Author: Huisman, B.A.
Title: Peer feedback on academic writing : effects on performance and the role of task-design
Issue Date: 2018-09-12

The impact of formative peer feedback on higher education students' academic writing: A meta-analysis

2

This chapter is an adapted version of:
Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2018). *The impact of formative peer feedback on higher education students' academic writing: A meta-analysis*. Manuscript submitted for publication.

Abstract

Peer feedback is frequently implemented with academic writing tasks in higher education (HE). However, despite a growing body of research and varying qualitative review studies, a quantitative synthesis is still lacking for the impact that peer feedback has on students' writing performance. This meta-analysis synthesized the results of 24 quantitative studies that reported on HE students' academic writing performance after formative peer feedback. Engagement in peer feedback was more effective than no feedback (g = 0.91 [0.41, 1.42]) and self-assessment (g = 0.33 [0.01, 0.64]), and similarly effective as teacher feedback (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. Findings and implications are discussed both for HE teaching practice and future research approaches and directions.

Keywords: Peer feedback, peer assessment, academic writing, higher education, meta-analysis

Introduction

Across higher education (HE) 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 (e.g., van Zundert, Sluijsmans, & van Merriënboer, 2010). Despite a growing body of research however (e.g., Evans, 2013; Gielen, Dochy, & Onghena, 2011; Topping, 1998; van Gennip, Segers, & Tillema, 2009), 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 oftentimes 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 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 as well as for HE teaching staff.

Generally speaking, there are at least two sets of arguments to support the implementation of peer feedback on writing in the HE context. The first relates to the learning benefits for students. Not only can students expect reliable assessments from their peers (Falchikov & Goldfinch, 2000), the very act of providing peer feedback can be beneficial as well (Cho & Cho, 2011; Cho & MacArthur, 2011; Lee, 2015; Lundstrom & Baker, 2009). Moreover, providing and utilizing feedback from peers can be considered an important skill for students' future academic or professional careers, and therefore can be considered an important learning goal within HE curricula (Liu & Carless, 2006). The second set of arguments 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 (Cho & MacArthur, 2010; Topping, 2009). Currently, more than half of the young people in OECD countries are expected to enroll 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 (Bailey & Garner, 2010; Ballantyne, Hughes, & Mylonas, 2002). Especially in the case of feedback on writing, being relatively time-consuming, such pressures on teaching staff increase the need for alternative formative feedback practices that are both effective and practically efficient.

Prior Research

To the best of 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 HE 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. As part of a larger focus on writing intervention treatments, Graham and Perin (2007) found a strong and positive impact on writing quality when comparing students that were engaged in 'peer assistance' with students that wrote alone. In their study, however, peer assistance also included students cooperating in both planning and composition phases, making it difficult to disentangle specific effects of peer feedback from those of a broader array of cooperative learning activities. For the HE 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 a subsequent increase in research on peer feedback in the thirteen following years, Gielen et al. (2011) deemed such a synthesis still unfeasible as a result of the variation in peer feedback practices and a general lack of comprehensive reporting on those practices:

> Topping was right about the expansion of peer assessment research. The number of studies since his review has doubled, or even tripled. Despite the large number of studies available today, however, the type of review, synthesis or meta-analysis that Topping anticipated is still not feasible today. [...] And the variation in peer assessment practices has only expanded, even rendering his typology inadequate to capture the diversity of peer assessment today. (p. 150)

Around the same time some qualitative review studies into peer feedback have been 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 et al. (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 Strijbos & Sluijsmans, 2010; Topping, 1998, 2010). What these and other review studies (e.g., van Gennip et al., 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 et al., 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 formative peer feedback and students' academic writing performance within the HE context for two reasons. First and foremost, the development of academic writing skills is considered important across HE 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 HE 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

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 modify his or her thinking or behavior for the purpose of learning'. Hence, peer feedback is formative in the sense that it can be utilized by the peer to improve subsequent learning (as measured in the particular study). 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.

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: planning, sentence generation, and revision. Therefore, the current study focuses on HE writing assignments that include such features of the writing process, for example lab 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 HE students. Two sets of research questions are addressed.

Peer feedback effectiveness. Peer feedback has traditionally been compared to alternative feedback sources such as that from teaching staff, both in terms of its outcomes (Cho & Schunn, 2007; Topping, 1998) and in terms of the reliability and validity of these outcomes (e.g., Falchikov & Goldfinch, 2000). Indeed, comparing the effectiveness of a particular practice to practically feasible alternatives is informative for teachers in HE. 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?

Exploration of practically applicable design variables. Gielen et al. (2011) provided an overview of 20 variables that could be considered important for the design and implementation of peer feedback tasks. As the current study's second central aim is to be of practical value for HE 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 HE teaching staff. For this purpose, and borrowing from planned behavior theory (Ajzen, 1991), six HE teachers were interviewed and performed a card-sorting task to rank Gielen et al.'s (2011) variables from 1 (*completely uncontrollable*) to 5 (*completely controllable*). These teachers were from different institutes and disciplines and all were experienced with incorporating peer feedback into their teaching practice. Their perceptions of controllability were then cross-referenced with the prevalence of these design variables across the included studies. This resulted in a focus on two variables that both were reported in the included studies and that were perceived as controllable by the HE teachers: 'student output' (referring to the quantitative/qualitative nature of the peer feedback) and 'assessor constellation' (the number of peer reviewers in particular). Hence, the second set of research questions investigates the impact of peer feedback on academic writing in relation to: (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.

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 HE 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) as measured in the particular study. 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 that (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 & Wilson, 1993).

Search Strategies

Search terminology and databases. The systematic search was conducted via EBSCOhost (including Academic Search Premier, ERIC, PsycARTICLES, Psychology and Behavioral Sciences Collection, and PsycINFO) and Web of Science. Search terms were determined through two complimentary steps. First, prior review studies (e.g., Falchikov & Goldfinch, 2000; Topping, 1998; van Gennip et al., 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 HE 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 ≥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 & Koch, 1977). Disagreements were resolved between the first and second author, resulting in the retraction of one inclusion as was initially 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 having a reference group. As two articles (Sampson & Walker, 2012; Walker & Sampson, 2013) were based on the same data, the study with the largest sample size (Sampson & Walker, 2012) was retained, resulting in the final inclusion of 24 articles (8.4%). 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). Hence, 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 & Wilson, 2001; Wright, 2006) to account for potential a priori between-group differences. Alternatively, they were based on the groups' posttest scores (cf. Lazonder & 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, Keenan, Sandidge, Sippl, and Konheim-Kalkstein (2012), for example, students' papers were graded by two different graders, effectively resulting in two grade-sets for the same writing task. Hence, these grade-sets were averaged before calculating effect sizes.

For studies without a reference group, that is studies which 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 & Wilson, 2001, p. 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., Cheng, Liang, & Tsai, 2015; Sampson & Walker, 2012), effect sizes were based on 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, Hedges, Higgins, & Rothstein, 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, respectively, the nature of the peer feedback and the number of peers engaged with during peer feedback. The data was analyzed using the 'metafor' package (version 2.0-0, Viechtbauer, 2010) 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

Meta-Analytical Assessments of 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. Regarding the effects of peer feedback compared to no feedback, the only two studies including such a comparison (Cho & MacArthur, 2011; Tsai & Chuang, 2013) showed a large 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 (see Figure 1). Regarding the comparison between peer feedback and self-assessment, the composite effect size of the three available studies that directly make this comparison (Cahyono & Amrina, 2016; Diab, 2011; Stellmack et al., 2012) 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 (Birjandi & Tamjid, 2012; Cho & Schunn, 2007; Hartberg, Gunersel, Simpson, & Balester, 2008), resulting in an intermediate sized, though non-significant composite effect size of 0.46 [-0.44, 1.36]. Hence, 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. There was an additional study comparing peer feedback to feedback from teaching staff (Yang, Badger, & Yu, 2006) but no effect sizes were available or could be calculated. It did report 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 was only qualitative in nature. In only one study (Greenberg, 2015) peer feedback was instructed to be merely quantitative (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

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		95% CI	
								g	*SE*$_g$	*lower*	*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; NA = not applicable; n/a= not available or unknown; Gr = Grades/ranking only; Co = Comments only; Both = Grades/ranking + comments; d.o.c. = depending on condition

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 et al., 2015; Sampson & Walker, 2012; Cho & 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

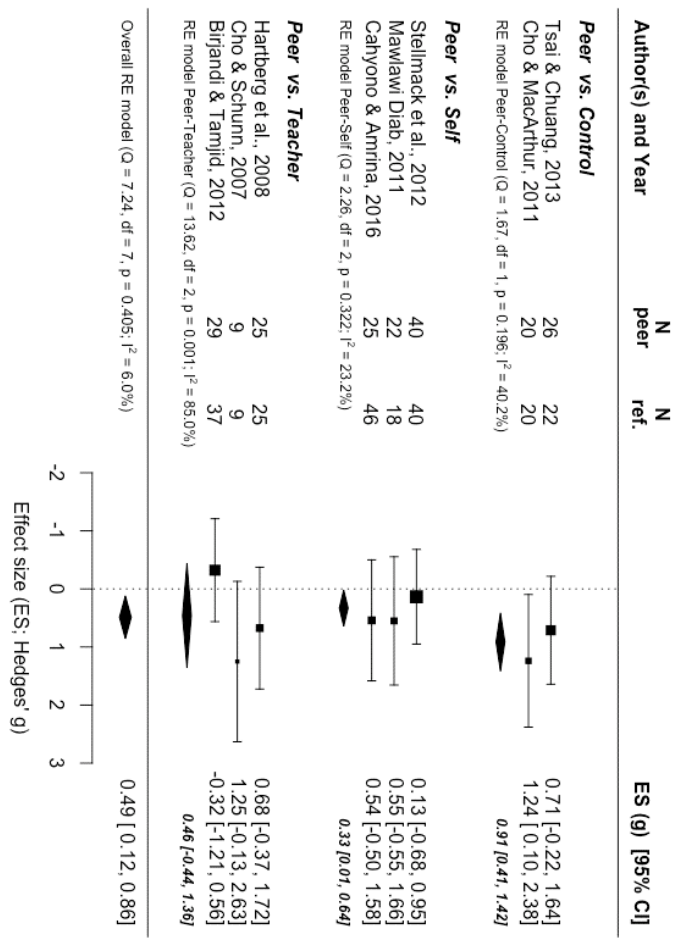

Figure 1. Peer feedback effect sizes for varying reference-group comparisons

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, & Mulder, 2016; Hu & Lam, 2010; Yoshizawa, Terano, & Yoshikawa, 2012; Crossman & Kite, 2012), the respective effect sizes ranged between small and intermediate: 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 peer feedback generally took place in-class (only Noroozi et al., 2016, was both in-class and online). In two studies (Crossman & Kite, 2012; Hu & Lam, 2010) peer feedback was face-to-face, allowing the possibility of 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. Indeed, the variation in students' writing improvement was moderated by the nature of the peer feedback ($\hat{\beta}_{FBnature} = 0.61$, $z = 2.02$; $Q_M(1)$, = 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 during peer feedback. 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 & Tamjid, 2012; Sampson & Walker, 2012) adopted a different procedure, with peer feedback on individual students' academic writing being provided in a groupwise 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 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

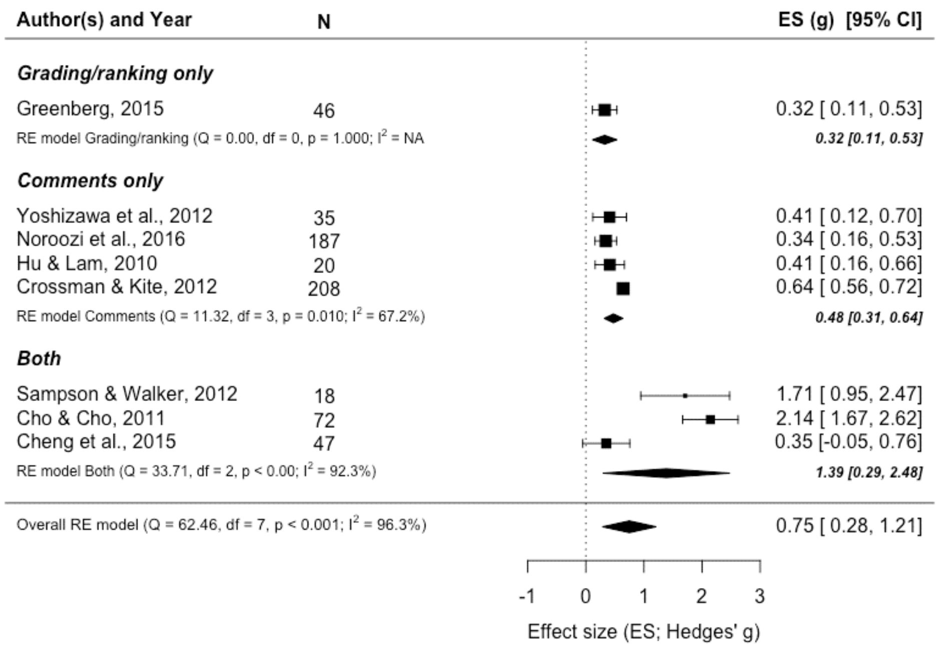

Figure 2. Peer feedback effect sizes for one-group pretest-posttest studies by nature of the peer feedback

sizes may have 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 & Lam, 2010; Yoshizawa et al., 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 & Lam, 2010, students were 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 et al., 2016; Cheng, et al., 2015; Crossman & Kite, 2012; Sampson & Walker, 2012; Cho & 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]). 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 & 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 ($\hat{\beta}_{NRpeers} = 0.60$, $z = 1.27$; $Q_M(1)$, $= 1.62$, $p = .202$, $I^2 = 96.2\%$).

Number of peer reviewers versus contact mode. One unanticipated but noticeable pattern that emerged across all the included studies relates the mode of contact between students (online versus in-class/face-to-face) to the number of peers engaged with. In 73 percent (8 out of 11) of the studies in which peer interaction was in-class/face-to-face, students engaged with no more than one peer, whereas in 90 percent (9 out of 10) of the studies in which peer interaction was online, students engaged with two or more peers. Hence, it appears that the online context facilitates or triggers the inclusion of multiple peers in the peer feedback process. Two studies (Ciftci & Kocoglu, 2012; Novakovich, 2016) directly compared the effects of contact mode (blogs versus face-to-face) on students writing performance. These studies reported large effect sizes for peer feedback through blogs versus face-to-face peer feedback (0.93 [0.38, 1.66] and 0.87 [0.32, 1.49], respectively). In the study by Novakovich (2016), students in both conditions engaged with three peers. For the study by Ciftci and Kocoglu (2012), however, it is unclear with how many peers a student engaged during

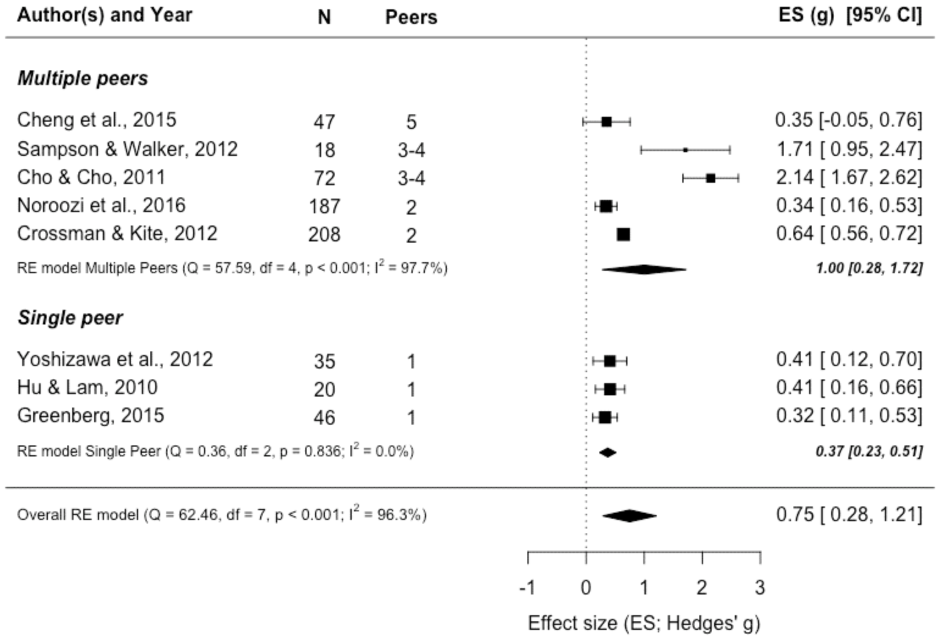

Figure 3. Peer feedback effect sizes for one-group pretest-posttest studies by number of peers engaged with

peer feedback. As a consequence, it still remains an open question how contact mode and the number of involved peers may confound in explaining students' academic writing performance.

Discussion

This study meta-analyzed the effect of peer feedback on the academic writing performance of HE students. Two sets of research questions were addressed. First, the effects of peer feedback on academic writing were analyzed 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. Insofar the limited number of studies allows for a generalization, this finding corroborates more descriptive conclusions of prior qualitative review studies. For example, van Zundert et al. (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 HE 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 HE students' academic writing performance.

The second comparison indicated larger writing improvements for students engaged in peer feedback than for students engaged in some form of self-assessment. However, 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, Hayes, Carey, Schriver, & Stratman, 1986; Patchan & 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, & Sluijsmans, 1999). Also, there 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 & Boud, 1989) and can be largely similar to peer- and teacher assessments with regard to specific aspects of writing assignments (Lindblom-Ylänne, Pihlajamäki, & Kotkas, 2006). In the context of online education, however, self-assessments may be biased (e.g., Admiraal, Huisman, & Pilli, 2015), which should at least prompt thoughtful considerations regarding the utilization of self-assessment for formal assessment procedures.

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. In fact, 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 observation as well. Still, these findings corroborate those of Topping's (1998) qualitative review. Also, these findings are in line with those of Cho and Schunn (2007). 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 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 & Goldfinch, 2000).

There are arguments in favor of teacher feedback (e.g., more expert knowledge) as well as arguments in favor of peer feedback, such as that it induces reflection (e.g., Nicol, Thomson, & Breslin, 2014) and that assessor status may affect critical appraisal of the feedback by the recipient (Strijbos, Narciss, & 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. Hence, 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 HE 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 et al., 2011) in explaining HE students' academic writing performance. Our analysis focused on two specific design variables that HE teachers identified 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 & 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 & Macfarlande-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 & Carless, 2006). At least at first, these two findings appear at odds. Some valuable insights are provided here by Nicol et al. (2014), who reported the arguments of students that either were in favor of or against peer grading. Students in favor 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 that 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 indeed this is the case, it may be possible to have 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 (Cheng et al., 2015; Cho & Cho, 2011; Sampson & Walker, 2012), the weighting of peer grades unfortunately either varied or was unclear. Hence, the weighing 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 & Schunn, 2007) found no significant effects on writing improvement, however. 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) Also, feedback from multiple peers may be more valid and reliable and therefore be preferred over feedback from a single peer (Cho, Schunn, & 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 (e.g., Cho & MacArthur, 2011; Lundstrom & Baker, 2009) studies emphasizing the learning benefits of providing peer feedback. In that light, it seems logical to expect that students' writing improves more as the number of peers with whom they engage increases. 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., Gielen et al., 2011; Topping, 1998, 2010). 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 HE, the current study contributes to the literature by estimating the *extent to which* students' engagement in peer feedback improves their writing performance within this HE context. The results convey two different but interrelated observations. The first observation concerns peer feedback effectiveness on HE 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 et al., 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. Evidenced 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. 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 HE 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 HE 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
The authors would like to thank Stan van Ginkel for the fruitful brainstorms and discussions in the early stages of this study. Also, thanks go out to Benjamin Telkamp for inspiring confidence to continue using the R language.

Statement on Open Data

The anonymized data and syntaxes are accessible via the following link: [URL following upon publication]