LCA on the instructional design elements only, which revealed that the ten most important design elements in the first LCA were also the ten most important design elements in the second LCA. The five classes solution, reflecting latent patterns, that best describe the variability in peer assessment designs is illustrated in Table 11.

To our knowledge, only the review by Ashenafi (2017) alluded to the possibility of a (proto)typical peer assessment design in the context of higher education, stating: 'The most common implementation of peer assessment in higher education scenarios involves students making use of prespecified criteria to assess their peers and assign marks or grades, possibly providing additional written feedback' (p. 231). Yet, this (proto)typical design clearly does not match with any of the five latent design patterns we identified, despite the fact that around 84% of the studies we included have been conducted in higher education (including postgraduate). When we cross classified the clusters with the subject domains, we found that 'arts and humanities' was most prominent in classes 1 and 3 and 'natural sciences, mathematics, and statistics' in classes 2, 4, and 5. The two subject domains 'education' and 'social sciences, journalism, and information' showed a comparatively even contribution to each of the five classes. It is important to stress that these patterns are based on

what is most prominent, which means for each class that the vast majority of studies assigned to that class reflect the particular description of that pattern; but it does not mean that all of the studies do by default. It simply means that the pattern is the most dominant one for studies belonging to that class.

Importantly, the ten most prominent design elements that contributed to the five-class solution appear to belong to two core themes of peer assessment design (see Table 11). The first theme is *decisions concerning the implementation of peer assessment* (cf. Gielen et al., 2011). It includes the design elements: purpose (formative vs. summative), scope of involvement (e.g. provision and/or reception of peer assessment), the product/output of peer assessment (scoring and/or [written] comment), and whether the peer assessment involves revision. The second theme is *concerned with assessment groups and peer interaction within them* (i.e. 'composition of assessment groups' and 'interaction between peers'; Gielen et al., 2011). This theme includes the design elements: group constellation (intragroup vs. intergroup vs. both), unit of assessment assessor (group vs. individual vs. both), unit of assessment assessee (group vs. individual vs. both), matching (e.g. based on ability or gender), contact (e.g. face-to-face synchronous vs. online asynchronous), and privacy (e.g. anonymous vs. public). Hence, peer assessment designs in research studies are clearly not as diverse as they theoretically can be or as diverse as assumed and implied by Topping's initial typology and subsequent refinements (Adachi et al., 2018; Gielen et al., 2011; Topping, 1998, 2013, 2017, 2021a, b; Van den Berg et al., 2006a, b; Van Gennip et al., 2009). Indeed, Ashenafi (2017) also concluded from a narrative review that despite the growing number of peer assessment studies 'most studies have insignificant variations in the variables being studied and usually reach similar conclusions that neither strengthen nor contradict the findings of previous studies.' (p. 244). Across all clusters, the combined use of scoring and comments is comparatively less important, which is striking given that Ashenafi (2017) reported it to be among the most common practices of peer assessment in higher education. In addition, Double et al. (2020) found a positive effect for peer grading on performance in higher education. In our study, written comments contributed to the largest identified class and thus was more common in peer assessment research (see Table 11). A possible explanation is the closer alignment between peer feedback and the notion of 'assessment for learning' that is advocated for in the peer assessment literature (e.g. Panadero et al., 2018; Wiliam & Thompson, 2007). However, we acknowledge that the importance of some peer assessment design elements (e.g. anonymity), and consequently their common use in research and practice, might be theory driven rather than empirically supported (Double et al., 2020). Furthermore, matching based on 'ability and skills' is not dominant among any of the identified classes which is quite surprising given that ability and year—as an indicator of skill and experience—are the main elements of Topping's (1998) typology that he stated students should be matched on. Although, the 'matching' element was extended by Gielen et al. (2011) to include broader matching principles (e.g. random, subject, social constellation), we stress that 'ability and skills' is an important matching principle given there is accumulating evidence that it can affect the processes and outcomes of peer assessment (e.g. Alqassab et al., 2018; Huisman et al., 2018; Patchan & Schunn, 2016; Patchan et al., 2013).

Beyond Peer Assessment Design Elements: Study Type, Study Design, and Educational Level

Apart from the limited diversity with respect to peer assessment design elements, our observations with respect to the orientation of research in terms of study types, study designs, and level of education deserve further comment. First, although included studies were quite balanced in terms of a quantitative, qualitative, and mixed-methods orientation, our finding for study design is worrisome. Despite the multiple calls by, among others, Topping (1998), Strijbos and Sluijsmans (2010), and Van Zundert et al. (2010) for more quasi-experimental and experimental designs, both types, combined, only accounted for 24% of the included studies, whereas about 76% was coded as using pre-experimental research designs (i.e. studies conducted without any experimental manipulations). Irrespective of the added value of the holistic approaches to study peer assessment (which often employ qualitative methods) and the relative increase in (quasi-) experimental studies over the past decade (Double et al., 2020), we would like to repeat the call for quasi-experimental and experimental studies. This type of studies provides more control over conditions to determine the impact of specific peer assessment design elements on core peer assessment mechanisms and outcomes (cf. Double et al., 2020).

Furthermore, almost 84% of the studies we included were conducted in higher education (including postgraduate education), and this overrepresentation is in line with several recent reviews that included multiple levels of education, i.e. 94% of all included studies by Li et al. (2016) were in higher education, 74% in Panadero and Alqassab (2019), and 54% in Double et al. (2020). Despite the observations by Topping (1998, 2003, 2013, 2021a) and Van Zundert et al. (2010) of limited research on peer assessment in both primary and secondary education, with the recent reviews by Hoogeveen and Van Gelderen (2013) and Sanchez et al. (2017) as notable exceptions, our results confirm an overrepresentation of research in higher education. Thus, our findings predominantly show that diversity in research studies in higher education is limited; and for sure more limited than it can be in theory.

Limitations

First, our use of the ISCED-F 2013 classification for fields of education (UNESCO Institute for Statistics, 2014) might be considered too elaborate given its 11 main categories and that not all categories are applicable to primary education (e.g. health and welfare), but it served our main purpose to describe the subject domains of peer assessment research at all levels of education in a comprehensive way. Moreover, Double et al. (2020) distinguished 14 domains, but due to small sample sizes in nearly all categories, they compared 'writing' to 'domains other than writing' (other categories collapsed). Likewise, Li et al. (2016) opted for three main categories of 'social science/arts', 'science/engineering', and 'medical/clinical' given sample size concerns to limit the number of dummy-coded predictor variables in their meta-regression model. Although such a reduction in

the case of meta-analysis might be difficult to avoid, the field of peer assessment research might opt for a community effort to better understand their field, pooling resources to gradually cover and expand the subsample sizes to enable more elaborate comparisons between subject domains.

Secondly, we have taken great care to establish inter-rater reliability for the coding of the peer assessment design elements. While coding the studies, descriptions of study details were sometimes unclear or inaccurately reported, and peer assessment design elements were often vague, unclear, or had to be partly inferred. In most cases we achieved the minimum of 70% for inter-rater reliability, except for three of the twenty-eight categories (relation to staff assessment, individual constellation assessor, and individual constellation assessee). We are, regrettably, not alone in observing that the information on study details and peer assessment design elements in many studies is either too limited or needs to be inferred (cf. Adachi et al., 2018; Ashenafi, 2017; Double et al., 2020; Li et al., 2016; Panadero & Alqassab, 2019). This clearly signals that—despite the call by Topping (1998) nearly 25 years ago—we have yet to systematically report on our studies to enable more efficient extraction of information for the purposes of meta-analysis and systematic review. As most journals nowadays offer the option to add supplemental material, a systematic description—using Topping's (1998) typology or Gielen et al.'s (2011) typology, or even the coding scheme of the present review (Online Resource 3)—can easily be added to any empirical study, and the argument of limited space no longer applies.

Finally, the five identified classes—or in other words, latent patterns that reflect five 'prototypical peer assessment designs'—can be critiqued in that they consist of more basic or traditional design elements and they are far simpler than how peer assessment practices might be in reality. However, in our view, this only highlights that there are too few studies on other design elements to claim the purported diversity of peer assessment. Additionally, peer assessment designs in primary, secondary, or vocational education are likely to differ from designs implemented in higher education, for example, in terms of the composition of assessment constellations, complexity of the learning tasks, and degree to which peer assessment will be taking place online. For example, Double et al. (2020) found in their meta-analysis that peer grading was beneficial for higher education students' performance, but not for students in secondary or primary education. Latent patterns of peer assessment designs at these levels of education can, thus, be expected to look different—especially as the majority of studies we included were conducted in higher education (84%). While this is an important issue to be investigated by future research, still limited subsample sizes of peer assessment studies at primary (4%) and secondary (9%) education that we included, as well as the resultant limited occurrences of peer assessment design elements, would render unstable LCA results if analysed separately per educational level.

Implications for Research and Practice

The complexity of peer assessment makes it difficult to disentangle the effects of different design elements without conducting controlled experiments (Double et al.,

2020; Panadero & Alqassab, 2019), and it seems that advocating for the increasing peer assessment diversity does not help in elucidating the benefits of different peer assessment practices. In fact, the five classes and latent design patterns we have identified might act as a scaffold for future experimental studies and meta-analyses. Researchers can use this review to guide them in designing focused experimental studies to test the differential impact of peer assessment design elements that are more prominently used in the literature and whether for instance the effects of these design patterns would differ across subject domains or educational levels. This is especially important given that, currently, some decisions regarding the use of certain peer assessment design elements appear to be theoretically motivated rather than empirically supported (Double et al., 2020). Moreover, guided by their finding that students can benefit from peer assessment regardless of its design, Double et al. (2020) encouraged teachers to liberally implement peer assessment to accommodate their learners' needs. However, in our opinion, teachers would benefit from having access to some simple design patterns with inter-related design elements to facilitate the design of peer assessment activities in classrooms. For example, teachers can consult this review to identify proto-typical peer assessment designs that are more commonly used by research conducted in the subject domain they teach, or they can use Table 11 as a quick reference to design basic peer assessment activities.

Conclusion

Overall, our findings have shown that peer assessment designs in research studies are not as diverse as they theoretically can be and are often assumed or implied to be. Our review revealed that (a) the subject domain that most frequently implemented peer assessment was arts and humanities, (b) in most studies, the purpose of assessment was formative, and (c) almost half of the studies used a written assignment as an object of peer assessment. Further, the most commonly studied outcome of peer assessment was students' beliefs and perceptions, and 'ability and skills' and gender were the most investigated moderators/mediators of peer assessment. We found five latent and prototypical peer assessment design patterns to which the instructional design elements were the most important contributors, whereas variables that belong to context, the outcomes variables, and moderators/mediators did not contribute. We showed that the most prominent design elements belong to two core themes of peer assessment design: decisions concerning the implementation of peer assessment and assessment groups and peer interactions within them (Gielen et al., 2011). Although the diversity that we found is narrower than assumed in the literature, we continue to encourage researchers in the field of peer assessment to report study characteristics as well as peer assessment designs more elaborately and precisely in future research studies. Doing so will not only assist future meta-analyses and systematic reviews but will also help to accumulate evidence for elements that were comparatively less important in the identified latent designs but which are nonetheless relevant for the implementation of peer assessment practices.

Appendix

Table 13 Peer assessment (PA) design elements extracted from previous typologies and reviews and included in the present review

	Primary source(s)	Correspondence with other typologies in reviews	Modifications made for the present review
PA context			
Curriculum area/subject	Topping (1998)	In Gielen et al. (2011) part of 'setting'	We changed the label to 'subject domain'
Place/Time	Topping (1998)	In Gielen et al. (2011) part of 'setting' and 'context'	We combined 'place' and 'time' in a single variable
Setting	Gielen et al. (2011)	Does not occur before Gielen et al. (2011)	This is multidimensional in Gielen et al. (2011). We limited it to 'formal vs. informal'
Requirement	Topping (1998)	Identical in Van Gennip et al. (2009) Identical in Gielen et al. (2011)	-
Alignment	Gielen et al. (2011)	Does not occur before Gielen et al. (2011)	-
PA instructional design			
Purpose	Van Gennip et al. (2009)	In Topping (1998) called 'focus' In Gielen et al. (2011) called 'function' We opted for 'purpose' from Van Gennip et al. (2009) as it is a more contemporary terminology	-
Object	Gielen et al. (2011)	In Van Gennip et al. (2009) called 'objectives measured'	-
Product/output of PA	Topping (1998)	In Van Gennip et al. (2009) called 'outcomes' In Gielen et al. (2011) called 'output'	We added 'of PA' to emphasize that it refers to the PA itself, e.g., ratings, marks, grades, feedback
Relation to staff assessment	Topping (1998)	In Gielen et al. (2011) called 'relation to other assessments'	-
Official weight	Topping (1998)	In Gielen et al. (2011) part of 'function'	-
Directionality	Topping (1998)	Identical in Van Gennip et al (2009) Identical in Gielen et al. (2011)	We use 'unidirectional vs. bi-directional' by Strijbos et al. (2009) instead of the 'one-way vs. reciprocal vs. mutual' by Topping (1998)
Degree of interactivity	Strijbos et al. (2009)	Does not occur in Topping (1998), Van Gennip et al. (2009), or Gielen et al. (2011)	We added this variable to account for the nature of bi-directionality in terms of 'reactive vs. reciprocal vs. negotiated'
Frequency	Gielen et al. (2011)	Does not occur before Gielen et al. (2011), where it is called 'frequency and previous experience'	We retained 'frequency' as part of instructional design, but moved the 'previous experience' part to the main analysis category of 'mediators/moderators'
Group constellation	Strijbos et al. (2009)	Does not occur in Topping (1998), Van Gennip et al. (2009), or Gielen et al. (2011)	We added this variable to account for intra- and inter-group peer assessment

Table 13 (continued)

	Primary source(s)	Correspondence with other typologies in reviews	Modifications made for the present review
Constellation assessor	Topping (1998)	Identical in Van Gennip et al. (2009) In Gielen et al. (2011) called 'constellations of assessors and assessees'	-
Constellation assessee	Topping (1998)	Identical in Van Gennip et al. (2009) In Gielen et al. (2011) called 'constellations of assessors and assessees'	-
Unit of assessment (assessor)		Does not occur in Topping (1998), Van Gennip et al. (2009), or Gielen et al. (2011)	We added this variable to report group constellation due to the wide variety in constellations used
Unit of assessment (assessee)		Does not occur in Topping (1998), Van Gennip et al. (2009), or Gielen et al. (2011)	We added this variable to report group constellation due to the wide variety in constellations used
Privacy	Topping (1998)	Identical in Van Gennip et al. (2009) Identical in Gielen et al. (2011)	-
Contact	Topping (1998)	Identical in Van Gennip et al. (2009) Identical in Gielen et al. (2011)	-
Matching	Gielen et al. (2011)	Does not occur before Gielen et al. (2011); subsumes variables 'year' and 'ability' in Topping (1998) and Van Gennip et al. (2009)	This is multidimensional in Gielen et al. (2011). We included only the 'principle of matching'
Reward	Topping (1998)	Identical in Van Gennip et al. (2009) Identical in Gielen et al. (2011)	-
Format	Gielen et al. (2011)	Does not occur before Gielen et al. (2011)	-
Training	Gielen et al. (2011)	Does not occur before Gielen et al. (2011), where it is called 'training/guidance'	Gielen et al. (2011) use 'extent of training/guidance', whereas we treat it as dichotomous (yes/no)
Revision	Van Zundert et al. (2010)	Does not occur in Topping (1998), Van Gennip et al. (2009), or Gielen et al. (2011)	-
Scope of involvement	Gielen et al. (2011)	Does not occur before Gielen et al. (2011)	-

Since Van den Berg et al., (2006a, b) only clustered the original 17 variables by Topping (1998), we omitted this article to simplify the overview.

Acknowledgements The authors like to thank Nikolai Klitzing, Meike Faber, Fabian Kracher, Monique Messado, Elvetia Parker, Wai Shan Tong, and Alex Horton for their assistance in coding trails.

Author Contribution Maryam Alqassab: conceptualization, methodology, investigation, data curation, formal analysis, writing—original draft, project administration.
Jan-Willem Strijbos: conceptualization, methodology, writing—original draft, supervision.
Ernesto Panadero: supervision, writing—review and editing.
Javier Fernández Ruiz: formal analysis, writing—review and editing.
Matthijs Warrens: methodology, formal analysis, visualization, writing—original draft.
Jessica To: formal analysis, writing—review and editing.

Funding The first author's research was funded by the German Elite Network of Bavaria (Reference Number: K-GS-2012-209) during the initial conceptualization and data search stages (March 2014 – October 2016), and by the Spanish Ministry of Science and Innovation (Ministerio de Ciencia e Innovación) under the Juan de la Cierva Incorporación program (Reference number: IJC2020-043302-I) during the first and second revisions of this manuscript (April 2022 – August 2022). Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data Availability Dataset is available online via: https://osf.io/728sc/?view_only=cdc919edc1fe4dbf82775cc307094f62.

Declarations

Conflict of Interest The authors declare no competing interests.

References

Adachi, C., Tai, J., & Dawson, P. (2018). A framework for designing, implementing, communicating and researching peer assessment. *Higher Education Research & Development, 37*(3), 453–467. https://doi.org/10.1080/07294360.2017.1405913

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology, 95*(3), 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alqassab, M., Strijbos, J. W., & Ufer, S. (2018). Training peer-feedback skills on geometric construction tasks: Role of domain knowledge and peer-feedback levels. *European Journal of Psychology of Education, 33*(1), 11–30. https://doi.org/10.1007/s10212-017-0342-0

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

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345–370. https://doi.org/10.1007/BF02294361

Brennan, R. L., & Prediger, D. J. (1981). Coefficient kappa: Some uses, misuses, and alternatives. *Educational and Psychological Measurement, 41*, 687–699.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.
de Hei, M., Strijbos, J. W., Sjoer, E., & Admiraal, W. (2016). Thematic review of approaches to design group learning activities in higher education: The development of a comprehensive framework. *Educational Research Review, 18*, 33–45. https://doi.org/10.1016/j.edurev.2016.01.001
Dijkstra, J., Latijnhouwers, M., Norbart, A., & Tio, R. A. (2016). Assessing the "I" in group work assessment: State of the art and recommendations for practice. *Medical Teacher, 38*(7), 675–682. https://doi.org/10.3109/0142159X.2016.1170796
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*, 481–509. https://doi.org/10.1007/s10648-019-09510-3
Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322. https://doi.org/10.3102/00346543070003287
Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology, 43*(6), 543–549. https://doi.org/10.1016/0895-4356(90)90158-l
Forsell, J., Forslund Frykedal, K., & Hammar Chiriac, E. (2020). Group work assessment: Assessing social skills at group level. *Small Group Research, 51*(1), 87–124. https://doi.org/10.1177/1046496419878269
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
Hoogeveen, M., & Van Gelderen, A. (2013). What works in writing with peer response? A review of intervention studies with children and adolescents. *Educational Psychology Review, 25*(4), 473–502. https://doi.org/10.1007/s10648-013-9229-z
Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101–110. https://doi.org/10.1111/bjet.12520
Huisman, B., Saab, N., Van den Broek, P., & Van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896
Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed.). Sage Publications.
Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology, 4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863
Lejk, M., Wyvill, M., & Farrow, S. (1996). A survey of methods of deriving individual grades from group assessments. *Assessment & Evaluation in Higher Education, 21*(3), 267–280. https://doi.org/10.1080/0260293960210306
Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2016). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation in Higher Education, 41*(2), 245–264. https://doi.org/10.1080/02602938.2014.999746
Li, H., Xiong, Y., Hunter, C. V., Guo, X., & Tywoniw, 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
Lombard, M., Snyder-Duch, J., Bracken, C. C. (2004). P*ractical resources for assessing and reporting intercoder reliability in content analysis research projects*. Retrieved October 30, 2010, from http://matthewlombard.com/reliability/
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
Meijer, H., Hoekstra, R., Brouwer, J., & Strijbos, J. W. (2020). Unfolding collaborative learning assessment literacy: A reflection on current assessment methods in higher education. *Assessment & Evaluation in Higher Education, 45*(8), 1222–1240. https://doi.org/10.1080/02602938.2020.1729696
Nylund, K., Asparouhov, T., & Muthén, B. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modelling, 14*(4), 535–569. https://doi.org/10.1080/10705510701575396
Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*:n71. https://doi.org/10.1136/bmj.n71

Panadero, E. (2016). Is it safe? Social, interpersonal, and human effects of peer assessment: A review and future directions. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 247–266). Routledge.

Panadero, E., & Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. *Assessment & Evaluation in Higher Education, 44*(8), 1253–1278. https://doi.org/10.1080/02602938.2019.1600186

Panadero, E., Jonsson, A., & Alqassab, M. (2018). Providing formative peer feedback: What do we know? In A. A. Lipnevich & J. K. Smith (Eds.), *The Cambridge handbook of instructional feedback* (pp. 409–431). Cambridge University Press. https://doi.org/10.1017/9781316832134.020

Patchan, M. M., & Schunn, C. D. (2016). Understanding the effects of receiving peer feedback for text revision: Relations between author and reviewer ability. *Journal of Writing Research, 8*(2), 227–265. https://doi.org/10.17239/jowr-2016.08.02.03

Patchan, M. M., Hawk, B., Stevens, C. A., & Schunn, C. D. (2013). The effects of skill diversity on commenting and revisions. *Instructional Science, 41*(2), 381–405. https://doi.org/10.1007/s11251-012-9236-3

Ploegh, K., Tillema, H. H., & Segers, M. S. R. (2010). In search of quality criteria in peer assessment practices. *Studies in Educational Evaluation, 35*(2–3), 102–109. https://doi.org/10.1016/j.stueduc.2009.05.001

Sanchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049–1066. https://doi.org/10.1037/edu0000190

Schreiber, J. B. (2017). Latent class analysis: An example for reporting results. *Research in Social and Administrative Pharmacy, 13*(6), 1196–1201. https://doi.org/10.1016/j.sapharm.2016.11.011

Schwarz, G. (1978). Estimating dimensions of a model. Annals of Statistics 6(2): 461–464. https://www.jstor.org/stable/2958889

Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer- and co-assessment. *Learning Environments Research, 1*(3), 293–319. https://doi.org/10.1023/A:1009932704458

Speyer, R., Pilz, W., Van der Kruis, J., & Brunings, J. W. (2011). Reliability and validity of student peer assessment in medical education: A systematic review. *Medical Teacher, 33*, e572–e585. https://doi.org/10.3109/0142159X.2011.610835

Strijbos, J. W. (2016). Assessment of collaborative learning. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 302–318). Routledge.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265–269. https://doi.org/10.1016/j.learninstruc.2009.08.002

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., Segers, M. S. R., & Tillema, H. H. (2009). Fostering interactivity through formative peer assessment in web-based collaborative learning environments. In C. Mourlas, N. Tsianos, & P. Germanakos (Eds.), *Cognitive and emotional processes in web-based education: Integrating human factors and personalization* (pp. 375–395). IGI Global.

Strijbos, J. W., Sluijsmans, D., Stegmann, K. (2017). Inferring individual scores from group scores via peer assessment: Part 1 – Methodological review [Conference presentation]. In *Assessment and Evaluation SIG invited symposium EARLI 2017 conference, Tampere, Finland*. EARLI.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools, 44*(5), 423–432. https://doi.org/10.1002/pits.20234

Tillema, H., Leenknecht, M., & Segers, M. (2011). Assessing assessment quality: Criteria for quality assurance in design of (peer) assessment for learning – A review of research studies. *Studies in Educational Evaluation, 37*(1), 25–34. https://doi.org/10.1016/j.stueduc.2011.03.004

Topping, K. (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). Kluwer.

Topping, K. (2017). Peer assessment: Learning by judging and discussing the work of other learners. *Interdisciplinary Education and Psychology, 1*(1), 7. https://doi.org/10.31532/InterdiscipEducPsychol.1.1.007

Topping, K. (2021a). Peer assessment: Channels of operation. *Education Sciences, 11*(3), 91. https://doi.org/10.3390/educsci11030091

Topping, K. (2021b). Face-to-face peer assessment in teacher education/training: A review. *The Educational Review, 5*(5), 117–130. https://doi.org/10.26855/er.2021.05.002

Topping, K. (1998). Peer assessment between students in colleges and universities. Review of Educational Research 68(3): 249–276. http://www.jstor.org/stable/1170598

Topping, K. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 394–412). SAGE publications. https://doi.org/10.4135/9781452218649.n22

UNESCO Institute for Statistics (2014). *ISCED fields of education and training 2013 (ISCED-F 2013): Manual to accompany the international standadrd classification of education 2011*. UNESCO Institute for Statistics (UIS). https://doi.org/10.15220/978-92-9189-150-4-en

Van den Berg, I., Admiraal, W., & Pilot, A. (2006a). Peer assessment in university teaching: Evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19–36. https://doi.org/10.1080/02602930500262346

Van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*(3), 341–356. https://doi.org/10.1080/03075070600680836

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54. https://doi.org/10.1016/j.edurev.2008.11.002

Van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P. R. J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review, 20*, 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

Van Zundert, M., Sluijsmans, D. M. A., & Van Merriënboer, J. J. G. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction, 20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Vermunt, J. K., & Magidson, J. (2005). *Latent GOLD 4.0 User's Guide*. Statistical Innovations.

Warrens, M. J. (2008). On similarity coefficients for 2x2 tables and correction for chance. *Psychometrika, 73*, 487–502. https://doi.org/10.1007/s11336-008-9059-y

Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4*, 271–286.

Warrens, M. J. (2015). Five ways to look at Cohen's kappa. *Journal of Psychology & Psychotherapy, 5*(04), 1–4. https://doi.org/10.4172/2161-0487.1000197

Weinberger, A., & Fischer, F. (2006). A framework to analyze argumentative knowledge construction in computer-supported collaborative learning. *Computers and Education, 46*(1), 71–95. https://doi.org/10.1016/j.compedu.2005.04.003

Wiliam, D., & Thompson, M. (2007). Integrating assessment with learning: What will it take to make it work? In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 53–82). Erlbaum.

Authors and Affiliations

Maryam Alqassab[1,2] · Jan-Willem Strijbos[1,3] · Ernesto Panadero[4,5] · Javier Fernández Ruiz[6,7] · Matthijs Warrens[3] · Jessica To[8]

Jan-Willem Strijbos
j.w.strijbos@rug.nl

Ernesto Panadero
ernesto.research@gmail.com

Javier Fernández Ruiz
javier.fernandez.uam@gmail.com

Matthijs Warrens
m.j.warrens@rug.nl

Jessica To
jessica.to@nie.edu.sg

1 Department of Psychology, LMU Munich, Munich, Germany

2 Faculty of Education, Universidad de Las Palmas de Gran Canaria, Las Palmas de Gran Canaria, Spain

3 GION Institute for Educational Research, Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands

4 Faculty of Psychology and Education, Universidad de Deusto, Bilbao, Spain

5 Basque Foundation for Science, Bilbao, Spain

6 Faculty of Psychology, Universidad Autónoma de Madrid, Madrid, Spain

7 Department of Psychology, Sociology, and Philosophy, Faculty of Education, Universidad de León, León, Spain

8 National Institute of Education, Nanyang Technological University, Singapore, Singapore