An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation, and peer grading

Ernesto Panadero [1] & Maryam Alqassab [2]

Author Note

[1] Departamento de Psicología Evolutiva y de la Educación. Facultad de Psicología. Universidad Autónoma de Madrid, Spain.

[2] Centre for the Enhancement of Teaching and Learning, The University of Hong Kong.

Recommended citation: 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*, 1-26. doi:10.1080/02602938.2019.1600186
*2019 is the online first. Once published in a regular issue, year might change.

Correspondence concerning this article should be addressed to: Ernesto Panadero. Despacho 109. Departamento de Psicología Evolutiva y de la Educación. Facultad de Psicología. Universidad Autónoma de Madrid, 28049, Cantoblanco. Spain. E-mail: ernesto.panadero@uam.es. Phone (+34) 914973553.

Acknowledgements: First author funded by the Ministerio de Economía y Competitividad via Spanish Ramón y Cajal programme (Referencia RYC-2013-13469). **We would like to thank Gavin T. L. Brown and Joanna Tai for providing detailed feedback on an earlier version of the manuscript**.

Authors' short bio

Ernesto Panadero is a researcher at the Developmental and Educational Psychology Department, Universidad Autónoma de Madrid (funded by the Ramón y Cajal research program - 2013 call) and an honorary professor at Deakin University (Australia) at the Centre for Research in Assessment and Digital Learning.

ORCID number: http://orcid.org/0000-0003-0859-3616

Maryam Alqassab is a research consultant at the Centre for the Enhancement of Teaching and Learning, The University of Hong Kong.

ORCID ID: https://orcid.org/0000-0002-6574-5113

Abstract

Peer assessment has proven to have positive learning outcomes. Importantly, peer assessment is a social process and some claim that the use of anonymity might have advantages. However, the findings have not always been in the same direction. Our aims were (a) to review the effects of using anonymity in peer assessment on performance, peer feedback content, peer grading accuracy, social effects and students' perspective on peer assessment; and (b) to investigate the effects of four moderating variables (educational level, peer grading, assessment aids, direction of anonymity) in relation to anonymity. A literature search was conducted including five different terms related to peer assessment (e.g., peer feedback) and anonymity. Fourteen studies that used a control group or a within group design were found. The narrative review revealed that anonymous peer assessment seems to provide advantages for students' perceptions about the learning value of peer assessment, delivering more critical peer feedback, increased self-perceived social effects, a slight tendency for more performance, especially in higher education and with less peer assessment aids. Some conclusions are that (a) when implementing anonymity in peer assessment the instructional context and goals need to be considered, (b) existent empirical research is still limited to establish strong conclusions, and (c) future research should employ stronger and more complex research designs.

Keywords: anonymity; peer assessment; peer review; peer feedback; peer review; peer evaluation; peer grading.

An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation, and peer grading

Peer assessment is an activity that has many educational benefits for students such as gaining domain-specific skills (van Zundert, Sluijsmans, & van Merriënboer, 2010), higher academic performance (Dochy, Segers, & Sluijsmans, 1999), and development of assessment skills (Sluijsmans, Brand-Gruwel, van Merriënboer, & Martens, 2004). Unsurprisingly, peer assessment is inherently a social activity requiring a mutual relationship of trust between assessor and assessee (van Gennip, Segers, & Tillema, 2009; Panadero, 2016). It has been argued that the quality of peer assessment processes can be improved if peer assessment is anonymous (e.g. Yu & Liu, 2009; Li, 2017) because this allows both parties to focus on the quality of the performance being evaluated independent of social aspects between parties (e.g. Peterson & Peterson, 2011). However, the impact of anonymity is not always positive (e.g. Yu, 2012) and a recent review claimed that simplistic approaches to anonymity might not overcome the complex social processes involved in peer assessment (Panadero, 2016). Thus, the aim of this review is to establish, within the known literature, the effects of anonymity in peer assessment practices. To answer these questions this paper reports a systematic and rigorous search for experimental studies that have used anonymous peer assessment which were then reviewed analytically for the study design features and the structure of the peer assessment activities. The paper contributes to the research literature by establishing that there is consistency in how anonymity in peer assessment contributes to learning outcomes. Further, the paper identifies facets of research designs in peer assessment studies that need to be made more robust.

In this paper, we define 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" (Topping, 1998, p. 250). We include a variety of synonyms under this definition including: peer feedback, peer review, peer evaluation, and peer grading. Detailed discussion of the differences between these terms can be found elsewhere (e.g. see Liu & Carless, 2006; van Gennip, 2012). When describing how interactions among peers takes place, anonymity is an important facet (Gielen, Dochy & Onghena, 2011). Anonymity can be *unidirectional* (i.e., either the assessor or the assessee is anonymous), or *bidirectional* (i.e., both the assessor & the assessee are anonymous). Additionally, anonymity can be provided in different ways, such as random assignment (e.g., assigning an ID), assignment without any identification, or by use of pseudonyms (e.g., Yu & Wu, 2011). Anonymity can also be used as a scaffold during initial stages of peer assessment interventions that can be faded out later once students feel more confident (Rotsaert, Panadero, & Schellens, 2018).

It is important to stress that the importance of anonymity might depend of the purpose of peer assessment (i.e., summative vs. formative). Strijbos et al. (2009) claimed that anonymous peer assessment might be preferable if there were summative or high-stake implications. However, a recent meta-analysis has found non-anonymous peer grading to be more accurate (i.e., higher correlation with teachers' grade) (Li et al., 2015). Additionally, a review by Panadero (2016) has suggested that formative peer assessment that usually involves more interactions among assessors and assessees (i.e., less anonymous settings) seems to have more positive effects on interpersonal factors.

However, anonymity in peer assessment activities does not exist in isolation. Peer assessment takes place in settings and activities that involve: (I) decisions concerning the use of peer assessment; (II) link between peer assessment and other elements in the learning environment; (III) interaction between peers; (IV) composition

of assessment groups; (V) management of the assessment procedure; and (VI) contextual elements (Adachi, Tai &, Dawson, 2018). Because as many as 19 different variables have been identified within these six facets related to peer assessment, there are many different combinations (or constellations) of aspects that might impact the efficacy of anonymity. Thus, one of the goals of this study is to identify the conditions under which anonymity is more beneficial for peer assessment.

Importantly, the above-mentioned constellation variables, that refer to different implementation decisions that need to be taken (e.g., aids for peer assessment) might be key for the effects of anonymity. For example, the presence of peer grading in a peer assessment intervention might have a moderating effect over the impact of anonymity on peer assessment accuracy. Indeed, there are some claims that approaching anonymity has been simplistic and using this on/off approach to anonymity might be obscuring our understanding of its effects (Panadero, 2016). To illustrate, the impact of anonymity can be weakened by the lack of accountability for the assessor on the accuracy of the feedback or ratings s/he provides. This could be counteracted by actions like: (a) teacher oversighting of comments or even evaluating the reliability of the peer feedback/grades; (b) assessors getting graded by assessees for the helpfulness of the comments; or (c) training/community building to support openness of feedback. Thus, the effect of anonymity will depend on other decisions that the instructor or researchers take regarding other aspects of the peer assessment activity. In the present study we will explore the moderating effects of the educational level of intervention (higher education vs. school), the presence of peer grading and peer assessment aids (e.g., rubrics), and anonymity type (unidirectional vs. bidirectional).

Effects of anonymity on different peer assessment outcomes

Since peer assessment is an interpersonal process, anonymity has been studied as a way to make this process fairer, safer, and more powerful (van Gennip et al., 2009; Panadero, 2016). Anonymity can influence two main types of peer assessment outcomes, namely *cognitive outcomes* and *social-affective outcomes*. Among the investigated cognitive outcomes are: performance or achievement, accuracy of peer assessment, and peer feedback content. First, it has been claimed that anonymity can enhance performance and achievement because in anonymous constellations, the assessor will mainly focus on analyzing the task instead of the person behind, and the assessee will rely more on the received feedback because s/he might perceive it as less biased, and thus will implement more changes suggested by the assessor (Yu, 2012). Secondly, regarding peer grading accuracy and peer feedback content, it has been claimed that anonymity will at least have two positive effects: the assessor will not be biased by knowing the assessee and the assessor will not be affected by repercussions of her evaluation having a positive influence in her honest assessment (Lu & Bol, 2007; Peterson & Peterson, 2011).

The social-affective outcomes mostly studied in relation to anonymity are students' perspective on peer assessment and related variables including peer pressure and disapproval (Vanderhoven, Raes, Montrieux, Rotsaert, & Schellens, 2015), psychological safety (Rotsaert et al., 2018), affective reactions (Bloom & Hautaluoma, 1987), and factors related to group dynamics such as frequency of interactions (Yu & Sung, 2015). The general tendency in this group of studies is to assume that anonymity will alleviate social tension derived from having students assessing each other.

Research questions

The aim of this study is to determine the effects of anonymity on different outcome variables taking into account the effect of some moderating variables. The following research questions will be addressed:

RQ1: What are the effects of peer assessment anonymity on: performance/achievement (RQ1a); peer feedback content (RQ1b); accuracy of peer grading (RQ1c); social and interpersonal variables (RQ1d); and students' perspective on peer assessment (RQ1e).

RQ2: Are these observed effects moderated by the participants' educational level (school vs. higher education), presence of peer grading, peer assessment aids (e.g., rubric, training), or anonymity type (unidirectional vs. bidirectional)?

Method

Selection of studies

The search was conducted over two phases. In both phases the databases PsycINFO and ERIC were searched. The following keyword combinations were used in the search in the title, abstract, and keywords fields: peer assessment/peer feedback/peer review/peer evaluation/peer grading AND anony*/blind/confidential. No time limit was set. In phase 1, we only limited the search to peer-reviewed journal articles. However, to make the search more comprehensive, a second literature search was conducted in phase 2 including peer-reviewed conference proceedings, doctoral dissertations, and peer-reviewed journal articles published after the search performed in the 1st phase. Phase 1 was conducted in January 2018 and phase 2 was conducted in January 2019. A snowball procedure was also used in phase 1 through examining the reference lists of empirical and review articles for additional references. The inclusion criteria used in this review were: (a) the study had anonymity as in independent variable; (b) the study included empirical results of anonymity interventions in relation to outcome variables (e.g. performance, students' attitudes and perceptions); (c) the study had at least one

control group or a within-subject design with different interventional phases (e.g., anonymity, no anonymity); (d) the study had been conducted in formal learning settings; (e) the study had been peer-reviewed (journal articles, dissertations or conference proceedings); and (d) the study was published in English.

Figure 1 presents a flow chart of the systematic review. In total, 182 records were identified through database and manual searches in phase 1. After removing duplicates, the abstracts of the 131 remaining publications were screened in order to select only relevant studies resulting in 104 articles being excluded because they did not meet one of the inclusion criteria (e.g., review papers; anonymity as a research procedure in survey studies; studies about publishing peer-reviewed journal articles). Ten articles were identified initially through snowball search and did not meet the inclusion criteria. The full texts of 27 empirical studies were read and assessed. In phase 2, 21 records were identified through parallel database search and after removing duplicates 13 records remained. These were screened but were all excluded because they did not meet the inclusion criteria (see Figure 1). After excluding the studies not meeting all inclusion criteria, 14 studies were included in the review.

>>> Figure 1 about here <<<

The following information was collected from the selected articles: general information (authors' names, year of publication), aim, research questions, hypotheses, sample characteristics and country, peer assessment terminology used, presence of grading in the peer assessment activity, description of the peer assessment constellation, use of aids (e.g., rubrics, training), type of anonymity (uni/bidirectional), peer assessment task, independent variables, dependent variables, study design, type of research, gender differences, procedure, results, conclusions, and observations. Table 1 presents a summary with the relevant information from the included publications.

Insert table 1 around here

The size of the studies in the sample range from 32 to over 243 participants (see table 2). In total, the studies encompass 1224 students from primary school to higher education. Studies involving students from primary (2) and secondary education (2) are few, while studies from higher education are more evenly represented in the sample. The studies are concentrated in a few countries USA (6), China (4) and Belgium (3). There is a larger representation of females in the samples, especially in the higher education studies. Only one of the included studies explored gender differences, so unfortunately it was not possible to explore gender effects in this review. Bloom and Hautaluoma (1987) found that females had more positive affective reactions to an imaginary feedback situation when the source of the feedback (i.e. an "imaginary" supervisor) was credible and anonymous. Importantly, this gender comparison was not formulated as a research question but rather seems to be a "side effect" that the authors discovered. Considering this study and the differences found in self-assessment (Panadero, Jonsson & Botella, 2017) this is an under-researched topic.

Insert table 2 around here

Although domain subjects such as computers and language are more common in the selected studies, there is large variation in the content domains (e.g., geography, mathematics, and psychology). Nine studies in the sample used a quasi-experimental design, while the remaining studies used experimental designs.

Review method

Considering the published research, it was not feasible to conduct a meta-analysis for two reasons. First, due to the small number of studies found on most of the dependent variables and, second, the large diversity in designs in terms of the type of control groups and the complexity of the variables surrounding anonymity (e.g., type of

measured outcomes, tasks, directionality of anonymity, etc.). Therefore, we adopted a narrative review method using first an in-depth analysis of the results, and, second a count of the results directions (Cooper, 2010). We created specific tables for each research question to better interpret the results while comparing the studies.

Coding the articles

To ensure the accuracy of the information included in our tables used to analyze the results, one of the authors interpreted the information from our database and translated it into the categories of the tables (e.g., design, presences of grading); then the other author checked the interpretation. These tables were later used to deduce the direction of anonymity effect based on the four moderating variables from RQ2. The two authors independently, interpreted the direction of anonymity effect for each study by assigning one point if anonymity had one or more significant effects in one direction (e.g., positive), or half points if anonymity had more than one significant effects in different directions (e.g., positive for one variable & negative for another variable). If no significant effects were reported then the effect is coded as *neutral* using the same points system. Out of the 14 studies, there was first agreement for 12. In the two disagreements the authors discussed the findings and agreed on a decision. These scores can be found in Table 1.

Results

RQ1a: What are the effects of anonymity in peer assessment on performance/achievement?

Three of the selected studies explored the effects of anonymity on academic performance (see Table 3). One of the studies found a significant gain for the anonymous condition reporting two effect sizes of .19 and .14 (Lu & Bol, 2007); another study found a significant gain for one of the non-anonymous conditions (with

training) and the anonymous condition over the other non-anonymous condition (without training) reporting an effect size of .27 (Li, 2017); the last study did not find significant differences (not reporting of effect size; Yu, 2012). The sample sizes of the studies are similar (N = 92, 101, and 77).

Considering the constellations variables of the studies (Table 3), Lu and Bol (2007) implemented a more intensive and prolonged intervention with many occasions for peer grading. Compared to Lu & Bol (2007), Yu (2012) did not implement peer grading, the intervention was shorter and did not include repeated occasions. Li (2017) had similar implementation conditions to Yu (2012), though in this study peer grading was used.

>>> Table 3 about here <<<

Looking at the design characteristics of the three studies, they share quasi-experimental designs using intact classroom as study groups; all of them used some control variables to account for differences between conditions (e.g., Scholastic Assessment Test –SAT-, high school Grade Point average -GPA- or scores in previous task performances). However, the quality of these control variables varied. Lu and Bol (2007) included stronger control variables such as external performance measures (i.e. SAT and high school GPA) and they used a counterbalanced design. Yu (2012) and Li (2017) used previous performance measures on the specific task used for their studies, and Yu (2012) did not report those analyses. Therefore, the quality of experimental results for Lu and Bol (2007) can be emphasized over the other two as these authors had a stronger control variable method. Additionally, the measurement of performance was stronger in Lu and Bol (2007) with external examination and raters, followed by Li (2017) with the instructor and an external rater on classroom materials, and finally Yu (2012) who did not report how the classroom exams were scored. Consider these issues,

we could preliminarily conclude that there seem to be a slight empirical advantage on anonymity benefits, but again this is only based on three studies.

RQ1b: What are the effects of anonymity in peer assessment on peer feedback content?

Four studies explored the effects of anonymity on peer feedback content (see Table 4). Reaching clear conclusions is complicated for two reasons: (a) the different peer feedback categorizations used (e.g., cognitive/metacognitive; verification/justification) and (b) the low number of studies. Nevertheless, two studies pointed out that the anonymous conditions provided more critical peer feedback (Lu & Bol, 2007; Howard, Barrett, & Frick, 2010). In terms of the different peer feedback categories, one of the studies found that the anonymous condition provided more cognitive but less affective comments, and both conditions were equal in terms of metacognitive comments (Lin, 2018). The study by Rotsaert et al. (2018) used a fading anonymity approach: firstly, when peer assessment was anonymous the quality of peer feedback was high, and, secondly, when anonymity disappeared the quality remained stable over time. The sample sizes varied for the different studies (range from 32-92) but the results seem to be in the same direction.

Considering the constellations variables of the studies (Table 4) it is not possible to extract strong conclusions because: (a) three of them used peer assessment aids; (b) two included peer grading; (c) three of the studies had a high number of occasions to assess and receive assessment; (d) peer assessment tasks were different; and (e) three of the studies had more extended interventions while one only had a single lecture session.

>>> Table 4 about here <<<

In terms of the quality of the research designs, one of the studies has a threat to validity because it employed a single group experimental design (Rotsaert et al., 2018).

The study by Lin (2018) was a controlled experiment, and the remaining two studies involved between subjects quasi-experiments (Lu & Bol, 2007; Howard et al., 2010). Only one of the quasi-experimental studies used a control variable for the different conditions (Lu & Bol, 2007). Therefore, the strongest experimental results are coming from Lin (2018) and Lu and Bol (2007). Considering the different characteristics of these studies, a preliminary conclusion is that anonymity seems to have a positive impact on providing more critical peer feedback.

RQ1c: What are the effects of anonymity in peer assessment on the accuracy of peer grading?

Two studies found that participants in the anonymity condition provided lower scores than participants in the non-anonymous condition (Lu & Bol, 2007) or the instructor (Peterson & Peterson, 2011). Additionally, one study found that grades by participants in the non-anonymous condition were much closer to the teacher's grades (Vanderhoven et al., 2015), while another study found higher correlations in the anonymous condition (Güler, 2017). Vanderhoven et al. (2015) reported that there might have been a validity threat due to the teacher and students seeing each other's scores and influencing the process. Güler (2017) claimed to have solved this threat. The study by Omelicheva (2005) did not find differences. Studies' sample sizes ranged from 69 to 110 participants.

Exploring the studies constellations variables (Table 5): (a) all studies but one used peer assessment aids (Omelicheva, 2005); (b) all studies but one compared the peer grade to the teachers' grades (Lu & Bol, 2007); (c) in terms of times of peer assessment one study clearly provided more assessment opportunities (Vanderhoven et al., 2015), three studies had similar numbers of opportunities, and one study did not provide information regarding the number of peer assessments (Güler, 2017); (d) in

terms of the task, three studies used group activities on various topics (e.g., instructional design, research methods) and the two other studies used writing activities; and (e) in terms of duration of the intervention, the studies used varying intervention occasions ranging from single session (e.g., Omelicheva, 2005) to a whole semester (e.g., Lu & Bol, 2007). Therefore, there is more homogeneity in this group of studies when it comes to the implementation variables than the two other groups presented in the previous sections.

>>> Table 5 about here <<<

Regarding the research designs, one of the studies had an experimental design (Omelicheva, 2005), and the remaining four studies had between subjects quasi-experimental designs with only one of them using a strong control method (Lu & Bol, 2007). Another study controlled for gender and post-intervention scores (Peterson & Peterson, 2011) and the other two studies did not report any control (Güler, 2017; Vanderhoven et al., 2015). Therefore, Omelicheva (2005) has the strongest experimental design. Given the characteristics of the studies, there are mixed results and it is not possible to establish a clear direction of the effect of anonymity from the analyzed studies.

RQ1d: What are the effects of anonymity in peer assessment on social and interpersonal variables?

Only three studies explored the impact of anonymity on social and interpersonal variables. This is in itself an interesting result because a large proportion of the claims for anonymous peer assessment are based on its positive effects on social and interpersonal variables, yet empirical evidence is very limited. One of the three studies did not find social differences using a sociograms methodology (Yu & Sung, 2015). A second study did not find differences in peer pressure, but assessors in the anonymous

condition felt more comfortable (Raes, Vanderhoven, & Schellens, 2013). Finally, Vanderhoven et al. (2015) found that participants in the anonymous condition experienced less peer pressure and fear of disapproval. Therefore, when it comes to the self-reported social aspects, anonymity seems to improve these (e.g., more comfort), but a more objective social measurement (sociograms) did not support that direction.

Exploring the studies constellations variables (Table 6): (a) all of the studies provided the same peer assessment aids (i.e., rubric/criteria & training); (b) two studies included peer grading; (c) two studies included a high number of peer assessments (around 24 & 30 assessments; Raes et al., 2013; Vanderhoven et al., 2015) the other study did not report the number of peer assessments (Yu & Sung, 2015); (d) two of the studies used group work as a task and the third study used individual tasks (questions generation); and (e) two of the studies used a similar length for the intervention (Raes et al., 2013; Yung & Sung, 2015) and the other study did not report the duration (Vanderhoven et al., 2015). It is then difficult to have a clear conclusion about constellations variables effects, as for example, the two more positive for anonymity were longer and more intensive, but they also included peer grading which might be a stressor for participants.

>>> Table 6 about here <<<

Regarding their experimental designs, all of the studies were quasi-experimental with no control over variables to assure comparability among groups. Therefore, there is a threat to validity as the comparability among conditions was not checked for.

RQ1e: What are the effects of anonymity on students' perspective on peer assessment?

Nine studies investigated at least one measure on self-reported students' perspectives. Most studies measured multiple perspectives, including

attitudes/perceptions towards peer assessment, perceived fairness of peer assessment, perceptions of learning activity/classroom climate, perceptions of assessor, psychological safety, etc. (see Table 7). Only three out of the nine studies reported no significant effect of anonymity on any measure (Bloom & Hautaluoma 1987; Yu, 2012; Güler, 2017). Considering the significant results (see Table 7), anonymity appears to have a positive effect when the measured students' perspective relates to the peer assessment or the learning activity, while it seems to have a negative effect on factors related to interpersonal characteristics in peer assessment (e.g., perceptions of assessor, fairness, psychological safety, fear of disapproval). That is, when these variables were compared, students in the anonymous condition seemed to report fewer positive perspectives than those in the non-anonymous condition when the measured variables were related to others in the peer assessment process. The most commonly measured variable is (positive) attitudes or perceptions towards peer assessment (e.g., "I liked assessing my peer during oral peer feedback"; Raes et al., 2013). Three studies used the same scale (i.e., positive attitudes towards peer assessment), and only two of them found a positive effect for anonymity (Raes et al, 2013; Vanderhoven et al., 2015), while Güler (2017) did not. However, in the latter the students in the anonymous condition saw the provided peer feedback by the non-anonymous group, unlike the other studies using the same scale (i.e., Raes et al, 2013; Vanderhoven et al., 2015). A fourth study that used a different scale also reported no significant effect (Yu, 2012).

>>> Table 7 about here <<<

Regarding the constellations variables, all studies except one (Bloom & Hautaluoma 1987), used at least one peer assessment aid (rubric, criteria, and/or training). One study suggested that training addressing the importance of formative peer assessment prior to peer assessment has a better effect on perceptions than anonymity

(Li, 2017). The duration of the peer assessment activity in most of the studies ranged between three to seven weeks. Additionally, anonymity was bidirectional in one study only (Yu, 2012) and unidirectional (known identity of assessee) in the remaining studies. Therefore, the constellations variables were relatively homogeneous among this group of studies.

In terms of the design, two studies implemented experimental designs (Bloom & Hautaluoma 1987; Lin, 2018), one study used single group repeated measures design (Rotsaert et al., 2018), and the rest of the studies relied on quasi-experimental between subjects design. Only three studies considered control variables: gender (Bloom & Hautaluoma 1987; Yu & Wu, 2011) and pre-intervention perceptions/attitudes (Yu & Wu, 2011; Yu, 2012). The sample sizes for studies ranged from N = 32 to 243 participants.

RQ2: Are these observed effects moderated by the participants' educational level in the intervention, presence of peer grading, use of peer assessment aids, or anonymity type?

In terms of the different educational levels, there seems to be an alignment of research results between the two primary, the two secondary and the ten higher education studies (See Table 8). Importantly, because of the nature of this review and the low number of studies that have explored anonymity empirically at this point, this result is preliminary and will need to be further tested in the future. According to the limited existent evidence –only two studies in primary and two in secondary education-, it seems that anonymity might have more positive results in higher education.

>>> Table 8 about here <<<

Regarding the rest of the explored moderating variables (presence of peer grading, assessment aid, & direction of anonymity), when peer grading is implemented

a negative effect of anonymity is likely to be found, same as with a higher number of peer assessment aids (See Table 8). The type of anonymity implemented does not seem to make any difference in the results.

Discussion

The aim of this review was to explore the effects of anonymous peer assessment –i.e. and similar concepts as peer feedback, peer grading etc.- on five outcome variables and how these effects might be moderated by four variables. A narrative review approach was used to explore the effects of anonymity. As has been shown, there is a considerable range of variety in the constellations variables of peer assessment and the vast majority of the studies adopted quasi-experimental designs. Needless to say, the number of the existing empirical studies is low when it comes to the effects of anonymity of different peer assessment in outcomes (except for perceptions of peer assessment). Therefore, drawing conclusions from the presented studies should be done carefully.

First, regarding the effects of anonymity on academic performance the results are mixed: one study showed better results for anonymity, another for anonymity and non-anonymity+training, and the third showed no significant results. Due to the small number of studies on this topic –three- we could not identify clear patterns for peer assessment constellations other that all of them use some peer assessment aid (e.g., training). Nevertheless, a very preliminary conclusion could be that anonymity might have better effects on performance as in none of the three studies was the anonymity group outperformed by the non-anonymous group. A more solid conclusion is that the low number of studies investigating this effect seems to reflect a lack of interest or vision in the field that anonymity might affect more than just the social aspects of peer assessment. It is crucial for future studies to include academic performance as one of

their outcome variables to better understand the effect of anonymity on learning from peer assessment.

Second, regarding peer feedback content and despite the differences in peer feedback content operationalization, there seems to be two general tendencies. Students in anonymous conditions tend to, first, deliver more critical peer feedback and, second, deliver different types of peer feedback compared to students in non-anonymous conditions. One explanation for anonymity allowing students to be more critical is that they might not fear any retaliation, and therefore they might be more willing to point out weaknesses. Exploring the effects of anonymity in the quality of peer feedback is crucial as it has been found that it can be of similar quality to teachers' feedback, that this quality can be increased by using scaffolding interventions, or that the conditions surrounding the intervention (i.e., peer assessment constellation) are key (Panadero, Jonsson, & Alqassab, 2018). Therefore, the need to continue exploring it in further research.

Third, when it comes to peer grading accuracy, two studies found that participants in anonymous conditions provided lower scores (Lu & Bol, 2007; Peterson & Peterson, 2011). This finding suggests that the participants in the anonymous condition were more critical in a similar manner to its effect on peer feedback content. Additionally, while one study found that the grades provided by anonymous participants were closer to teachers' grades (Güler, 2017), another found the contrary (Vanderhoven et al., 2015) but the experimental procedure in the latter presents a threat to validity. Nevertheless, a meta-analysis by Li et al. (2015) found that non-anonymous peer rating was highly correlated to teachers' rating, suggesting and together with our mixed results this suggests that non-anonymity is better to increase students' peer grading accuracy when compared to teachers' assessment.

It is important to highlight two aspects here. Firstly, as discussed above almost all of the analyzed studies (i.e., four out of five) used peer assessment aids which might have influenced the accuracy of peer assessment in all conditions. For example, research on peer assessment with grading aids such as rubrics showed that this type of aids can increase accuracy (Panadero, Romero, & Strijbos, 2013). Therefore, the combination of anonymity and grading aids might be crucial to improve accuracy. Secondly, the effects of holding assessors accountable for their peer grading accuracy should be considered. If there are no repercussions for assessors' lack of accuracy (e.g., overinflating) in the case of non-anonymous peer assessment, s/he could lean towards overscoring because there is nothing to be won from being accurate on the expense of looking harsh to the assessee. This issue needs to be explored in future research because it could be a key explanation for the peer assessment scoring deviation.

Fourth, regarding social effects of anonymity, there seem to be a slight vantage for the use of anonymity in the self-reported social effects such as comfort or less peer pressure (Raes et al., 2013; Vanderhoven et al., 2015). However, a more objective social measure (sociograms) showed no significant differences (Yu & Sung, 2015). Importantly, none of these studies implemented control variables to compare among the groups which is a crucial limitation. These findings are appealing because one of the main theoretical claims for using anonymity is that it might decrease negative social effects (Cheng & Warren, 1997; Ainsworth et al., 2011). However, the existent empirical evidence analyzed here does not fully support this claim. Or alternatively, the lack of anonymity might not always create the assumed social struggles. Due to absence of control baseline measures in the analyzed studies, it is difficult to disentangle this effect. Additionally, we need stronger evidence on whether anonymity is the solution for undesirable social effects such as peer pressure. This can only be revealed using

stronger research designs that take into account the peer assessment constellation variables and use more objective measurements than self-reported social variables.

Fifth, regarding the effects of anonymity on students' perspective on peer assessment, the results, although mixed, suggest that this effect depends on the target of the measured perspective. Anonymity seems to have a positive effect on students' perceptions or attitudes related to the peer assessment activity, but a negative effect on those related to others in the peer assessment activity (e.g., perceptions of assessor, psychological safety). This finding is crucial given that one of the main claimed purposes of using anonymity in peer assessment is to reduce the negative effects resulting from the interpersonal nature of peer assessment (Cheng & Warren, 1997; Vanderhoven et al., 2015; Rotsaert et al., 2018) so that the assessor and the assessee focus on the peer assessment product/output and not on the person behind it (Panadero, 2016).

Investigating the effects of the four moderating variables (i.e., education level, presence of grading and assessment aids, and direction of anonymity) revealed some preliminary effects for these variables. There seems to be a more positive effect on using anonymity when employed in higher education context which could be a result of the higher grade and competitiveness of that educational level compared to school education. However, given the low number of peer assessment studies conducted with school students this pattern remains questionable. Future studies should consider empirically testing how the educational level (school vs. higher education) might moderate the effect of anonymity on different peer assessment outcomes.

Additionally, we found a pattern indicating that the use of peer grading or additional assessment aids are more likely to be associated with negative effects of anonymity. This is not surprising for the assessment aids given that the students might

not need anonymity when enough support is provided for them. For instance, when training was provided to students in non-anonymous peer assessment condition in Li's (2017) study the students outperformed those in the anonymous condition. One would assume that anonymity would have more positive effect in the presence of grading. However, the suggested tendency in this review was in the other direction. One reason for this might be because most of the studies included in this review that used peer grading also implemented some sort of assessment aid. Thus, it is difficult to disentangle the moderating effects of peer grading and assessment using the data provided in this review. Empirical studies are required to investigate the interaction between these two constellation variables in relation to the effects of anonymity on peer assessment outcomes.

Some considerations

At this point, it is essential to reconsider whether concealing the identity of the assessor/assessee is always the right solution to ameliorate students' interpersonal experiences in peer assessment. Strijbos et al. (2009) already questioned the role of anonymity in formative peer assessment activities that focus on supporting students' cognitive, interpersonal, and self-regulation skills. According to the authors, it is important that the assessor and the assessee are known and that they even interact when there is critical feedback and negative affect, so that via the interaction these tensions can be cleared out. A review also pointed out that formative peer assessment interventions that usually require more interaction might have better social and interpersonal results (Panadero, 2016).

Some of the observed lack of positive effects of anonymity can be explained by the literature in computer mediated communications, as some theories point out the negative effects of anonymity. Howard, Barrett, and Frick (2010) argued for three main

aspects. First, according to them and following the social network theory, members in a community help other members to reach their goals, thus breaking these connections via anonymity might come with the cost of decreasing the quality of learning. Second, anonymity in computer mediated communications results in conformity, anti-social online behavior, and other de-individuating effects. Finally, the Social Identity De-individuation Effect explores the cost of anonymity, for example, higher group conformity, members being less motivated, or even anti-social communication, all of these being effects teachers do not want for their groups.

It is also unclear whether concealing the identity of the assessor/assessee really prevents the learner from thinking about and constructing the identity of their counterparts using other cues in the learning environment especially when the focus of peer assessment is not merely providing grades. Indeed, the Interactional Framework of feedback by Strijbos and Müller (2014), postulates that during feedback provision and reception the representations of the assessor/assessee (i.e., how each perceives his/her peer) are activated and they influence the composition of the feedback message as well as its utilization by the recipient. Thus, when it comes to interpersonal factors in peer assessment, then the person can be expected to trust feedback and evaluations from someone they know rather than from someone with an *assumed* level of knowledge or interpersonal skills.

Additionally, all of the analyzed studied (except one) used peer assessment aids including criteria, rubric, and/or training. Thus, the effect of anonymity on students' perspectives on peer assessment might be moderated by the effect of peer assessment aids. Yet, this should be empirically tested in future research. Importantly, Li (2017) showed that an additional training focusing on promoting formative purpose of peer assessment had a better effect on students' perceptions of peer assessment than

anonymity, supporting the argument that students should be prepared for the interpersonal nature of peer assessment instead of being protected from it.

A final crucial point is that peer assessment might need accountability pressure and some freedom for the assessor to be honest and veridical. Ideally, anonymity should provide such freedom, but the lack of accountability might counteract the previous effect. The reason is simple: if the assessors do not have any responsibility on providing veridical information, why would they risk giving feedback that might be received by the assessee as harmful? (Panadero, 2016). This combination of lack of accountability pressure and freedom might be influencing the interpersonal effects of peer assessment and, finally, its effects. Therefore, future interventions need to address the 'weakness' of anonymity via other implementation conditions (e.g., teacher oversight of comments, reviewers getting graded by authors for the helpfulness of the comments) OR address the weaknesses on non-anonymity (e.g., training/community building to support openness of feedback), then the contrasts of anonymous vs. non-anonymous can be further pushed around.

General conclusions and directions for future research

The findings of the analyzed studies on anonymity in peer assessment revealed four main issues. First, the empirical evidence on peer assessment anonymity suggests that the effect of anonymity on different peer assessment outcomes is far more complex than it has been assumed. For three out of our five observed variables (i.e., performance, peer feedback content, and social effects), the results are mixed with a slight tendency towards anonymity. For peer grading accuracy, an emergent trend seems to exist (i.e., lower grades awarded in anonymous activities) but this preliminary conclusion is only based on two studies and, when put in light of the stronger evidence from a previous meta-analysis showing that anonymity is worse for accuracy (Li et al., 2015), this

finding might be negative as anonymous assessors might be grading harder but not producing accurate assessment. It is only in the area where there has been more research (i.e., perspectives on peer assessment) that a clearer pattern can be identified with students' perceptions related to others in the peer assessment activity being adversely influenced by anonymity and those related to learning and peer assessment activities positively influenced.

Therefore, though our findings should be regarded as preliminary, in general anonymity seems to have mixed results when supporting peer assessment interventions outcomes. Importantly, our results as well as previous discussions on this topic suggest that non-anonymous versions of peer assessment might be needed for deeper formative interventions (Strijbos et al., 2009), that formative peer assessment practices seem to be better for interpersonal factors (Panadero, 2016), and that anonymity does not seem to improve peer grading accuracy either (Li et al., 2015) probably because students are not held accountable for their actions. Accordingly, anonymity should not be regarded as the solution to peer assessment implementation problems.

Second, research using anonymity seems to overlook the role of peer assessment constellations. Implementing peer assessment involves various options in terms of the design of the peer assessment activity that the available studies barely reflected upon. When a teacher implements peer assessment, the presence or absence of anonymity is likely to be affected by other factors such as, level of education, whether the assessment includes peer grading, the length of the intervention, the difficulty of the task, the number of occasions in which peer assessment will be delivered or received, etc. Thus, future research should start reporting all the constellation variables as suggested by Topping (1998), and most importantly take them into consideration during intervention designs so that later research can better replicate and interpret the results.

In considering the constellations, it would be important to take into account that privacy is only one of the components of the interaction between peers. According to Gielen et al. (2011), privacy has three dimensions: (a) anonymity of assessor/assessee, (b) teacher presence, and (c) whether the output of peer assessment is confidential or public. Thus, privacy includes more aspects other than whether the assessor and assessee are anonymous, though it is anonymity that has been studied in more detail. Therefore, the effects of the other two components need to be explored.

Third, many of the studies' results are affected by the social and learning context that are not controlled for, as many of the available studies are quasi-experimental. It is already problematic that the number of experimental designs used is low, yet it is even more problematic that many of the quasi-experimental designs do not use strong control variables.

Fourth, there is the classical tension between "*laboratory*" studies in which only one or two variables are manipulated where we would obtain more knowledge about how the individual variables modify the effects (e.g., just anonymity), and *naturalistic* studies in which peer assessment is implemented in the classroom with many variables playing a role, but with difficulties to disentangle different effects. We propose here two solutions. First, laboratory studies need to manipulate more variables in combination with anonymity. For example, anonymity x peer grading, anonymity x gender, anonymity x training, anonymity x rubrics, etc. Secondly, the naturalistic studies should, first, always report the results using Topping's constellation taxonomy so that it is easier to replicate the study and stronger conclusions can be made in reviews like ours. Additionally, stronger designs with more control for confounding variables should be adopted by classroom studies.

Limitations and educational implications

Although this review systematically investigated the effect of anonymity on different peer assessment outcomes, it has four main limitations. First, our conclusions are rather preliminary due to the low number of the available empirical studies currently. Second, we included a number of studies that used quasi-experimental designs with no control variables or single group designs. However, given the nature of studies conducted currently in the field and the small number of empirical studies those studies provided useful insights on anonymity and guidelines for future research. Third, in addition to the in-depth analysis of the studies results, for the second research question we used vote counting which is a technique with limitations when extracting conclusions as compare with stronger methods -i.e. meta-analysis- (Cooper, 2010 p. 157-158). This was done due to the inability of sufficient studies to conduct a meta-analysis. And, fourth, while exploring the moderating effect of the purpose of assessment (formative or summative) would have been very relevant, it was not possible because of lack of information in the included publications. Most papers do not explicitly state the purpose of peer assessment. While the use of peer grading could be a key feature to consider the peer assessment as summative, some studies employ peer grading formatively (e.g., using iterative assessment).

The general educational implication extracted from this review is that anonymity should be implemented with care and knowing that is not the panacea for peer assessment complexity. If the teachers want their students to learn from peer assessment, then allowing non-anonymous face to face interactions might better enrich the process. This formative use in combination with training and practice –i.e., prolonged interventions- can reduce the interpersonal challenges associated with peer assessment. Finally, even if the main intention is to use peer assessment for grading

purposes, then non-anonymous peer assessment produces closer results to teachers' scores Li et al. (2015).

Conclusion

Our review has shown mixed results for anonymity in peer assessment interventions. One of the main factors influencing these results is that anonymity has been approached as the solution to a quite complex challenge: the interpersonal effects in peer assessment. Obviously, asking peers to asses each other usually comes with tolls to pay as performing such assessment generates a number of cognitive, motivational, and emotional tensions. Expecting anonymity to ameliorate all of that on its own is most likely not realistic. Yet, most of the current peer assessment research has adopted a simplistic approach to investigating anonymity. Therefore, we need to start exploring anonymity in interaction with other crucial variables (e.g. aids, length of intervention, practice) as these interactions might have a stronger impact on peer assessment outcomes (Evans, 2013; Topping, 2010). If we continue oversimplifying anonymity impact ignoring its interactions with other constellations variables then we will lose information about the complex dynamics that seem to be in place. It is then time to bring our anonymity research to the next level and take it to more complex, and finally, higher places.

References

(Marked with an * the studies included in the review).

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. doi:10.1080/07294360.2017.1405913

Ainsworth, S., Gelmini-Hornsby, G., Threapleton, K., Crook, C., O'Malley, C., & Buda, M. (2011). Anonymity in classroom voting and debating. *Learning and*

Instruction, 21(3), 365-378. doi:http://dx.doi.org/10.1016/j.learninstruc.2010.05.001

* Bloom, A. J., & Hautaluoma, J. E. (1987). Effects of message valence, communicator credibility, and source anonymity on reactions to peer feedback. *The Journal of Social Psychology, 127*(4), 329-338. doi:10.1080/00224545.1987.9713712

Cheng, W., & Warren, M. (1997). Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233-239. doi:10.1080/03075079712331381064

Cooper, H. (2010). *Research synthesis and meta-analysis*. Thousand Oaks, California: SAGE.

Dochy, F., Segers, M., & Sluijsmans, D. (1999). The use of self-, peer- and co-assessment in higher education. A review. *Studies in Higher Education, 24*(3), 331-350. doi:10.1080/03075079912331379935

Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research, 83*(1), 70‑120. doi:10.3102/0034654312474350

Gielen, S., Dochy, F., & Onghena, P. (2011). An inventory of peer assessment diversity. *Assessment & Evaluation In Higher Education, 36*(2), 137-155. doi:10.1080/02602930903221444

* Güler, C. (2017). Use of WhatsApp in higher education: What's up with assessing peers anonymously? *Journal of Educational Computing Research, 55*(2), 272-289. doi:10.1177/0735633116667359

* Howard, C. D., Barrett, A. F., & Frick, T. W. (2010). Anonymity to promote peer feedback: Pre-service teachers' comments in asynchronous computer-mediated communication. *Journal of Educational Computing Research, 43*(1), 89-112. doi:10.2190/EC.43.1.f

Li, H., Xiong, Y., Zang, X., Kornhaber, M. L., Lyu, Y., Chung, K. S., & Suen, H. K. (2015). Peer assessment in the digital age: A meta-analysis comparing peer and teacher ratings. *Assessment & Evaluation In Higher Education,* 1-20. doi:10.1080/02602938.2014.999746

* Li, L. (2017). The role of anonymity in peer assessment. *Assessment & Evaluation In Higher Education,* 1-12. doi:10.1080/02602938.2016.1174766

* Lin, G. Y. (2018). Anonymous versus identified peer assessment via a Facebook-based learning application: Effects on quality of peer feedback, perceived learning, perceived fairness, and attitude toward the system. *Computers & Education, 116,* 81-92. doi:https://doi.org/10.1016/j.compedu.2017.08.010

Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290.

* Lu, R., & Bol, L. (2007). A comparison of anonymous versus identifiable e-peer review on college student writing performance and the extent of critical feedback. *Journal of Interactive Online Learning, 6*(2).

* Omelicheva, M. Y. (2005). Self and peer evaluation in undergraduate education: Structuring conditions that maximize its promises and minimize the perils. *Journal of Political Science Education, 1*(2), 191-205. doi:10.1080/15512160590961784

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 Human and Social Conditions in Assessment* (pp. 247-266). New York: Routledge.

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*: Cambridge University Press.

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74-98. doi:https://doi.org/10.1016/j.edurev.2017.08.004

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on construct validity of peer assessment, perceived fairness and comfort, and performance. *Studies In Educational Evaluation, 39*(4), 195-203. doi:10.1016/j.stueduc.2013.10.005

* Peterson, C. H., & Peterson, N. A. (2011). Impact of peer evaluation confidentiality on student marks. *International Journal for the Scholarship of Teaching and Learning, 5*(2).

* Raes, A., Vanderhoven, E., & Schellens, T. (2013). Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. *Studies in Higher Education*, 1-16. doi:10.1080/03075079.2013.823930

* Rotsaert, T., Panadero, E., & Schellens, T. (2018). Anonymity as an instructional scaffold in peer assessment: Its effects on peer feedback quality and evolution in students' perceptions about peer assessment skills. *European Journal of Psychology of Education, 33*, 75–99. doi:10.1007/s10212-017-0339-8

Sluijsmans, D. M. A., Brand-Gruwel, S., van Merriënboer, J. J. G., & Martens, R. L. (2004). Training teachers in peer-assessment skills: Effects on performance and perceptions. *Innovations in Education and Teaching International, 41*(1), 59-78. doi:10.1080/1470329032000172720

Strijbos, J. W., & Müller, A. (2014). Personale faktoren im feedbackprozess. In H. Ditton & A. Müller (Eds.), *Feedback und rückmeldungen: Theoretische grundlagen, empirische befunde, praktische anwendungsfelder* [Feedback and evaluation: Theoretical foundations, empirical findings, practical implementation] (pp. 87–134). Münster, Germany: Waxmann.

Strijbos, J. W., Ochoa, T. A., Sluijsmans, D. M. A., Segers, M., & 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). Hersey, PA: IGI Global.

Topping, K. J. 2010. Methodological Quandaries in Studying Process and Outcomes in Peer Assessment. *Learning and Instruction* 20 (4):339‑343. doi:10.1016/j.learninstruc.2009.08.003

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276.

van Gennip, N. (2012). *Assessing together. Peer assessment from an interpersonal perspective.* (PhD), Universiteit Leiden.

van Gennip, N., Segers, M., & 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. doi:10.1016/j.edurev.2008.11.002

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

* Vanderhoven, E., Raes, A., Montrieux, H., Rotsaert, T., & Schellens, T. (2015). What if pupils can assess their peers anonymously? A quasi-experimental study.

Computers & Education, 81, 123-132. doi:http://dx.doi.org/10.1016/j.compedu.2014.10.001

* Yu, F. Y. (2012). Any effects of different levels of online user identity revelation? *Journal of Educational Technology & Society, 15*(1), 64-77.

Yu, F. Y., & Liu, Y. H. (2009). Creating a psychologically safe online space for a student-generated questions learning activity via different identity revelation modes. *British Journal of Educational Technology, 40*(6), 1109-1123. doi:10.1111/j.1467-8535.2008.00905.x

* Yu, F. Y., & Sung, S. (2015). A mixed methods approach to the assessor's targeting behavior during online peer assessment: Effects of anonymity and underlying reasons. *Interactive Learning Environments*, 1-18. doi:10.1080/10494820.2015.1041405

* Yu, F. Y., & Wu, C. P. (2011). Different identity revelation modes in an online peer-assessment learning environment: Effects on perceptions toward assessors, classroom climate and learning activities. *Computers & Education, 57*(3), 2167-2177. doi:10.1016/j.compedu.2011.05.012

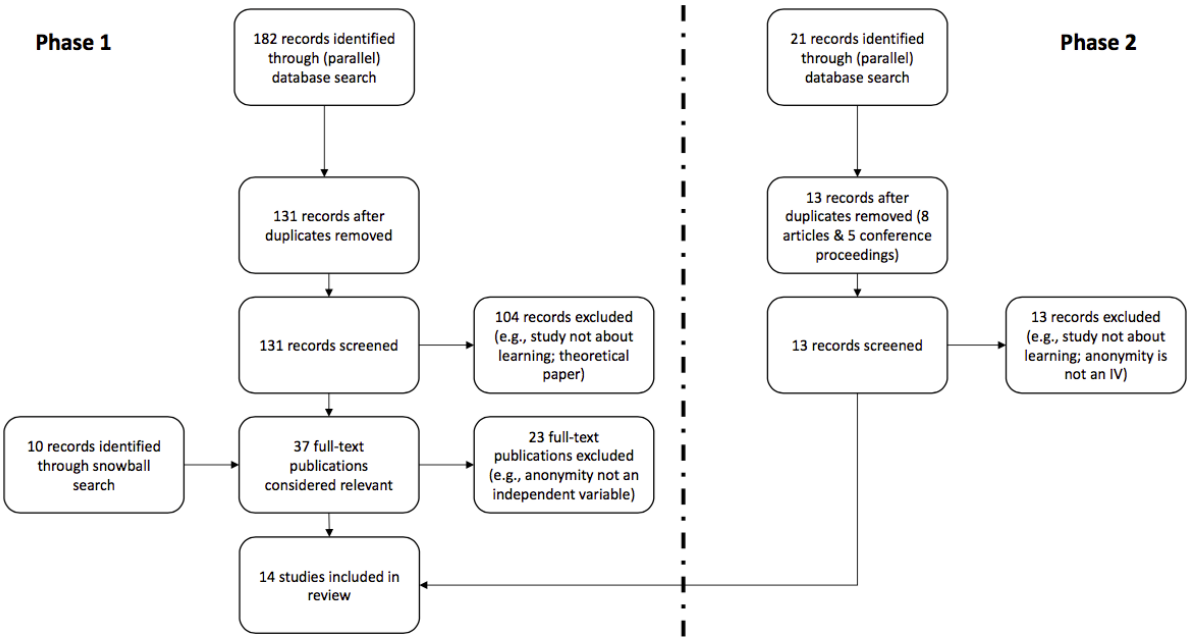

Figure 1. A flow chart showing the number of records identified within the two phases of the performed literature search in PsycINFO and ERIC

Table 1

Summary of aim, design, terminology (of PA used), grades (inclusion in PA), PA aids (use), type (of anonymity), and results and conclusions of the studies included in the review

Study	Aim	Design (a) & Terminology (b)	PA grading (a), PA aids (b), & type (c)	Results & conclusions
Bloom & Hautaluoma 1987	To investigate how intentions to improve and affective reactions to simulated peer feedback received via superiors is influenced by (a) valence of peer feedback, (b) credibility of supervisor, and (c) anonymity/identity of peer sources	(a) Experimental factorial design (2 x 2 x 2); (b) Peer feedback	(a) No; (b) No; (c) Unidirectional (Identity of fictional peer assessor)	No significant effect of anonymity was found on intentions to improve or affective reactions. Positive peer feedback (valence) leads to more positive reactions, but negative peer feedback leads to greater intentions to improve. **Direction of results: 1 neutral**
Omelicheva, 2005	To examine the influence of (a) presence/absence of assessment criteria, (b) anonymity/non-anonymity, and (c) trigger of students' motivation (strong/weak) on the accuracy of self and peer marking	(a) Experimental factorial design (2 x 2) – two experiments; (b) Peer assessment	(a) Yes; (b) No; (c) Bidirectional	Experiment 1: Anonymous PA was more reliable than non-anonymous PA. Experiment 2: No significant differences were found in PA reliability between the anonymous and non-anonymous conditions. **Direction of results: ½ positive for anonymity, ½ neutral**
Lu & Bol, 2007	To examine the impact of anonymous peer review on (a) writing performance and (b) critical peer feedback	(a) Quasi-experimental between-subjects design – two studies; (b) Peer review	(a) Yes; (b) Yes-training; (c) Bidirectional	Students in the anonymous condition performed significantly better in the post-test than students in the non-anonymous condition. Anonymous group provided significantly more negative comments and lower ratings than the non-anonymous group. **Direction of results: 1 positive for anonymity**
Howard et al., 2010	To investigate the impact of anonymity on peer feedback frequency, nature and pattern	(a) Quasi-experimental between-subjects design; (b) Peer feedback	(a) No; (b) No; (c) Bidirectional	Anonymous peers wrote significantly more comments, more negative comments, and more irrelevant comments. The probabilities of critical comments followed by positive reactions, and negative reactions followed by suggestions were significantly higher for the anonymous condition. Overall, anonymous students seemed to provide more critical feedback. **Direction of results: 1 positive for anonymity**
Peterson & Peterson, 2011	To investigate the impact of confidentiality on PA of team work, and how PA can influence the grades individuals receive on group projects.	(a) Quasi-experimental between-subjects design; (b) Peer evaluation, peer rating	(a) Yes; (b) Yes – evaluation form; (c) Unidirectional (identity of assessor)	Confidentiality in PA results in lower peer ratings and lower overall individual mark. **Direction of results: ½ positive for anonymity** (half point though they did not check the direction against the teacher's scores because the majority of research shows students tend to overscore when peer grading)

Study	Aim	Design (a) & Terminology (b)	PA grading (a), PA aids (b), & type (c)	Results & conclusions
Yu & Wu, 2011	To investigate the impact of identity revelation (real name / anonymity / nickname / self-choice) on (a) perceptions towards assessor, (b) perceptions of classroom clime, and (c) attitudes towards learning activity	(a) Quasi-experimental between-subjects design (4 conditions); (b) Peer assessment	(a) Yes – Ratings; (b) Yes – criteria & training; (c) Bidirectional	Different identity modes lead assesses to view their assessors differently. Participants in the real-name and self-choice conditions perceived their assessors significantly more positively than those in the anonymous and nickname conditions. **Direction of results: 1 negative for anonymity**
Yu, 2012	To investigate the impact of different levels of identity revelation (real name / nickname / anonymity) on (a) academic performance, (b) attitudes towards PA, (c) perceptions towards interacting with assessors and towards the learning activity and PA	(a) Quasi-experimental between-subjects design (3 conditions); (b) Peer assessment	(a) No; (b) Yes – criteria & training; (c) Bidirectional	No evidence was found to support that various identity revelation modes differentially influenced academic performance, attitudes towards PA, perceptions towards interacting with assessors, or perceptions towards the learning activity and PA. **Direction of results: 1 neutral**
Raes et al., 2013	To investigate the effect of increasing levels of anonymity on (a) perceived peer pressure, (b) comfort, and (b) perceptions towards PA	(a) Quasi-experimental mixed-design; (b) Peer assessment	(a) Yes; (b) Yes – rubric & training; (c) Unidirectional (identity of assessor)	No evidence was found that less peer pressure was felt while providing anonymous PA compared to public oral feedback. Students felt more comfortable providing anonymous PA. Students perceived greater added value of providing written feedback following anonymous PA and non-anonymous oral feedback when the written feedback was anonymous. **Direction of results: ½ neutral, ½ positive for anonymity**
Vanderhoven et al., 2015	To investigate the impact of providing anonymity during PA on (a) undesirable social effects, (b) validity of PA, and (c) teachers' experiences of PA	(a) Quasi-experimental between-subjects design (2 conditions); (b) Peer assessment	(a) Yes; (b) Yes – rubric & training; (c) Unidirectional (identity of assessor)	Students in the anonymous condition experienced significantly less peer pressure and less fear of disapproval, and they reported more positive attitudes towards PA. No differences were found in levels of discomfort between the anonymous and the non-anonymous conditions. Non-anonymous PA is more valid than anonymous PA when compared to teacher assessment. **Direction of results: ½ positive for anonymity, ½ negative for anonymity**
Yu & Sung, 2015	To investigate the effect of identity revelation (real-name vs. anonymous) on online interaction behaviour and social dynamics of the group (measured by frequency of PA)	(a) Repeated measures design; (b) Peer assessment	(a) No; (b) Yes – marking criteria & training; (c) Bidirectional	No evidence was found that anonymity increased interactions between peers (measured by frequency of PA) **Direction of results: 1 neutral**
Güler, 2017	To examine the use of WhatsApp application as an anonymous vs. non-anonymous PA tool on (a) perceived fairness and (b) positive attitude towards PA, and (c) accuracy of PA	(a) Quasi-experimental between-subjects design (2 conditions); (b) Peer assessment	(a) Yes; (b) Yes – rubric & training; (c) Unidirectional (identity of assessor)	No evidence was found that anonymity had an effect on perceived PA fairness, or on positive attitude towards PA. **Direction of results: 1 neutral**

Study	Aim	Design (a) & Terminology (b)	PA grading (a), PA aids (b), & type (c)	Results & conclusions
Li, 2017	To investigate the effect of anonymity and training (as an alternative in non-anonymous situations) on (a) performance and (b) perceptions of PA	(a) Quasi-experimental between-subjects design (3 conditions); (b) Peer assessment, peer review	(a) Yes; (b) Yes - training & marking criteria; (c) Bidirectional	Students in the anonymous and non-anonymous training conditions performed significantly better than students in the non-anonymous condition. Students in the non-anonymous training condition perceived significantly less pressure and more value/usefulness of PA compared to students in the anonymous and the non-anonymous conditions. **Direction of results: ½ neutral, ½ negative for anonymity**
Lin, 2018	To investigate the effects of anonymity on (a) distribution of peer feedback types (affective, cognitive & metacognitive) and (b) perceived learning, perceptions, and attitudes towards Facebook-based learning application	(a) Experimental between-subjects design (2 conditions); (b) Peer assessment, peer feedback	(a) No; (b) Yes – criteria; (c) Unidirectional (identity of assessor)	Anonymity can lead to providing more cognitive and less affective peer feedback comments. Students in the anonymous condition reports significantly more perceived learning, less fairness of PA, and more positive attitudes towards the learning application compared to the non-anonymous condition. **Direction of results: ½ positive for anonymity, ½ neutral**
Rotsaert et al., 2018	To investigate the impact of gradual fading of anonymity on (a) quality of peer feedback, (b) perceived improvement of peer feedback skills, and (c) perceptions related to PA (importance of anonymity, interpersonal variables, conceptions of PA)	(a) Quasi-experimental repeated measures design (anonymous phase: 2 sessions; non-anonymous phase: 2 sessions); (b) Peer assessment, peer feedback	(a) Yes; (b) Yes – rubric & criteria; (c) Unidirectional (identity of assessor)	Peer feedback quality increased in the earlier anonymous phase and remained comparable later in the non-anonymous sessions. Perceived improvement of peer feedback increased significantly from session 1 to session 2 but did not change thereafter. Students' preferences for anonymity significantly decreased after the non-anonymous phase. Significantly more psychological safety was reported after the 1st non-anonymous session. Students' trust in self as assessor improved from 1st to 2nd session but did not change from the anonymous to the non-anonymous phase. Students' fear of disapproval significantly decreased after the non-anonymous sessions. Students' conceptions of PA improved after the anonymous phase but did not improve in the non-anonymous phase. **Direction of results: ½ negative for anonymity, ½ neutral**

Table 2

Description of sample size, educational level, gender distribution, subject domain, and country

Study	Sample size	Ed. level	Gender distribution	Subject domain	Country
Bloom & Hautaluoma 1987	96	Higher Ed.	50% female	Psychology	USA
Omelicheva, 2005	110	Higher Ed.	46% female first study and 40% second	Political Sciences	USA
Lu & Bol, 2007	92	Higher Ed.	Not reported	English	USA
Howard et al., 2010	72	Higher Ed.	74% female	Technology for preservice teachers	USA
Peterson & Peterson, 2011	86	Higher Ed.	88% female	Education/ Research Methods	USA
Yu & Wu, 2011	243	Primary Ed.	51% female	Science	China
Yu, 2012	101	Secondary Ed.	Not reported	Science & Technology	China
Raes et al., 2013	51	Higher Ed.	92% female	Education/ Instructional Design	Belgium
Vanderhoven et al., 2015	69	Secondary Ed.	72% female	Presentation Skills	Belgium
Yu & Sung, 2015	65	Primary Ed.	52% female	Science	China
Güler, 2017	84	Higher Ed.	43% female	Instructional Design & Computer Education	Turkey
Li, 2017	77	Higher Ed.	72% female	Technology Application	USA
Lin, 2018	32	Higher Ed.	72% female	Education	China
Rotsaert et al., 2018	46	Higher Ed.	84% female	Education/ Instructional Design	Belgium

Table 3
Comparison peer assessment anonymity studies effect on performance

Study	Design	PA Aid (a); Pa grading (b)	Number PA	Intervention length	Control variables	Results
Lu & Bol, 2007	Quasi-experimental between subjects intact classroom groups (two conditions)	(a) Training; (b) Yes	Two provided and two received per nine assignments	A semester, eight occasions for rehearsal	SAT, GPA high school, plus counterbalanced scheduling	The anonymous condition outperformed *Performance was measured via two timed essays scored by two professional raters*
Yu, 2012	Quasi-experimental between subjects intact classroom groups (three conditions)	(a) Criteria & training; (b) No	Each participant assessed at least one peer. No clear information on number of assessors per task	Six weeks	Students' scores on the first biology exam were used as covariate	No difference *Performance was measured via two exams (not reported who scored them)*
Li, 2017	Quasi-experimental between subjects intact classroom groups (three conditions)	(a) Training, marking criteria & rubric. One non-anonymous condition extra training on PA concerns and strategies to avoid them; (b) Yes	Provided to two peers. No clear information on number of assessors per task	Multiple sessions (length not specified)	Pre peer assessment scores used as a covariate	The non-anonymous with training and the anonymous conditions outperformed the non-anonymous no training condition *Performance was measured via the WebQuests designed by the students by the instructor and an independent grader*

Table 4
Comparison peer assessment anonymity studies effect on peer feedback content and quality

Study	Design	PA Aid (a), PA grading (b) & Task (c)	Number PA	Intervention length	Control variables	Results
Lu & Bol, 2007	Quasi-experimental between subjects (two conditions)	(a) Training; (b) Yes; (c) Writing	Two provided and two received per 9 assignments	Whole semester, eight occasions	SAT, GPA high school, plus counterbalanced scheduling	The anonymous group provided more frequent negative comments and lower ratings
Howard et al., 2010	Quasi-experimental between subjects (two conditions)	(a) No; (b) No; (c) Website design	Not reported	One lecture (50 minutes)	No	RQ1: both conditions equal number of comments, but anonymous more words and utterances RQ2: both conditions equal number of positive words, but anonymous group more critical feedback and off topic words RQ3: anonymous condition used more two patterns: positive comments followed by critical feedback, and critical feedback followed by suggestion for design change
Lin, 2018	Experimental design between subjects (two conditions)	(a) Criteria; (b) No; (c) Micro-teaching demonstration	Participants assessed 5/6 times and received peer feedback from five peers	Six last sessions. Several occasions to give, one to receive	- (experimental design with randomized assignment of participants)	Anonymous group provided higher cognitive comments (subcategories "vague suggestions" and "extension") and lower affective comments (subcategories "supportive" and "oppositional"). No difference in metacognitive comments
Rotsaert et al., 2018	Quasi-experimental within subjects (one group pre-post comparison no counterbalanced)	(a) Rubrics and criteria; (b) Yes; (c) Conducting workshop	Assessor seven times, assessee one	Four weeks, several occasions of assessor practice	- (only one condition)	The quality of the peer feedback was higher faster in the anonymous phase, but over time, the feedback quality in the non-anonymous sessions was comparable

Table 5
Comparison peer assessment anonymity studies effect on peer grading accuracy

Study	Design	PA Aid (a); PA grading compared (b); Task (c)	Number PA	Intervention length	Control variables	Results
Omelicheva 2005	Experimental factorial design 2 x 2 in both studies	(a) No; (b) Instructor; (c) Essay	Four provided and four received in each experiment	Single occasion in each experiment	- (experimental assigned at random)	Both experiment results did not reach significance. In experiment 1 anonymous was more reliable, in experiment 2 the non-anonymous
Lu & Bol, 2007	Quasi-experimental between subjects (two conditions)	(a) Training; (b) Average of two peers scores; (c) Writing	Two provided and two received per nine assignments	Whole semester, eight occasions	SAT, GPA high school, plus counterbalanced scheduling	Anonymity group awarded lower ratings. Small difference
Peterson & Peterson, 2011	Quasi-experimental between subjects	(a) Evaluation form; (b) *Instructor**; (c) Team work	Provided 2/3 and received depending on the group size. Done twice	Two group work tasks by the end of the semester	Group work scores - but they were post-measured- and the gender	They created a correction system. Anonymous condition was below the corrected group work grade. Non-anonymous condition was above
Vanderhoven et al. 2015	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Teacher; (c) Group oral presentation	Each scored aprox. 10 times, received scores from around 30 peers once	Multiple sessions (length not specified)	No	Strong correlation between students' and teachers' scores in the non-anonymous condition (r = .93); moderate correlation in the anonymous condition (r = .42)
Güler, 2017	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Teacher; (c) Group workshop design	Not mentioned	Last 7 weeks of semester	No	Strong correlation between students' and teachers' scores in the anonymous condition (r = .78); moderate correlation in the non-anonymous condition (r = .48)

* Direct communication with first author.

Table 6
Comparison peer assessment anonymity studies effect on social effects

Study	Design	PA Aid (a); PA grading (b); Task (c)	Number PA	Intervention length	Control variables	Results
Raes et al. 2013	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Yes; (c) Group workshop	Each assessed 7/8 teams and received for their group performance at least 24 evaluations	Three weeks	No	There was no difference between conditions in peer pressure however the anonymous participants felt more comfortable giving feedback
Vanderhoven et al. 2015	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Yes; (c) Group oral presentation	Each scored aprox. 10 times, received scores from around 30 peers once	Multiple sessions (length not specified)	No	Participants in the anonymity condition experience less peer pressure and fear of disapproval *Nevertheless accuracy was lower*. There was no diferences between conditions on assessee's comfort with peer assessment
Yu & Sung, 2015	Quasi-experimental within subjects (One condition pre-post)	(a) Criteria & training; (b) No; (c) Construction of science questions	Not reported but seems like several	Two weeks	- (the two classroom groups were put in the only condition; no counterbalance)	Through the use of sociograms it was found that there were no distinguishable differences between the different identity revelation modes in both classes

Table 7
Comparison studies effect on perceptions

Study	Design	PA Aid (a), PA grading (b) & Task (c)	Number PA (a); Intervention length (b)	Perceptions/ reaction type	Control variables	Results
Bloom & Hautaluoma 1987	Experimental between subjects (eight conditions)	(a) No; (b) No; (c) Fictional work performance (scenario)	(a) Once (scenario); (b) Single occasion	Intentions to improve Affective reactions	Stratified random sampling (equal gender split)	No significant effect of anonymity
Yu & Wu, 2011	Quasi-experimental between subjects design (four conditions)	(a) Training; (b) Yes (rating); (c) Science questions generation	(a) Assessed 2 peers at least; No information on assessors per task; (b) Six weeks	Perceptions towards: assessors, classroom climate, and learning activity	Pre-intervention perceptions used as covariate; Gender differences tested	Participants in the real-name and self-choice conditions perceived their assessors more positively than those in the anonymous and nickname conditions; No effects on perceptions of classroom climate or activity
Yu, 2012	Quasi-experimental between subjects (three conditions)	(a) Criteria & training; (b) No; (c) Science questions generation	(a) Assessed 1 peer at least; No clear information on number of assessors per task; (b) Six weeks	Attitudes towards PA Perceptions towards: assessors, interaction with assessors, and learning activity	Pre-intervention perceptions used as covariate	No significant effects
Raes et al. 2013	Quasi-experimental between-within subjects (two conditions)	(a) Rubric & training; (b) Yes; (c) Group workshop	(a) Each assessed 7/8 teams; Received for their group performance at least 24 evaluations; (b) Three weeks	Positive attitudes towards PA Perceived added value	No	More positive attitudes towards the anonymous PA; Perceived greater added value of providing written feedback following anonymous PA and non-anonymous oral feedback when the written feedback was anonymous
Vanderhoven et al. 2015	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Yes; (c) Group oral presentation	(a) Each scored aprox. 10 times; Received scores from around 30 peers once; (b) Multiple sessions (length not specified)	Positive attitudes towards PA	No	More positive attitudes were reported by the anonymous condition
Güler, 2017	Quasi-experimental between subjects (two conditions)	(a) Rubric & training; (b) Yes; (c) Group workshop design	(a) Not mentioned; (b) Last 7 weeks of semester	Perceived fairness Positive attitudes towards PA	No	No significant effects
Li, 2017	Quasi-experimental between subjects (three conditions)	(a) Training, marking criteria & rubric; (b) Yes; (c) WebQuest	(a) Each provided feedback to 2 peers; No clear information on number of assessors per task; (b) Multiple sessions (length not specified)	Perceived value/usefulness of PA Perceived tension/pressure in the process	No	Students in the non-anonymous training condition perceived less pressure and more value/usefulness of PA compared to students in anonymous and non-anonymous conditions
Lin, 2018	Experimental design between subjects (two conditions)	(a) Criteria; (b) No; (c) Micro-teaching demonstration session	(a) Each assessed 5/6; Received peer feedback from 5 peers; (b) Six last class sessions. Several occasions to	Perceived learning Perceived fairness of PA	Randomized assignment of participants	Students in the anonymous condition reported more perceived learning, less fairness of PA, and more positive attitudes towards the learning

			give, one to receive from five assessors	Perceptions towards online system		application compared to the non-anonymous condition
Rotsaert et al., 2018	Quasi-experimental within subjects (one group)	(a) Rubrics and criteria; (b) Yes; (c) Conducting workshop	(a) Each assessed 7 times; Received one; (b) Four weeks, several occasions of assessor practice	Perceived improvement of peer feedback skills Perceptions related to PA (importance of anonymity, interpersonal variables) Conceptions of PA	Only one condition (pre-post comparison not counterbalanced)	Perceived improvement of peer feedback increased significantly from 1st to 2nd session but did not change thereafter; Students' preferences for anonymity significantly decreased after the non-anonymous phase; More psychological safety was reported after the 1st non-anonymous session; Trust in self as assessor improved from 1st to 2nd session but did not change from the anonymous to the non-anonymous phase; Fear of disapproval decreased after the non-anonymous sessions; Conceptions of PA improved after the anonymous phase but did not improve in the non-anonymous phase

Table 8
Direction of results organized by four moderating variables

Educational level		
Primary	Positive (anonymity)	
	Neutral	1
	Negative (anonymity)	1
Secondary	Positive (anonymity)	½
	Neutral	1
	Negative (anonymity)	½
Higher education	Positive (anonymity)	4
	Neutral	4 ½
	Negative (anonymity)	1
Peer grading		
No	Positive (anonymity)	1 ½
	Neutral	3 ½
	Negative (anonymity)	
Yes	Positive (anonymity)	3
	Neutral	3
	Negative (anonymity)	2 ½
PA aids		
None	Positive (anonymity)	1 ½
	Neutral	1 ½
	Negative (anonymity)	
One	Positive (anonymity)	2
	Neutral	½
	Negative (anonymity)	
Two	Positive (anonymity)	1
	Neutral	3 ½
	Negative (anonymity)	2 ½
Anonymity type		
Unidirectional	Positive (anonymity)	2
	Neutral	3 ½
	Negative (anonymity)	1
Bidirectional	Positive (anonymity)	2 ½
	Neutral	3
	Negative (anonymity)	1 ½