The impact of formative peer feedback on higher education students' academic writing: a Meta-Analysis

Bart Huisman, Nadira Saab, Paul van den Broek & Jan van Driel

© 2018 Informa UK Limited, trading as Taylor & Francis Group

Published online: 29 Dec 2018.

Submit your article to this journal ↗

Article views: 1944

View Crossmark data ↗

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=caeh20

Routledge
Taylor & Francis Group

The impact of formative peer feedback on higher education students' academic writing: a Meta-Analysis

Bart Huisman[a] , Nadira Saab[a] , Paul van den Broek[a] and Jan van Driel[b]

[a]Leiden University, The Netherlands; [b]University of Melbourne, Australia

ABSTRACT

Peer feedback is frequently implemented with academic writing tasks in higher education. However, a quantitative synthesis is still lacking for the impact that peer feedback has on students' writing performance. The current study conveyed two types of observations. First, regarding the impact of peer feedback on writing performance, this study synthesized the results of 24 quantitative studies reporting on higher education students' academic writing performance after peer feedback. Engagement in peer feedback resulted in larger writing improvements compared to (no-feedback) controls ($g = 0.91$ [0.41, 1.42]) and compared to self-assessment ($g = 0.33$ [0.01, 0.64]). Peer feedback and teacher feedback resulted in similar writing improvements ($g = 0.46$ [-0.44, 1.36]). The nature of the peer feedback significantly moderated the impact that peer feedback had on students' writing improvement, whereas only a theoretically plausible, though non-significant moderating pattern was found for the number of peers that students engaged with. Second, this study shows that the number of well-controlled studies into the effects of peer feedback on writing is still low, indicating the need for more quantitative, methodologically sound research in this field. Findings and implications are discussed both for higher education teaching practice and future research approaches and directions.

Introduction

Across disciplines, peer feedback is frequently implemented as an instructional method with academic writing assignments. In part, this is supported by prior qualitative review studies indicating that peer feedback can improve domain specific skills (van Zundert, Sluijsmans and van Merriënboer 2010). Despite a growing body of research however (e.g., Topping 1998; van Gennip, Segers and Tillema 2009; Gielen, Dochy and Onghena 2011; Evans 2013), a quantitative synthesis of the research is still lacking for the impact that peer feedback has on students' academic writing performance. Consequently, the extent to which peer feedback can improve students' writing is still unknown. The current meta-analysis has two central aims. First, it investigates the impact that peer feedback has on students' academic writing performance as compared to two feasible alternatives: self-assessment and feedback from teaching staff. Second, it aims to gain more insight into the role that the design of peer feedback tasks can have on students' learning outcomes. Specifically, it explores the extent to which students' writing

CONTACT Bart Huisman b.a.huisman@iclon.leidenuniv.nl

performance is moderated by two variables that are important for the design and implementation of peer feedback: the nature of the peer feedback and the number of peers engaged with. This way, this study aims to be informative for both academic researchers in the field and higher education teaching staff.

Generally speaking, there are at least two arguments to support the implementation of peer feedback on writing in the higher education context. The first relates to the learning benefits for students. Students may expect reliable assessments from their peers (Falchikov and Goldfinch 2000), although students' conceptions of validity may differ from that of teachers (or researchers) and reliability may be contingent on, for example, the number of peer reviews (Cho, Schunn and Wilson 2006; Cho and Schunn 2007). The very act of providing peer feedback can be beneficial as well (Lundstrom and Baker 2009; Cho and MacArthur 2011; Cho and Cho 2011; Lee 2015), for example because it requires students to actively consider the assessment criteria (Huisman et al. 2018). Moreover, providing and utilizing feedback from peers can be considered an important skill for students' future academic or professional careers, and can therefore be considered an important learning goal within higher education curricula (Liu and Carless 2006). The second argument relates to the logistic and economic benefits of peer feedback, and revolves around the notion that peer feedback can be available in greater volume and with greater immediacy compared to teacher feedback (Topping 2009; Cho and MacArthur 2010). Currently, more than half of the young people in Organisation for Economic Cooperation and Development countries are expected to enrol in a bachelor's program or equivalent at some point in their life (OECD 2016), an upward trend that started over a decade ago. This can affect student-to-teacher ratios and corresponding workloads for academic staff (Ballantyne, Hughes and Mylonas 2002; Bailey and Garner 2010). Especially in the case of feedback on writing, being relatively time-consuming, such pressures on teaching staff increase the need for alternative feedback practices that are both effective and practically efficient.

Prior research

To our knowledge, a quantitative synthesis or meta-analysis for the impact of peer feedback on students' writing performance has not yet been published for the higher education context. As a consequence, the extent to which peer feedback can improve students' writing is still unknown. For adolescent students (Grade 4-12) at least one prior meta-analysis has been conducted (Graham and Perin 2007). As part of a larger focus on writing intervention treatments, this meta-analysis found a strong and positive impact on writing quality when comparing students who were engaged in 'peer assistance' with students who wrote alone (weighted effect size 0.75 [0.54-0.97]). In the Graham and Perin (2007) study, however, peer assistance also included students cooperating in planning, revision and composition phases. Hence, peer assistance reflected a broader set of cooperative activities, of which peer feedback was merely one. This makes it difficult to disentangle the specific effects of peer feedback from the other cooperative learning activities. For the higher education context, a relatively early and often cited qualitative review that partly focuses on peer assessment of writing is that by Topping (1998). Topping concluded that peer assessment appears to yield outcomes that are at least comparable to teacher assessment, but noted that most of the research was descriptive in nature. In particular, he found eleven references that reported specifically on writing outcomes consisting of three peer-reviewed journal articles, six doctoral dissertations, and two conference papers. Given the early stage of the research field and the variance in reported peer feedback practices, Topping (1998) acknowledged it was too early for a best-evidence synthesis or meta-analysis. Despite an increase in research on peer feedback in the thirteen following years, Gielen, Dochy and Onghena (2011) deemed such a synthesis still unfeasible.

Around the year 2010, some qualitative review studies into peer feedback were published. Among others, these reviews have provided descriptive accounts of the effects of peer feedback and updated our knowledge regarding the variables important in designing and implementing peer feedback. In their review on effective peer assessment processes, for example, van Zundert, Sluijsmans and van Merriënboer (2010) investigated which factors and processes influenced three different outcome variables of peer assessment: the psychometric quality of peer assessment, domain-specific skills and peer assessment skills. They concluded that training and experience in peer assessment positively related to all three outcome measures. The majority of the included studies were case studies, interventions were often not described specifically, and specific causal inferences were generally lacking. Therefore, the authors cautioned that the share of (quasi-)experimental studies was small and stressed the need for more controlled studies with specific variable descriptions (see also Topping 1998, 2010; Strijbos and Sluijsmans 2010). What these and other review studies (e.g., van Gennip, Segers and Tillema 2009) have in common, is that they do not focus on one specific object of assessment *within* a particular educational context, such as primary, secondary or higher education. This may not yet have been feasible because of the diversity in reported peer feedback practices in which many factors interrelate (e.g., Gielen, Dochy and Onghena 2011). For example, providing and receiving peer feedback on an oral presentation or on a written essay involves different feedback criteria and interpersonal communication.

As these aspects probably interrelate with students' prior experience and educational level, these determine how and to what extent students need to be trained or guided and what may be expected in terms of learning outcomes. Hence, a more specific focus on one particular object of assessment within one particular educational context is required if we want to move from relatively general conclusions towards specific syntheses of empirical evidence. The current study specifically focuses on the relation between peer feedback and students' academic writing performance within the higher education context for two reasons. First and foremost, the development of academic writing skills is considered important across higher education disciplines and institutes. Second, peer feedback research often focuses on academic writing and is conducted in various research domains. Consequently, a meta-analysis on the impact that peer feedback has on higher education students' writing performance appears to be relevant across both educational and research disciplines, and simultaneously appears to be practically feasible given the anticipated number of studies published.

Definitions

Formative peer feedback

Based on the definition of peer assessment by Topping (1998) and the definition of formative feedback by Shute (2008), formative peer feedback in this study is defined as '*all* task-related information that a learner communicates to a peer of similar status which can be used to improve his or her academic writing performance'. Hence, peer feedback is formative in the sense that it can be utilized by the peer to improve subsequent writing. In addition, this definition encompasses all types of information, including basic peer feedback such as grades or ordinal rankings. This allows us to cover the literature on both peer feedback and peer assessment. In this study, 'peer feedback' refers to *formative* peer feedback unless stated otherwise.

Academic writing

According to Hayes and Flower (1987), critical features of the writing process include that it is goal directed, that writing goals are hierarchically organized, and that these goals are accomplished through three recursive processes: (a) planning, (b) sentence generation and (c) revision.

Therefore, the current study focuses on higher education writing assignments that include such features of the writing process, for example laboratory reports and (sections of) papers.

Research questions

The current study synthesizes the available empirical, quantitative research regarding the impact of peer feedback on the academic writing performance of higher education students. Two sets of research questions are addressed.

Peer feedback effectiveness

Peer feedback has conventionally been compared to alternative feedback sources such as teaching staff, both in terms of its outcomes (e.g., Topping 1998; Cho & Schunn 2007) and in terms of the reliability and validity of these outcomes (e.g., Falchikov and Goldfinch 2000). Indeed, comparing the effectiveness of a particular practice to practically feasible alternatives is informative for teachers in higher education. Therefore, the current study's first set of research questions addresses the impact of peer feedback compared to baseline and two frequently available alternatives: To what extent does engagement in peer feedback improve students' writing performance in comparison to: (a) receiving no feedback at all, (b) self-assessment and (c) feedback from teaching staff (i.e. subject-matter experts or trained teaching assistants)?

Exploration of practically applicable design variables

The second set of research questions investigates the impact of peer feedback on academic writing in relation to two of the variables that Gielen, Dochy and Onghena (2011) mentioned in their review: (d) the nature of the peer feedback (qualitative comments, quantitative grades/ranks, or a combination of both) and (e) the number of peers that students engaged with during peer feedback. Gielen, Dochy and Onghena provided an overview of 20 variables that could be considered important for the design and implementation of peer feedback tasks. Among others, these included variables related to the interaction between peers (e.g., anonymity), how groups are composed (e.g., random or friendship matching) and how the assessment procedure is managed (e.g., training or guidance). As the current study's second central aim is to be of practical value for higher education teaching staff, we focused on those design variables that were both sufficiently available for analysis and that, above all, are practically applicable and adaptable by higher education teaching staff.

For this purpose, six higher education teachers from different institutes and disciplines, all experienced with incorporating peer feedback into their teaching practice, were interviewed and performed a card-sorting task to rank Gielen et al.'s variables from 1 (*completely uncontrollable*) to 5 (*completely controllable*). Borrowing from planned behaviour theory (Ajzen 1991; Ajzen and Fishbein 2005), these perceptions of controllability were then cross-referenced with the prevalence of these design variables across the included studies. This resulted in the focus on two variables that were reported in the included studies and perceived as controllable by the higher education teachers: 'student output' (the quantitative/qualitative nature of the peer feedback) and 'assessor constellation' (the number of peer reviewers in particular).

Method

Focus and inclusion criteria

Following on Topping's (1998) review, the timespan of the search was set to range between 1 January 1998 and 31 October 2016. Given the focus on empirical evidence for the effects of peer

feedback on higher education students' academic writing performance, articles were considered for inclusion when they: (1) were published in English language, peer reviewed academic journals, (2) were empirical in nature, and (3) reported on higher education students. In addition, articles were required to: (4) report on *formative* peer feedback (5) in relation to quantitative measures of academic writing performance. Here, peer feedback was considered formative when students had the opportunity to utilize the peer feedback to improve their writing (e.g., Sadler 1989; Wingate 2010). Finally, (6) the effects on students' writing performance should be attributable to the peer feedback process. Specifically, this means that: (a) no parallel, confounding feedback sources such as teacher feedback or automated feedback were reported, and (b) writing performance was measured both before and after formative peer feedback. One exception to this pretest-posttest criterion were posttest-only designs in which a priori between-group differences were tested to be absent or could be assumed to be minimal, for example by testing between-group similarities based on a relevant proxy, through (quasi-)random allocation of participants into groups or conditions, or through blocked grouping procedures.

Finally, from a methodological perspective, (c) the presence of a reference group was considered highly desirable for attributing writing performance effects to preceding peer feedback processes. Nevertheless, given that the proportion of studies that met all but this final criterion was relatively large, the inclusion of studies that adopted a one-group pretest-posttest design was considered informative. These one-group pretest-posttest studies were incorporated separately into the second set of research questions, both because they reflect different types of effects compared to the studies with a reference group (within-group writing improvement versus between-group comparisons of writing improvement, respectively) and because they tend to overestimate treatment effects compared to studies that do include reference groups (Lipsey and Wilson 1993).

Search strategies

Search terminology and databases

The systematic search was conducted via EBSCOhost (including Academic Search Premier, ERIC, PsycARTICLES, Psychology and Behavioural Sciences Collection, and PsycINFO) and Web of Science. Search terms were determined through two complimentary steps. First, prior review studies (Topping 1998; Falchikov and Goldfinch 2000; van Gennip, Segers and Tillema 2009) were inspected with respect to the search terms used for the independent variable 'peer feedback' and the dependent variable 'academic writing performance'. This resulted in four search terms for the independent variable: *peer feedback*, *peer assessment*, *peer evaluation* and *peer review*, and in eight search terms for the dependent variable: *writing skill**, *writing competen**, *writing proficiency*, *writing performance*, *writing ability*, *writing quality*, *writing achievement* and *essay*. Second, an informal member check with two researchers in the field was conducted to verify our overview of the seminal and/or recent academic literature. This resulted in an additional fifth search term for the independent variable: *peer revision*, and a ninth search term for the dependent variable: *text*.

Article selection

The search yielded a total of 934 unique hits across search engines. A manual assessment of titles and abstract with respect to the higher education context resulted in a selection of 289 articles, of which 287 full-texts (99.3%) were retrieved. These full-texts were assessed by the first two authors with respect to the inclusion criteria, and agreement was determined between the first author (assessing all 287 articles) and second author (assessing a subset of 45 articles). A 'proportional random selection' procedure was applied, meaning that a $\geq$ 15% random selection was drawn separately out of the included and excluded articles, as assessed by the first author.

Importantly, the second author was blinded for the first author's inclusion-exclusion ratio. Inter-rater agreement for the decision on inclusion was calculated to be κ = .81 [.55, 1.00], which may be considered substantial (Landis and Koch 1977). Disagreements were resolved between the first and second author, resulting in the retraction of one inclusion judged by the first author.

Given the substantial inter-rater agreement, the first author's decision on inclusion was followed for the remaining 242 articles. Uncertainties by either of the two authors were resolved through team discussion. In total, 25 articles proved eligible for inclusion, 16 of which had a reference group. As two articles (Sampson and Walker 2012; Walker and Sampson 2013) were based on the same data, only the study with the largest sample size (Sampson and Walker 2012) was retained. Hence, 24 articles (8.4%) were ultimately included in the current study. Among the 16 included articles with a reference group, the data reported in 3 articles was insufficient to calculate an effect size and supplementary data could not be retrieved via the articles' authors (see Table 1 for a complete overview). These articles were not incorporated in the meta-analyses, although they were included in the qualitative analysis.

Statistical methods

Computation of effect sizes

For studies including a reference group, effect sizes (standardized mean differences) were computed based on reported group means and standard deviations. When either of these was missing, effect sizes were based on inferential statistics instead. Where possible, effect sizes were based on gain scores (e.g., Lipsey and Wilson 2001; Wright 2006) to account for potential a priori between-group differences. Alternatively, they were based on the groups' posttest scores (cf. Lazonder and Harmsen 2016) provided groups did not significantly differ at pretest. When multiple types of between-group comparisons were reported, reference groups were averaged where conceptually feasible to retain as much of the available data as possible. Alternatively, the comparison that best fitted the goals of this meta-analysis was included. If averaging was conceptually unfeasible and the relative fit of the different comparisons with the current study's goals was considered to be arbitrary, one comparison was randomly chosen by rolling a dice.

In case academic writing performance after peer feedback was measured multiple times within one assignment and effect sizes could not be based on repeated measurement statistics due to the insufficiently available statistics or data, between-group comparisons were based on final posttest-scores in case groups tested similar at the first pretest measure (before peer feedback). In case academic writing performance after peer feedback was measured multiple times at different assignments, average pretest and posttest scores were created to facilitate a single between-group comparison. Finally, in case multiple types of scores were simultaneously reported as indicators of students' writing performance, scores were averaged into composite scores of academic writing performance. In the study by Stellmack et al. (2012), for example, students' papers were graded by two different graders, effectively resulting in two grade-sets for the same writing tasks. These grade-sets were averaged before calculating effect sizes.

For studies without a reference group, i.e. studies that adopted a one-group pretest-posttest design, effect sizes (standardized gain scores) were computed based on reported pretest and posttest scores or gain scores (see Lipsey and Wilson 2001, 44). In case effect sizes or their standard errors were missing, these were computed using reported inferential statistics where possible (e.g., Greenberg 2015). When pretest-posttest correlations were missing, could not be computed, and proved not retrievable via the article's author(s), this correlation was assumed zero, resulting in conservative estimates of standard errors for these effect sizes. In case multiple rounds of peer feedback and revision were reported and effect sizes could not be based on repeated measures statistics (e.g., Sampson and Walker 2012; Cheng, Liang and Tsai 2015), effect sizes were based on

Table 1. Included Studies: Characteristics and Effect Sizes.

Author(s) & Year	Object assessed	N*	Ref. group	Reference group comparison*	# FB Peers	Peer interaction	FB type	Effect size g	Effect size SEg	95% CI lower	95% CI upper
Birjandi & Tamjid (2012)	Essay	66	Y	Peer vs. Teaching staff	Group	Offline	Both	−0.32	0.20	−0.72	0.08
Cahyono & Amrina (2016)	Essay	71	Y	Peer vs. Self	1	Offline	Both	0.54	0.28	−0.13	1.10
Cho & Schunn (2007)	Paper	18†	Y	Peer vs. Teaching staff	6	Online	Both	1.25	0.50	0.28	2.22
Cho & MacArthur (2011)	Lab report	40	Y	Peer (providing) vs. No-FB control	3	Online	Both	1.24	0.34	0.57	1.90
Ciftci & Kocoglu (2012)	Essay	30	Y	Blog (exp) vs. F2F (ref)	n/a	Online vs. Offline	Co	0.93	0.38	0.19	1.66
Gunersel et al. (2008)	Essay	47	Y	Low (exp) vs. High performers (ref)	3	Online	Both	n/a	n/a	n/a	n/a
Hartberg et al. (2008)	Abstract	50	Y	Peer vs. Teaching staff	3	Online vs. Offline	Both	0.68	0.29	0.11	1.24
Lee (2015)	Buss. plan	96	Y	Reciprocal (exp) vs. Receiving only (ref)	1	Online	Both	0.47	0.22	0.04	0.89
Diab (2011)	Essay	40	Y	Peer vs. Self	1	Offline	Both	0.55	0.32	−0.07	1.17
Novakovich (2016)	Essay	42	Y	Blog (exp) vs. F2F (ref)	3	Online vs. Offline	Co	0.87	0.32	0.25	1.49
Stellmack et al. (2012)	Report	80	Y	Peer vs. Self	1	Offline	Co	0.13	0.17	−0.20	0.47
Tsai & Chuang (2013)	Essay	48	Y	Peer vs. No-FB control	3	Online	Co	0.71	0.22	0.27	1.15
Wong & Storey (2006)	Essay	36	Y	Peer vs. Self	1	Offline	Co	n/a	n/a	n/a	n/a
Xiao & Lucking (2008)	Wiki	232	Y	FB + grades (exp) vs. Grades (ref)	3	Online	d.o.c.	0.50	0.13	0.24	0.76
Yang et al. (2006)	Essay	79	Y	Peer vs. Teaching staff	1	Offline	Co	n/a	n/a	n/a	n/a
Yang & Meng (2013)	Paper	50	Y	Low (exp) vs. High performers (ref)	3	Online	Co	1.30	0.31	0.70	1.91
Cheng et al. (2015)	Report	47	N	NA	5	Online	Both	0.35	0.21	−0.05	0.76
Cho & Cho (2011)	Lab report	72	N	NA	3-4	Online	Both	2.14	0.24	1.67	2.62
Crossman & Kite (2012)	Proposal	208	N	NA	2	Offline	Co	0.64	0.04	0.56	0.72
Greenberg (2015)	Report	46	N	NA	1	Offline	Gr	0.32	0.11	0.11	0.53
Hu & Lam (2010)	Report	20	N	NA	1	Offline	Co	0.41	0.13	0.16	0.66
Noroozi et al. (2016)	Essay	187	N	NA	2	Online	Co	0.34	0.09	0.16	0.53
Sampson & Walker (2012)	Report	18	N	NA	Group (3-4)	Offline	Both	1.71	0.39	0.95	2.47
Yoshizawa et al. (2012)	Essay	35	N	NA	1	Offline	Co	0.41	0.15	0.12	0.70

Note: * = selected comparisons; † = conservative estimate of N = 9 per group; SA = self-assessment; SP = single peer; MP = multiple peers; T = teaching staff; d.o.c. = depending on condition.

NA = not applicable; n/a = not available or unknown; Gr = Grades/ranking only; Co = Comments only; Both = Grades/ranking + comments;.

averaged gain scores and pooled standard errors. For *all* estimated effect sizes reported in the current study, a correction for sample size was applied (Hedges' g, see Borenstein et al. 2009)

Data analysis

Consistent with the research questions, three separate meta-analyses were conducted for the studies that included a reference group: (a) peer feedback versus no-feedback control, (b) peer feedback versus self-assessment, and (c) peer feedback versus feedback from teaching staff. Given the variability in the studies' disciplinary contexts and their differing designs of the peer feedback process, random effects models were fitted for research questions (a), (b) and (c). Two mixed-effects model analyses were conducted for research questions (d) and (e) to explore the moderating role of the nature of the peer feedback and the number of peers engaged with during peer feedback, respectively. The data was analysed using the 'metafor' package (version 2.0-0, Viechtbauer 2010; see also Polanin, Hennessy and Tanner-Smith 2017) in R (version 3.4.2, R Core team 2017). Effect sizes were weighted by their studies' sample size by assigning inverse variance weights, and restricted maximum likelihood estimation (REML) was used to estimate residual heterogeneity (see Raudenbush 2009).

Results

Peer feedback effectiveness

The first set of research questions investigated the impact that engaging in peer feedback has on students' academic writing performance: (a) in comparison to receiving no feedback at all, (b) in comparison to self-assessment and (c) in comparison to feedback from teaching staff (see Figure 1). Regarding the effects of peer feedback compared to no feedback, the only two studies including such a comparison (Cho and MacArthur 2011; Tsai and Chuang 2013) showed a large
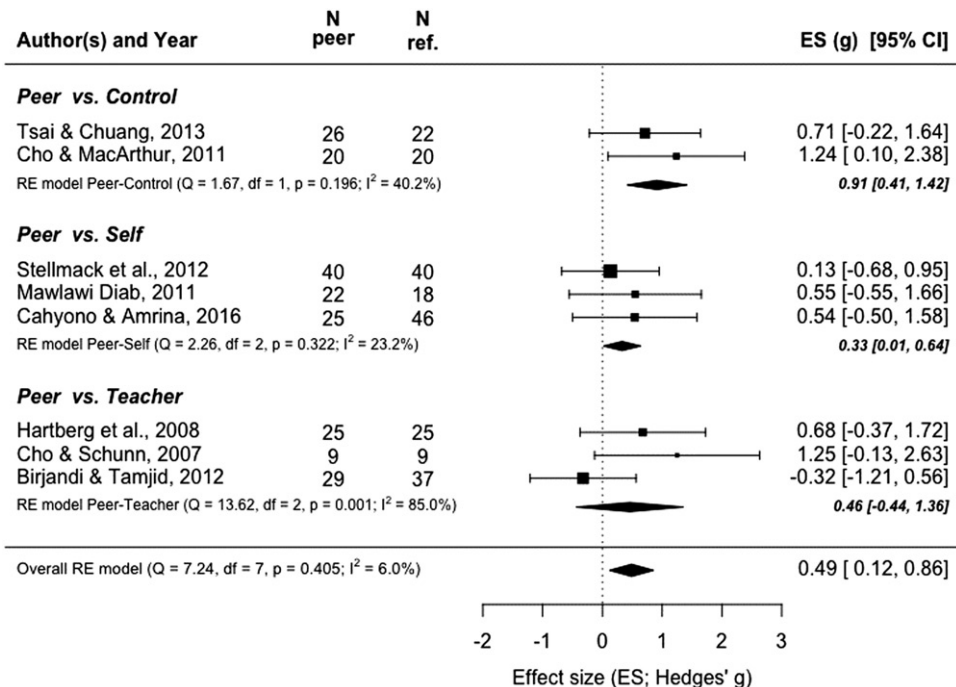

Author(s) and Year	N peer	N ref.		ES (g) [95% CI]
Peer vs. Control				
Tsai & Chuang, 2013	26	22		0.71 [-0.22, 1.64]
Cho & MacArthur, 2011	20	20		1.24 [0.10, 2.38]
RE model Peer-Control (Q = 1.67, df = 1, p = 0.196; I² = 40.2%)				*0.91 [0.41, 1.42]*
Peer vs. Self				
Stellmack et al., 2012	40	40		0.13 [-0.68, 0.95]
Mawlawi Diab, 2011	22	18		0.55 [-0.55, 1.66]
Cahyono & Amrina, 2016	25	46		0.54 [-0.50, 1.58]
RE model Peer-Self (Q = 2.26, df = 2, p = 0.322; I² = 23.2%)				*0.33 [0.01, 0.64]*
Peer vs. Teacher				
Hartberg et al., 2008	25	25		0.68 [-0.37, 1.72]
Cho & Schunn, 2007	9	9		1.25 [-0.13, 2.63]
Birjandi & Tamjid, 2012	29	37		-0.32 [-1.21, 0.56]
RE model Peer-Teacher (Q = 13.62, df = 2, p = 0.001; I² = 85.0%)				*0.46 [-0.44, 1.36]*
Overall RE model (Q = 7.24, df = 7, p = 0.405; I² = 6.0%)				0.49 [0.12, 0.86]

Effect size (ES; Hedges' g)

Figure 1. Effect sizes for varying reference-group comparisons.

composite effect (0.91 [0.41, 1.42]), suggesting that students' engagement in a peer feedback process improves their writing performance as compared to when no feedback is provided at all.

Regarding the comparison between peer feedback and self-assessment, the composite effect size of the three available studies that directly make this comparison (Diab 2011; Stellmack et al. 2012; Cahyono and Amrina 2016) was small but significant (0.33 [0.01, 0.64]). This suggests that students improve their writing performance more after having engaged in peer feedback than after having engaged in a form of self-assessment. Although effect sizes could not be calculated for the study by Wong and Storey (2006), their findings were in line with these results, suggesting larger writing improvements for students engaged in peer feedback as compared to self-assessment.

The third comparison was that between peer feedback and feedback from teaching staff. Here, the direction of effects was mixed across the three studies (Cho and Schunn 2007; Hartberg et al. 2008; Birjandi and Tamjid 2012), resulting in an intermediate sized, though non-significant composite effect size of 0.46 [-0.44, 1.36]. Based on this small sample of studies, students' writing performance does not appear to be differentially affected by peer feedback and feedback from teaching staff. For one study (Yang and Meng 2013) no effect sizes were available or could be calculated. This study compared peer feedback to feedback from teaching staff, and reported larger writing improvement after feedback from teachers than after feedback from peers.

Exploration of practically applicable design variables

Nature of the peer feedback

Across all included studies, the nature of the peer feedback included both a qualitative component such as written comments and a quantitative component such as grades or rankings in eleven studies (46%). In another eleven studies, peer feedback only consisted of peer comments. In only one study (Greenberg 2015) peer feedback was instructed to be merely quantitative (e.g., scores; see Table 1). The remaining study by Xiao and Lucking (2008) is the only included study directly comparing the nature of peer feedback. Specifically, 114 students provided and received ratings and comments, whereas 118 students provided and received ratings only. After the peer feedback phase, students that exchanged both peer comments and grades outperformed those that had only exchanged peer grades (0.50 [0.24, 0.76]). The results of this study by Xiao and Lucking (2008) suggest that the combination of qualitative and quantitative peer feedback is more effective in improving students' writing performance than quantitative peer feedback alone.

Among the studies without a reference group, three studies included both qualitative peer comments as well as quantitative peer grading or ranking (Cheng, Liang, and Tsai 2015; Sampson and Walker 2012; Cho and Cho 2011). Their respective effect sizes ranged between small (0.35 [-0.05, 0.76]) and large (1.71 [0.95, 2.47] and 2.14 [1.67, 2.62]), which weighted into a composite effect size of 1.39 [0.29, 2.48]. In all three studies, the peer feedback processes involved three or more students in reviewing a single peers' written work. Furthermore, peer feedback was anonymous, and all three studies incorporated some form of guidance or instructions with regard to the assignment criteria. Sampson and Walker (2012) differed from the other two studies in two respects: peer feedback was conducted in-class on hard-copies as opposed to online, and peer feedback was provided by groups of three to four students instead of by multiple students individually.

In the four one-group pretest-posttest studies that included peer comments without peer grading or ranking (Noroozi, Biemans and Mulder 2016; Hu and Lam 2010; Yoshizawa, Terano and Yoshikawa 2012; Crossman and Kite 2012), the respective effect sizes ranged between small and intermediate (Cohen, 1988): 0.34 [0.16, 0.53], 0.41 [0.16, 0.66], 0.41 [0.12, 0.70] and 0.64 [0.56, 0.72]. These weighted into a composite effect size of 0.48 [0.31, 0.64]. In all these four studies, peer feedback was guided in one way or another, students engaged with one or two peers, and
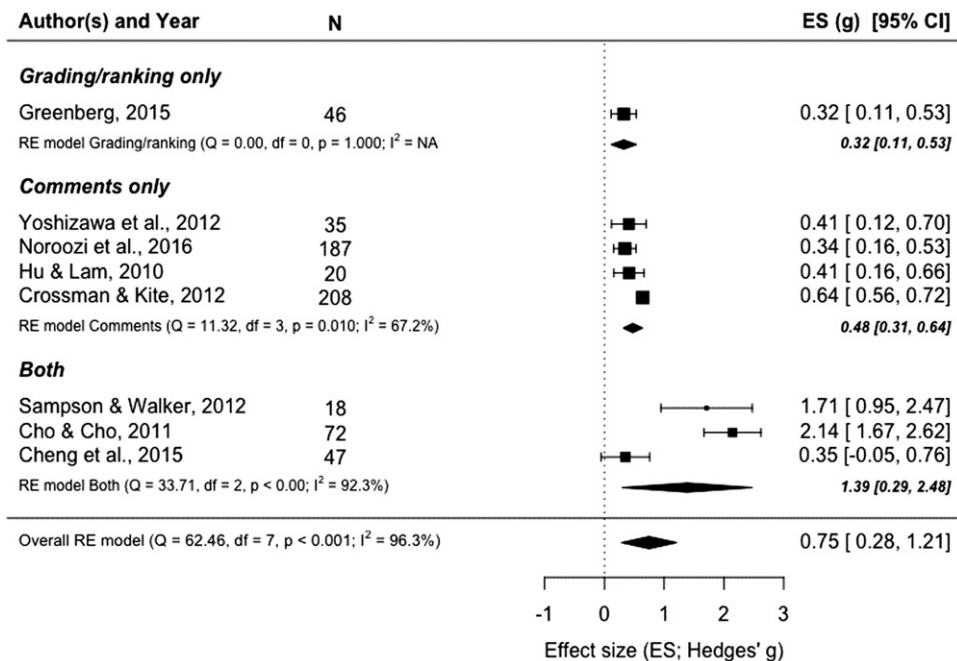
Figure 2. Effect sizes for one-group pretest-posttest studies by nature of the peer feedback.

peer feedback generally took place in-class (only Noroozi, Biemans and Mulder 2016, was both in-class and online). In two studies (Crossman and Kite 2012; Hu and Lam 2010) peer feedback was face-to-face, allowing for peer dialogue. In the remaining one-group pretest-posttest study (Greenberg 2015), peer feedback only consisted of scores based upon a thematic three-point rating scale, for which an effect size of 0.32 [0.11, 53] was reported. Peer feedback in this study was an anonymous, in-class process that was guided by a scoring form.

Summarizing, a direct comparison regarding the nature of peer feedback by Xiao and Lucking (2008) suggests that peer feedback including comments *in addition to* grades improves students' writing more than peer feedback that includes grades alone. This pattern appears to be confirmed within the group of studies that did not include a reference group; large effect sizes were more frequently present and more substantial in the studies where peer feedback simultaneously included both comments and grades (see Figure 2). A moderator analysis was conducted to test the extent to which the nature of the peer feedback related to students' writing improvement. The variation in students' writing improvement indeed was moderated by the nature of the peer feedback ($\beta_{FBnature}$ = 0.61, $z = 2.02$; $Q_M(1)$, = 4.10, $p = .043$, $I^2 = 95.5\%$), such that a combination of both comments and grades resulted in larger writing improvements than either comments or grades alone.

Number of peers engaged with

Across all included studies, the number of peers with whom students engaged during the peer feedback process ranged between one and six, with the mode being three. Two studies (Birjandi and Tamjid 2012; Sampson and Walker 2012) adopted a different procedure, with peer feedback on individual students' academic writing being provided in a group-wise manner (see Table 1).

Among the included studies with a reference group, the only one that directly assessed students' writing improvement in relation to the number of peer reviewers is Cho and Schunn (2007). These authors compared the writing improvement of students that either received feedback from a single expert, a single peer or six peers. Only one between-group comparison
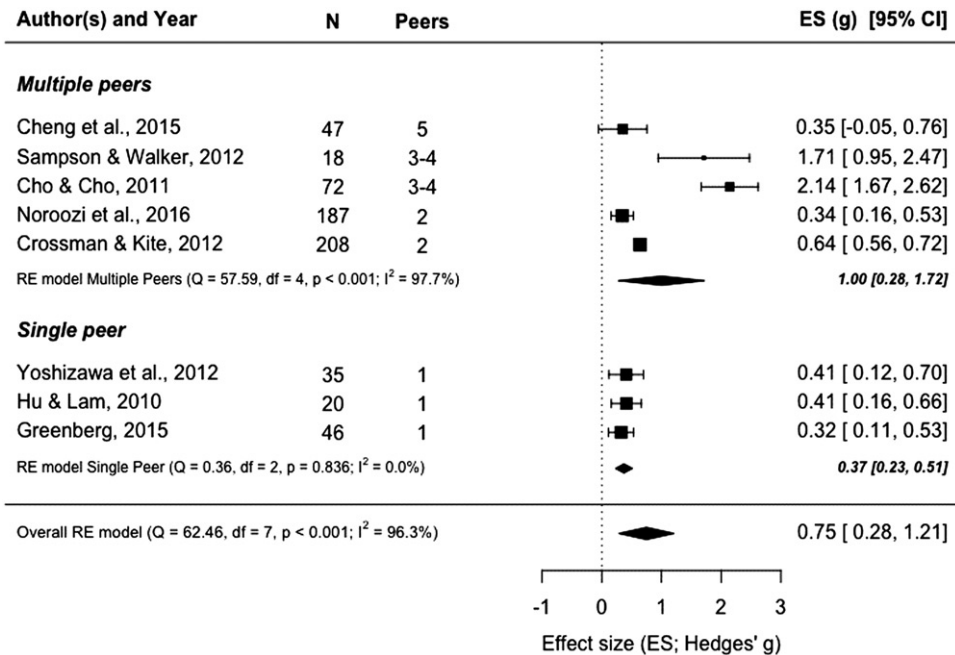
Author(s) and Year	N	Peers		ES (g) [95% CI]
Multiple peers				
Cheng et al., 2015	47	5		0.35 [-0.05, 0.76]
Sampson & Walker, 2012	18	3-4		1.71 [0.95, 2.47]
Cho & Cho, 2011	72	3-4		2.14 [1.67, 2.62]
Noroozi et al., 2016	187	2		0.34 [0.16, 0.53]
Crossman & Kite, 2012	208	2		0.64 [0.56, 0.72]
RE model Multiple Peers (Q = 57.59, df = 4, p < 0.001; I² = 97.7%)				**1.00 [0.28, 1.72]**
Single peer				
Yoshizawa et al., 2012	35	1		0.41 [0.12, 0.70]
Hu & Lam, 2010	20	1		0.41 [0.16, 0.66]
Greenberg, 2015	46	1		0.32 [0.11, 0.53]
RE model Single Peer (Q = 0.36, df = 2, p = 0.836; I² = 0.0%)				**0.37 [0.23, 0.51]**
Overall RE model (Q = 62.46, df = 7, p < 0.001; I² = 96.3%)				0.75 [0.28, 1.21]

Effect size (ES; Hedges' g)

Figure 3. Effect sizes for one-group pretest-posttest studies by number of peers engaged with.

appeared significant: students receiving feedback from six peers improved their writing to a larger extent than students receiving feedback from a single expert. However, no significant difference in writing improvement was found for students receiving feedback from one versus six peers. There did appear to be an upward trend in writing improvement as the number of peers increased, but small sample sizes limited the generalizability of this trend. Clearly, conclusions regarding the effect that the number of peer reviewers has on students' writing improvement cannot be drawn based on this single study.

For the eight studies without a reference group, students in three studies engaged with no more than one peer during peer feedback (Greenberg 2015; Hu and Lam 2010; Yoshizawa, Terano and Yoshikawa 2012). The respective effect sizes for these studies ranged between small (0.32 [0.11, 0.53]) and intermediate (0.41 [0.16, 0.66] and 0.41 [0.12, 0.70]), weighting into a composite effect size of 0.37 [0.23, 0.51]. The between-study differences included students' anonymity (only in Hu and Lam (2010) were students aware of each other's identities) or the nature of the peer feedback (in Greenberg 2015, peer feedback was restricted to rubric scores). However, there were at least as many commonalities. In all three studies, peer feedback occurred in-class, was performed in writing without opportunity for peer dialogue and included some form of guidance with respect to the assessment criteria. In the other five studies adopting a one-group pretest-posttest design (Noroozi, Biemans and Mulder 2016; Cheng, Liang and Tsai 2015; Crossman and Kite 2012; Sampson and Walker 2012; Cho and Cho 2011), students engaged with multiple peers during peer feedback. The respective effect sizes for these five studies ranged from small to large (0.34 [0.16, 0.53], 0.35 [-0.05, 0.76], 0.64 [0.56, 0.72], 1.71 [0.95, 2.47] and 2.14 [1.67, 2.62], respectively). The weighted composite effect size for these five studies was 1.00 [0.28, 1.72]. In all five studies, peer feedback was guided by explicit criteria and/or rubrics. In all but one of these studies (the exception being Crossman and Kite 2012), peer feedback was performed in writing without opportunity for peer dialogue.

Insofar it is possible to distinguish patterns relating the number of peer reviewers to the magnitude of students' writing improvement, effect sizes appear to be larger in the studies where

peer feedback was provided by multiple peers (see Figure 3). A moderator analysis tested the extent to which students' writing improvement varied as a result of their engagement with either one or multiple peers. Between these eight studies, this did not appear to be the case ($\beta_{NRpeers}$ = 0.60, z = 1.27; $Q_M(1)$, = 1.62, p = .202, I^2 = 96.2%).

Discussion

This study meta-analysed the effect of peer feedback on the academic writing performance of higher education students. Two sets of research questions were addressed. First, the effects of peer feedback on academic writing were analysed in comparison to baseline (no feedback) or to the effects of two alternative feedback sources (self or teacher). Second, the moderating role of two peer feedback 'design variables' in explaining students' writing improvement were explored: the nature of peer feedback and the number of peers with whom students engaged.

Peer feedback effectiveness

Regarding the first comparison, a large effect size indicated that students improved their writing more when they engaged in peer feedback than when they did not provide and/or receive any type of feedback. The limited number of studies limits the extent to which this finding can be generalized. Nevertheless, this finding corroborates more descriptive conclusions of prior qualitative review studies. For example, van Zundert, Sluijsmans and van Merriënboer (2010) concluded that peer feedback can stimulate the development in domain-specific skills. However, the studies in their analysis included students from both primary education and higher education contexts and concerned diverse outcome measures (e.g., academic writing, science activity design). The current study adds to the research by providing a baseline estimate for the effect that peer feedback has on higher education students' academic writing performance.

The second comparison indicated larger writing improvements for students engaged in peer feedback than for students engaged in self-assessment (e.g., rubric-guided self-assessment, such as in Stellmack et al. 2012). This effect size was notably smaller than the prior baseline comparison. Both these observations can be aligned with prior research findings. First, the observation that the effect size for peer feedback is larger than that for self-assessment may be explained by inherently different characteristics of the two feedback processes. For example, peers may introduce students to ideas and arguments from very different perspectives, which is increasingly the case as multiple peers become involved. Reversely, peer feedback can expose students to an array of alternative approaches, ideas and writing styles, which may have more impact than having one model answer (McConlogue 2015). The act of providing peer feedback also requires students to actively (re)consider the assignment criteria, which may improve their own subsequent writing performance (Flower et al. 1986; Patchan and Schunn 2015).

Second, there is the observation that the effect of peer feedback was smaller when compared to self-assessment than when compared to baseline. It seems plausible that self-assessment does account for some variation in effects of students' writing performance. For example, self-assessment may improve learning by triggering students to reflect upon their learning process (Dochy, Segers and Sluijsmans 1999). There also is evidence that self-assessments can be relatively reliable indicators of performance. For example, self-assessment can correlate with holistic assessments by teaching staff (e.g., Falchikov and Boud 1989) and can be largely similar to peer and teacher assessments with regard to specific aspects of writing assignments (Lindblom-Ylänne, Pihlajamäki and Kotkas 2006). In the context of online education, however, self-assessments may be biased (e.g., Admiraal, Huisman and Pilli 2015), which should at least prompt thoughtful considerations regarding the utilization of self-assessment for formal assessment procedures (e.g., Matsuno 2009).

The third comparison contrasted peer feedback with feedback from teaching staff and did not indicate a systematic difference with respect to the impact on students' academic writing. Given the low number of quantitative studies that incorporated such direct comparisons and the variability in the individual effect sizes of those studies, caution is required in generalizing this finding as well. Still, these findings corroborate those of Topping's (1998) qualitative review, and are in line with those of Cho and Schunn (2007) as well. One comparison that these authors reported, which was not included in the current study's quantitative analyses as a result of the random selection for an interrelated comparison, concerned that between feedback from a single peer versus feedback from a single expert. Cho and Schunn (2007) reported a similar impact on students' writing improvement for both conditions, which aligns with prior studies reporting high correlations between peer and teacher judgements (e.g., Falchikov and Goldfinch 2000).

There are arguments in favour of teacher feedback (e.g., more expert knowledge) as well as arguments in favour of peer feedback. For example, peer feedback may induce reflection (e.g., Nicol, Thomson and Breslin 2014). That the assessor status of a peer is different from that of a teacher or an expert may enhance critical appraisal of the feedback by the recipient (Strijbos, Narciss and Dünnebier 2010). Based on the diverse nature and implications of these arguments, we conceive this comparative question of effectiveness as requiring contextualization depending on characteristics of the learning environment, the task and the learning goals. For example, the argument that peer feedback is more available and faster (e.g., Topping 1998) seems tied to both the student-to-teacher ratio within a particular learning environment as well as the size and complexity of the writing task. From our perspective, the question whether peer feedback or teacher feedback is most efficient can hardly be considered without taking into account the reality constraints with which higher education teaching staff are confronted in their teaching practice. This raises the issue of practical applicability.

Exploration of practically applicable design variables

The second set of research questions investigated the role of specific peer feedback design variables (see Gielen, Dochy and Onghena 2011) in explaining higher education students' academic writing performance. Our analysis focused on two specific design variables that higher education teachers perceived as controllable: the nature of the peer feedback and the number of peers that students engaged with during peer feedback.

Regarding the nature of the peer feedback, a differentiation was made between either grading or ranking only, qualitative commenting only, or a combination of both. The composite effect size for studies that simultaneously included both grades and comments was large, whereas the effect size was intermediate for studies in which only comments were provided. The only included study directly investigating the relation between the nature of the peer feedback and students' writing performance (Xiao and Lucking 2008) reported an intermediate effect size for the combination of both comments and grades as opposed to grading only. A moderation analysis in the current study indicated that the nature of the peer feedback indeed moderated the effects of peer feedback on students' writing performance. Specifically, a combination of both comments and grades tended to result in larger writing improvements than either comments or grades alone. This is in line with the conclusion by Sadler (1989). Sadler argues that students benefit from feedback on academic tasks when they know: (1) what good performance is, (2) how their current performance relates to good performance and (3) how to close the gap between current and good performance (see also Nicol and Macfarlane-Dick 2006).

Possibly, students perceive some type of holistic assessment in addition to comments as helpful in determining how their current performance relates to their aspired level of performance. At the same time, students can also have reservations about peer grading (e.g., Liu and Carless 2006). At least at first, these two findings appear at odds. Some valuable insights are provided

here by Nicol, Thomson and Breslin (2014), who reported the arguments of students who either were in favour of or against peer grading. Students in favour of peer grading mentioned that a grade would give them a 'more accurate picture of how they were doing' (p. 109). In contrast, the students who were against peer grading mentioned issues relating to the limited expertise of their peers and their subsequent concerns of accuracy and fairness. One conclusion could be that students' valuation of peer grades is contingent on the role that these grades play in formal assessment. If this is the case, it may be possible to have the best of both worlds by incorporating peer grading in a 'no stakes' manner (i.e. by making clear that peer grades are purely formative and do not weigh into students' final grade).

For the three studies in the moderator analyses that included both comments and grades (Cho and Cho 2011; Sampson and Walker 2012; Cheng, Liang and Tsai 2015), the weighting of peer grades unfortunately either varied or was unclear. Hence, the weighting of peer grades may be one feature to investigate for future research. At minimum, future peer feedback studies should be clear about the role that peer grades and comments have in students' formal assessment when investigating how the nature of peer feedback influences students' writing performance.

Peer feedback could involve a single peer or multiple peers. A large effect size was found when students that engaged with multiple peers, whereas a small effect size was found when students engaged with only one peer. The only included study directly comparing the effects of feedback from one peer versus multiple peers (Cho and Schunn 2007) found no significant effects on writing improvement. A non-significant trend in that direction was visible, but generalizability was limited due to small sample sizes in their particular study. We also did not find that the number of peers with whom a student engaged significantly moderated writing performance. Although the direction of the effect suggested that engagement with multiple peers positively influences writing performance, the limited number of studies restricts making statistical inferences. More research is required to estimate the reliability of this trend.

If future research would indicate that this trend is reliable, that conclusion would be supported by prior research. For example, the perspectives of multiple peers may be especially beneficial to students' conceptions of how their text is perceived by a target audience (e.g., Schriver 1989). Feedback from multiple peers may be more valid and reliable and therefore be preferred over feedback from a single peer (Cho, Schunn and Wilson 2006; Evans 2013). If future research would show that this trend is not reliable, we would consider this at least somewhat surprising. Consider for example Schriver's (1989) 'audience conception' argument as well as prior theoretical (e.g., Flower et al., 1986) and empirical (Lundstrom and Baker 2009; Cho and MacArthur 2011) studies emphasizing the learning benefits of providing peer feedback. In that light, it seems logical to expect that an increasing number of peer reviewers increases the likelihood of receiving qualitatively good peer feedback, which in turn can be expected to improve students' writing performance. In order to more confidently make inferences, however, more well-controlled, quantitative studies are needed to assess the effects that the number of involved peers has on students' writing performance.

Implications and limitations

Research

To our knowledge, this study is the first to follow up on multiple calls for a quantitative research synthesis for the effects of peer feedback (e.g., Topping 1998, 2010; Gielen, Dochy and Onghena 2011). The current study accomplished this by focusing on one specific object of assessment, academic writing, within one specific educational context, higher education. By specifically focusing on studies that reported quantitative measures of writing performance in higher education, the current study contributes to the literature by estimating the *extent to which* students'

engagement in peer feedback improves their writing performance within this higher education context. The results convey two different but interrelated observations.

The first observation concerns peer feedback effectiveness on higher education students' academic writing performance: engaging in peer feedback appears to improve students writing more than engaging in no feedback at all (large effect size) or than students engaging in self-assessment (small effect size), whereas peer feedback appears similarly effective as feedback from teaching staff. The second observation concerns the limited number of studies that was considered eligible for inclusion. As has been reported by prior review studies (e.g., van Zundert, Sluijsmans and van Merriënboer 2010), research into peer feedback often involves case studies and globally described interventions, limiting the extent to which inferences can be drawn for what caused the outcomes.

As shown by the relatively small number of included studies (24, 8.4% out of all the retrieved full-texts), the proportion of well-controlled, quantitative studies still appears to be limited at the time of writing. This signals a limitation for the area of peer feedback research and, consequently, for the current study as well. Additionally, within this limited set of included studies, this meta-analysis could not always include all relevant data. For the Cho & MacArthur (2011) study specifically, it was necessary to (randomly) exclude one of two 'control' conditions; a peer feedback condition was compared to a 'reading only' and a 'no-feedback' condition, and as such would weigh in twice if both comparisons were included. The random selection of one comparison (in this case: peer feedback versus no-feedback) means that our findings may have varied if the dice landed differently. The limited number of included studies has direct implications for the estimated effect sizes reported in the current study, in particular with respect to the confidence with which these can be generalized. Therefore, we hereby reiterate calls by for example Strijbos and Sluijsmans (2010) for more well controlled, (quasi-)experimental peer feedback studies in which variables related to the design of the task, the intervention and the peer feedback process are well described. To facilitate the process of cumulative knowledge building in this area, the data, syntax and logbook for this study are provided as openly accessible materials online.

Teaching

The exploration of the two practically applicable peer feedback design variables was intended to be informative for higher education teaching staff. Regarding the first variable, the moderating effect of the nature of peer feedback suggests that a combination of both comments and grades result in larger writing improvement by students than peer feedback involving either comments or grades only. Regarding the second variable, a non-significant pattern indicated that students may benefit from engaging with multiple peers as opposed to engaging with one peer. We consider it plausible that future research will prove these patterns to be reliable, for example because the directions of the effects are in line with varying theoretical rationales. The limited number of studies should prompt a degree of caution with respect to their generalizability, however, especially in the case of non-significant patterns. If these patterns prove reliable, that evidently would suggest higher education teaching staff to design peer feedback as including both peer feedback comments as well as grades or rankings, and to have students engage with multiple peers.

Acknowledgements

Data availability and deposition statement

The study logbook and the anonymized data and syntaxes (including additional funnel plots etc.) are accessible via the following link: https://osf.io/ajsbg

References *marked with an asterisk indicate studies included in the analyses*

ORCID

Bart Huisman http://orcid.org/0000-0003-3634-3729
Nadira Saab http://orcid.org/0000-0003-0751-4277
Paul van den Broek http://orcid.org/0000-0001-9058-721X
Jan van Driel http://orcid.org/0000-0002-8185-124X

References

Admiraal, W., B. Huisman, and O. Pilli. 2015. "Assessment in Massive Open Online Courses." *The Electronic Journal of e-Learning* 13 (4):207–16.

Ajzen, I. 1991. "The Theory of Planned Behavior." *Organizational Behavior and Human Decision Processes* 50 (2): 179–211. doi:10.1016/0749-5978(91)90020-T.

Ajzen, I., and M. Fishbein. 2005. "The influence of attitudes on behavior." In *The handbook of attitudes*., edited by D. Albarracin, B. T. Johnson and M. P. Zanna, 173–221. NJ: Lawrence Erlbaum Associates: Mahwah.

Bailey, R., and M. Garner. 2010. "Is the Feedback in Higher Education Assessment Worth the Paper It Is Written on? Teachers' Reflections on Their Practices." *Teaching in Higher Education* 15 (2):187–98. doi:10.1080/13562511003620019.

Ballantyne, R., K. Hughes, and A. Mylonas. 2002. "Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process." *Assessment & Evaluation in Higher Education* 27 (5):427–41. doi:10.1080/0260293022000009302.

* Birjandi, P., and N. H. Tamjid. 2012. "The Role of Self-, Peer and Teacher Assessment in Promoting Iranian EFL Learners' Writing Performance." *Assessment & Evaluation in Higher Education* 37 (5):513–33. doi:10.1080/02602938.2010.549204.

Borenstein, M., L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. 2009. *Introduction to Meta-analysis, introduction to Meta-Analysis*. West Sussex, UK: John Wiley & Sons, Ltd.

* Cahyono, B. Y., and R. Amrina. 2016. "Peer Feedback, Self-Correction, and Writing Proficiency of Indonesian EFL Students." *Arab World English Journal* 7 (1):178–93. doi:10.24093/awej/vol7no1.12.

* Cheng, K. H., J. C. Liang, and C. C. Tsai. 2015. "Examining the Role of Feedback Messages in Undergraduate Students' Writing Performance during an Online Peer Assessment Activity." *Internet and Higher Education* 25: 78–84. doi:10.1016/j.iheduc.2015.02.001.

Cho, K., and C. MacArthur. 2010. "Student Revision with Peer and Expert Reviewing." *Learning and Instruction* 20 (4): 328–38. doi:10.1016/j.learninstruc.2009.08.006.

* Cho, K., and C. MacArthur. 2011. "Learning by Reviewing." *Journal of Educational Psychology* 103 (1):73–84. doi:10.1037/a0021950.

* Cho, K., and C. D. Schunn. 2007. "Scaffolded Writing and Rewriting in the Discipline: A Web-Based Reciprocal Peer Review System." *Computers & Education* 48 (3):409–26. doi:10.1016/j.compedu.2005.02.004.

Cho, K., C. D. Schunn, and R. W. Wilson. 2006. "Validity and Reliability of Scaffolded Peer Assessment of Writing from Instructor and Student Perspectives." *Journal of Educational Psychology* 98 (4):891–901. doi:10.1037/0022-0663.98.4.891.

* Cho, Y. H., and K. Cho. 2011. "Peer Reviewers Learn from Giving Comments." *Instructional Science* 39 (5):629–43. doi:10.1007/s11251-010-9146-1.

Ciftci, H., and Z. Kocoglu. 2012. "Effects of Peer e-Feedback on Turkish EFL Students' Writing Performance." *Journal of Educational Computing Research* 46 (1):61–84. doi:10.2190/EC.46.1.c

Cohen, J. C. 1988. *Statistical power analysis for the behavioural sciences*. (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

* Crossman, J. M., and S. L. Kite. 2012. "Facilitating Improved Writing among Students through Directed Peer Review." *Active Learning in Higher Education* 13 (3):219–29. doi:10.1177/1469787412452980.

* Diab, N. M. 2011. "Assessing the Relationship between Different Types of Student Feedback and the Quality of Revised Writing." *Assessing Writing* 16 (4):274–92. doi:10.1016/j.asw.2011.08.001.

Dochy, F., M. Segers, and D. Sluijsmans. 1999. "The Use of Self-, Peer and co-Assessment in Higher Education: A Review." *Studies in Higher Education* 24 (3):331–50. 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.

Falchikov, N., and D. Boud. 1989. "Student Self-Assessment in Higher Education: A Meta-Analysis." *Review of Educational Research* 59 (4):395–430. doi:10.3102/00346543059004395.

Falchikov, N., and J. Goldfinch. 2000. "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks." *Review of Educational Research* 70 (3):287–322. doi:10.2307/1170785.

Flower, L., J. R. Hayes, L. Carey, K. Schriver, and J. Stratman. 1986. "Detection, Diagnosis, and the Strategies of Revision +." *Composition.*" *College Composition and Communication* 37 (1):16–55. doi:10.2307/357381.

Gielen, S., F. Dochy, and P. Onghena. 2011. "An Inventory of Peer Assessment Diversity." *Assessment & Evaluation in Higher Education* 36 (2):137–55. doi:10.1080/02602930903221444.

Graham, S., and D. Perin. 2007. "A Meta-Analysis of Writing Instruction for Adolescent Students." *Journal of Educational Psychology* 99 (3):445–76. doi:10.1037/0022-0663.99.3.445.

* Greenberg, K. P. 2015. "Rubric Use in Formative Assessment: A Detailed Behavioral Rubric Helps Students Improve Their Scientific Writing Skills." *Teaching of Psychology* 42 (3):211–7. doi:10.1177/0098628315587618.

Gunersel, A. B., N. J. Simpson, K. J. Aufderheide, and L. Wang. 2008. "Effectiveness of Calibrated Peer Review for Improving Writing and Critical Thinking Skills in Biology Undergraduate Students." *Journal of the Scholarship of Teaching and Learning* 8 (2):25–37.

* Hartberg, Y., A. B. Gunersel, N. J. Simspon, and V. Balester. 2008. "Development of Student Writing in Biochemistry Using Calibrated Peer Review." *Journal of the Scholarship of Teaching and Learning* 8 (1):29–44.

Hayes, J. R., and L. S. Flower. 1987. "On the Structure of the Writing Process." *Topics in Language Disorders* 7 (4): 19–30.

* Hu, G., and S. T. E. Lam. 2010. "Issues of Cultural Appropriateness and Pedagogical Efficacy: Exploring Peer Review in a Second Language Writing Class." *Instructional Science* 38 (4):371–94. doi:10.1007/s11251-008-9086-1.

Huisman, B. A., N. Saab, J. H. van Driel, and P. W. van den Broek. 2018. "Peer Feedback on Academic Writing: Undergraduate Students' Peer Feedback Role, Peer Feedback Perceptions and Essay Performance." *Assessment & Evaluation in Higher Education* 36 (7):955–68. doi:10.1080/02602938.2018.1424318

Landis, J. R., and G. G. Koch. 1977. "Measurement of Observer Agreement for Categorical Data." *Biometrics* 33 (1): 159–74. doi:10.2307/2529310.

Lazonder, A. W., and R. Harmsen. 2016. "Meta-Analysis of Inquiry-Based Learning: Effects of Guidance." *Review of Educational Research* 86 (3):681–718. doi:10.3102/0034654315627366.

* Lee, C.-Y. 2015. "The Effects of Online Peer Assessment and Family Entrepreneurial Experience on Students' Business Planning Performance." *TOJET: The Turkish Online Journal of Educational Technology* 14 (1):123–32.

Lindblom-Ylänne, S., H. Pihlajamäki, and T. Kotkas. 2006. "Self-, Peer- and Teacher-Assessment of Student Essays." *Active Learning in Higher Education* 7 (1):51–62. doi:10.1177/1469787406061148.

Lipsey, M. W., and D. B. Wilson. 1993. "The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis." *American Psychologist* 48 (12):1181–209. doi:10.1037/0003-066X.48.12.1181.

Lipsey, M. W., and D. B. Wilson. 2001. *Practical Meta-analysis, applied social research methods series. Thousand oaks.*, Calif: SAGE Publications, Inc.

Liu, N., and D. Carless. 2006. "Peer Feedback: The Learning Element of Peer Assessment." *Teaching in Higher Education* 11 (3):279–90. doi:10.1080/13562510600680582.

Lundstrom, K., and W. Baker. 2009. "To Give Is Better than to Receive: The Benefits of Peer Review to the Reviewer's Own Writing." *Journal of Second Language Writing* 18 (1):30–43. doi:10.1016/j.jslw.2008.06.002.

Matsuno, S. 2009. "Self-, Peer-, and Teacher-Assessments in Japanese University EFL Writing Classrooms." *Language Testing* 26 (1):75–100. doi: 10.1177/0265532208097337.

McConlogue, T. 2015. "Making Judgements: investigating the Process of Composing and Receiving Peer Feedback." *Studies in Higher Education* 40 (9):1495–506. doi:10.1080/03075079.2013.868878.

Nicol, D. J., and D. Macfarlane-Dick. 2006. "Formative Assessment and Self-Regulated Learning: A Model and Seven Principles of Good Feedback Practice." *Studies in Higher Education* 31 (2):199–218. doi:10.1080/03075070600572090.

Nicol, D. J., A. Thomson, and C. Breslin. 2014. "Rethinking Feedback Practices in Higher Education: A Peer Review Perspective." *Assessment & Evaluation in Higher Education* 39 (1):102–22. doi:10.1080/02602938.2013.795518.

* Noroozi, O., H. Biemans, and M. Mulder. 2016. "Relations between Scripted Online Peer Feedback Processes and Quality of Written Argumentative Essay." *Internet and Higher Education* 31:20–31. doi:10.1016/j.iheduc.2016.05.002.

* Novakovich, J. 2016. "Fostering Critical Thinking and Reflection through Blog-Mediated Peer Feedback." *Journal of Computer Assisted Learning* 32 (1):16–30. doi:10.1111/jcal.12114

OECD. 2016. *"Education at a glance 2016: OECD indicators.".* In. OECD Publishing, Paris.

Patchan, M. M., and C. D. Schunn . 2015. "Understanding the Benefits of Providing Peer Feedback: How Students Respond to Peers' Texts of Varying Quality." *Instructional Science* 43(5): 591-614. doi: 10.1007/s11251-015-9353-x.

Polanin, J. R., E. A. Hennessy , and E. E. Tanner-Smith . 2016. "A Review of Meta-Analysis Packages in R." *Journal of Educational and Behavioral Statistics* 42 (2):206–42. doi: 10.3102/1076998616674315.

R Core team 2017. *"R: a language and environment for statistical computing.".* In. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W. 2009. "Analyzing effect sizes: Random-effects models." In *The handbook of research synthesis and Meta-analysis.,* 2nd ed., edited by Harris Cooper, Larry V. Hedges, Jeffrey C. Valentine, Harris Cooper, Larry V. Hedges and Jeffrey C. Valentine, 295–315. New York, NY, US: Russell Sage Foundation.

Sadler, D. R. 1989. "Formative Assessment and the Design of Instructional-Systems." *Instructional Science* 18 (2): 119–44. doi:10.1007/Bf00117714.

* Sampson, V., and J. P. Walker. 2012. "Argument-Driven Inquiry as a Way to Help Undergraduate Students Write to Learn by Learning to Write in Chemistry." *International Journal of Science Education* 34 (10):1443–85. doi:10.1080/09500693.2012.667581.

Schriver, K. 1989. "Evaluating Text Quality: The Continuum from Text-Focused to Reader-Focused Methods." *IEEE Transactions on Professional Communication* 32 (4):238–55. doi:10.1109/47.44536.

Shute, V. J. 2008. "Focus on Formative Feedback." *Review of Educational Research* 78 (1):153–89. doi:10.3102/0034654307313795.

* Stellmack, M. A., N. K. Keenan, R. R. Sandidge, A. L. Sippl, and Y. L. Konheim-Kalkstein. 2012. "Review, Revise, and Resubmit: The Effects of Self-Critique, Peer Review, and Instructor Feedback on Student Writing." *Teaching of Psychology* 39 (4):235–44. doi:10.1177/0098628312456589.

Strijbos, J. W., S. Narciss, and K. Dünnebier. 2010. "Peer Feedback Content and Sender's Competence Level in Academic Writing Revision Tasks: Are They Critical for Feedback Perceptions and Efficiency?" *Learning and Instruction* 20 (4):291–303. doi:10.1016/j.learninstruc.2009.08.008.

Strijbos, J. W., and D. Sluijsmans. 2010. "Unravelling Peer Assessment: Methodological, Functional, and Conceptual Developments." *Learning and Instruction* 20 (4):265–9. doi:10.1016/j.learninstruc.2009.08.002.

Topping, K. J. 1998. "Peer Assessment between Students in Colleges and Universities." *Review of Educational Research* 68 (3):249–76. doi:10.3102/00346543068003249.

Topping, K. J. 2009. "Peer Assessment." *Theory into Practice* 48 (1):20–7. doi:10.1080/00405840802577569.

Topping, K. J. 2010. "Methodological Quandaries in Studying Process and Outcomes in Peer Assessment." *Learning and Instruction* 20 (4):339–43. doi:10.1016/j.learninstruc.2009.08.003.

* Tsai, Y.-C., and M.-T. Chuang. 2013. "Fostering Revision of Argumentative Writing through Structured Peer Assessment." *Perceptual and Motor Skills* 116 (1):210. doi:10.2466/10.23.PMS.116.1.210-221.

van Gennip, N. A. E., M. S. R. Segers, and H. H. Tillema. 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., D. Sluijsmans, and J. van Merriënboer. 2010. "Effective Peer Assessment Processes: Research Findings and Future Directions." *Learning and Instruction* 20 (4):270–9. doi:10.1016/j.learninstruc.2009.08.004.

Viechtbauer, W. 2010. *"Conducting Meta-Analyses in R with the Metafor Package."* Journal of Statistical Software 2010 36 (3):48. doi: 10.18637/jss.v036.i03.

Walker, J. P., and V. Sampson. 2013. "Argument-Driven Inquiry: Using the Laboratory to Improve Undergraduates' Science Writing Skills through Meaningful Science Writing, Peer-Review, and Revision." *Journal of Chemical Education* 90 (10):1269–74. doi:10.1021/ed300656p.

Wingate, U. 2010. "The Impact of Formative Feedback on the Development of Academic Writing." *Assessment & Evaluation in Higher Education* 35 (5):519–33. doi:10.1080/02602930903512909.

* Wong, H., and P. Storey. 2006. "Knowing and Doing in the ESL Writing Class." *Language Awareness* 15 (4): 283–300. doi:10.2167/la365/0.

Wright, D. B. 2006. "Comparing Groups in a before–after Design: When t Test and ANCOVA Produce Different Results." *British Journal of Educational Psychology* 76 (3):663–75. doi:10.1348/000709905X52210.

* Xiao, Y., and R. Lucking. 2008. "The Impact of Two Types of Peer Assessment on Students' Performance and Satisfaction within a Wiki Environment." *Internet and Higher Education* 11 (3-4):186–93. doi: 10.1016/j.iheduc.2008.06.005.

* Yang, M., R. Badger, and Z. Yu. 2006. "A Comparative Study of Peer and Teacher Feedback in a Chinese EFL Writing Class." *Journal of Second Language Writing* 15 (3):179–200. doi:10.1016/j.jslw.2006.09.004.

* Yang, Y. F., and W. T. Meng. 2013. "The Effects of Online Feedback Training on Students' Text Revision." *Language Learning & Technology* 17 (2):220–38.

* Yoshizawa, S., T. Terano, and A. Yoshikawa. 2012. "Assessing the Impact of Student Peer Review in Writing Instruction by Using the Normalized Compression Distance." *IEEE Transactions on Professional Communication* 55 (1):85–96. doi:10.1109/tpc.2011.2172833.