Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/65378 holds various files of this Leiden University dissertation.

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

Peer feedback on academic writing

Effects on performance and the role of task-design

Peer feedback on academic writing

Effects on performance and the role of task-design

Proefschrift

ter verkrijging van

de graad van Doctor aan de Universiteit Leiden,

op gezag van Rector Magnificus prof. mr. C.J.J.M. Stolker,

volgens besluit van het College voor Promoties

te verdedigen op woensdag 12 september 2018

klokke 16.15 uur

door

Bart Ananda Huisman

geboren te 's-Gravenhage

in 1983

Promotores

Prof. Dr. J.H. van Driel, The University of Melbourne, Australia

Prof. Dr. P.W. van den Broek

Copromotor

Dr. N. Saab

Promotiecommissie

Prof. dr. W.F. Admiraal

Prof. dr. S.F. Akkerman

Prof. dr. ir. F.J.J.M. Janssen

Prof. dr. J.W. Strijbos, Rijksuniversiteit Groningen

Dr. A.W. Koornneef

Dr. D.M.A. Sluijsmans, Hogeschool Zuyd

Table of contents

General Introduction

1

Context of the Study

Following the agreement between the Ministry of Education, Culture and Science and the Association of Universities in The Netherlands on improving study success at universities, Leiden University initiated the Taskforce Study Success in 2008. The taskforce's primary aim was to provide recommendations for increasing study success and decreasing student attrition across bachelor (BSc) programs, for which it proposed a variety of strategic measures in its 2009 report. These proposed measures related to issues such as assessment policy, student engagement, guidance, and supervision. Important aspects thereof included, among others, students' active participation, feedback, and systematic attention for the skills required to successfully write a BSc thesis. One aspect of Leiden University's measures included the facilitation of educational research to further support evidence-based decision making on these issues. In this context, the current thesis focused on the use of formative peer feedback to enhance students' performance on academic writing tasks. As academic writing is an integral part of curricula across disciplines, insights into the design of formative peer feedback and its effects on students' academic writing performance are informative for a broad array of higher education teachers and educational advisors.

Peer feedback is reported both as being reliable and as being beneficial to the students' development of domain specific skills (e.g., Nelson & Schunn, 2009; Cho & Schunn, 2007; Cho & MacArthur, 2010, 2011; van Zundert, Sluijsmans, & Merriënboer, 2010). In addition to the potential learning benefits, peer feedback can be a practical instructional method. Its potential practical advantages become increasingly salient when considering two interrelating factors: the relatively complex nature of academic writing tasks and higher education teaching staff's available time. The provision of formative feedback on complex, open-ended tasks is a time-consuming procedure for higher education teaching staff. Here, its immediacy, frequency, and volume (Topping, 1998) make peer feedback a potentially efficient alternative formative feedback practice. The salience of these practical benefits increases as student enrollment numbers increase. Both in OECD countries (OECD, 2016) and in the specific context of The Netherlands (CBS, 2017), participation in higher education has steadily increased between 2004 and 2014. Moreover, since the year 2000 the increase in student enrollment has led to increasing student-to-teacher ratios and, consequently, increasing time-pressure for higher education teaching staff (VAWO, 2015, 2015; SoFoKleS, 2016). The growing availability and user-friendliness of – often web-based – applications that facilitate the peer feedback process (see Luxton-Reilly, 2009, for an overview) partly ameliorates these effects of increasing student-to-teacher ratios. In addition, such applications can increase higher education (HE) teaching staff's options with respect to the instructional design of peer feedback tasks, facilitate access to empirical data, and as such stimulate empirical research. In summary, peer feedback could simultaneously be beneficial to students' academic writing performance and could be a practical alternative formative feedback practice for which the opportunity for empirical research is growing.

In this light, it is not surprising that research into peer feedback on academic writing in higher education has gained momentum in the last two decades (see Figure 1). Despite specific calls by, for example, Topping (1998) and Strijbos and Sluijsmans (2010), the proportion of well-controlled peer feedback studies that focus on students' learning outcomes has remained low. This thesis contributes to the knowledge base by investigating the relation between peer feedback and higher education students' writing performance through a quantitative lens. In doing so, this thesis aims to advance our knowledge on the extent to which peer feedback impacts students' academic writing and on how specific aspects of task design relate to student performance.
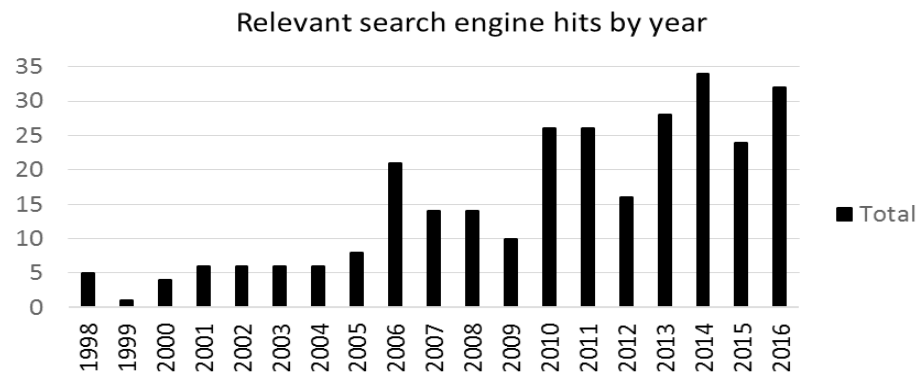
Relevant search engine hits by year

Figure 1. Journal articles on peer feedback with academic writing in higher education, sorted by year. Frequencies based on search criteria and initial selection of relevant articles in meta-analysis (Chapter 2).

Conceptual Framework

Terminology

Formative peer feedback. In his seminal article on peer assessment in colleges and universities, Topping (1998) defined peer assessment as 'an arrangement in which individuals consider the amount, level, value, worth, quality or success of the products or outcomes of learning of peers of similar status' (p. 250). The clarity of this definition with respect to the equal status between learners and the diversity of aspects that can be assessed is regarded as valuable for the current thesis. However, this thesis focuses on formative peer feedback, and Topping's definition does not explicate that the assessment is fed back to the assessee for the purpose of enhancing his or her learning. Therefore, this thesis additionally borrows from Shute (2008), who defined formative feedback as 'information communicated to the learner that is intended to modify his or her behavior for the purpose of improving learning' (p. 154). Taken together, formative peer feedback is defined here 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*. Because this definition encompasses *all* task-related information that is communicated between peers, formative peer feedback may

range from basic grades or rankings to elaborate comments, as long as it can be used to improve subsequent learning. Consequently, this definition of formative peer feedback encompasses both 'peer feedback' and 'peer assessment', insofar these reflect different practices in the literature.

Academic writing assignments were defined as writing assignments within the particular academic curriculum that included the hierarchical organization of writing goals that students could accomplish through iterative planning, writing and revising (e.g., Hayes & Flower, 1987). In the empirical studies of this thesis, these considered essay assignments on theoretical concepts (chapters three and five) or theoretical claims or dilemmas (chapter four). In the empirical on-campus studies (chapters three and five) these assignments were written in students' native language (Dutch). Students were required to write in English in the open online course (chapter four).

Framework of peer feedback design variables

In order to facilitate a systematic investigation of the effects of peer feedback, Topping (1998) provided a typology of 17 variables that he considered to be the most important varying parameters across studies. These variables addressed, for example, the directionality of contact between peers (e.g., one-way or reciprocal), the ability of peers (same or different), and how contact was arranged (e.g., distance or face-to-face). One conclusion of Topping was that future research would contribute most when it explicitly described participants' characteristics, the design of the peer feedback task and the research design. Later studies reorganized and/or expanded Topping's set of variables in different ways. For example, van den Berg, Admiraal and Pilot (2006b) rearranged these 17 variables to be included in one of four clusters, whereas van Gennip, Segers and Tillema (2009) grouped these same variables into three clusters. In what currently is the most comprehensive overview of variables that are relevant for the design of peer feedback tasks, Gielen, Dochy and Onghena (2011) expanded the total number of variables to 20. They rearranged these variables to fit one of five clusters: 1) decisions concerning the use of peer assessment (e.g., setting, objectives), 2) the link between peer assessment and other elements of

the learning environment (e.g., alignment with curriculum), 3) the interaction between peers (e.g., role of assessed student), 4) the composition of groups (e.g., ability matching) and 5) procedural management (e.g., feedback format, training/guidance).

Research Aims

The current thesis builds upon these developments and this framework with three general aims in mind. First, this thesis aims to investigate the empirical evidence for the effects that peer feedback has on students' academic writing performance, and to investigate what the role of specific peer feedback design variables is in explaining differences in students' writing performance. Hence, the overarching research question is: *To what extent does formative peer feedback impact higher education students' academic writing performance, and what is the influence of specific aspects of peer feedback task-design on writing performance?* The second aim of this thesis is to be of practical value for higher education teachers and educational advisors. This means that this thesis focuses on peer feedback design variables that tend to be relatively controllable for higher education teaching staff. The third aim concerns the research *approach,* and follows the calls by, for example, Topping (1998, 2010) and Strijbos and Sluijsmans (2010). Specifically, the current thesis aims to incorporate well-controlled research designs. Also, it aims to be optimally transparent in order to facilitate the process of cumulative knowledge building. This means that this thesis is as explicit as possible in its descriptions of participant characteristics, peer feedback task- and research design, and that the anonymized data and syntaxes are openly accessible online.

Overview of the Thesis

The overarching research question is addressed in chapters two through five, whereas chapter six builds on the implications of chapter five by describing an instrument to assess students' beliefs about peer feedback. Chapter two describes a meta-analysis that was conducted to assess the impact of formative peer feedback in higher education students' academic writing. A total of

287 full-texts were assessed, of which 24 articles were considered eligible for inclusion. Two sets of analyses were conducted. First, the impact of peer feedback was compared to a baseline (no feedback), to self-assessment, and to feedback from teaching staff. Second, two mixed-effects model analyses were conducted to assess the moderating role of two peer feedback design variables that were perceived as controllable by higher education staff: the nature of the peer feedback (grades, comments or a combination of both) and the number of peers a student engaged with during peer feedback (one or multiple).

Chapter three describes an empirical study investigating the effects of students' ability match on the nature of the peer feedback and subsequent writing performance. Contrasting two competing theoretical rationales in the literature, 94 first-year undergraduate students in an Education & Child Studies course were anonymously matched into either a homogeneous dyad (with a similar ability peer) or a heterogeneous dyad (with a different-ability peer). The relation between dyad composition and the nature of the peer feedback was explored, with the latter being defined in terms of peer feedback *aspects* (content, structure, style) and peer feedback *functions* (analysis, evaluation, explanation, revision). Finally, it was explored how dyad composition and peer feedback nature were related to students' writing performance.

Chapter four follows up on chapter three by exploring how the ability match between participants in a massive open online course (MOOC) related to their performance on a subsequent writing task. As the number of MOOCs continues to grow, and as the number of MOOC participants tends to run in the thousands, peer assessment is often applied because it is a scalable way to assess complex, open-ended tasks. In this chapter, post-hoc analyses were conducted based on the data from a MOOC on Terrorism and Counterterrorism by Leiden University. Based on prior performance indices within the MOOC, 565 participants were categorized as relatively low, intermediate or high ability authors. In contrast to the on-campus study in chapter three, participants wrote two different essays, each of which was assessed by 4-6 randomly assigned peers. Regression analyses were conducted to assess a) whether the average ability of peer reviewers had influenced the extent to which participants improved their

writing between the first and second essay and b) whether the average ability of peer reviewers differentially impacted the writing performance increase of low, intermediate and high ability authors.

Chapter five aims to disentangle the effects that either providing or receiving peer feedback can have on students' academic writing performance. More specifically, this chapter contrasted the writing performance increase of peer feedback providers and –receivers in the context of an authentic academic writing task on campus. In total, 83 first-year undergraduate students in an Education and Child Studies course were assigned to either the role of peer feedback provider or –receiver. All students first submitted a draft essay, engaged in anonymous, online peer feedback in accordance with their assigned role of peer feedback provider or –receiver, and submitted a final essay after the peer feedback phase. In addition to the direct relation between feedback role and writing performance, the role of the peer feedback nature and students' perceptions thereof were investigated. In particular, it was explored which peer feedback aspects and functions influenced the extent to which students perceived the received peer feedback to be adequate, and to what extent they were willing to improve their writing based upon it. The outcomes and implications are discussed in relation to the training of students for the peer feedback process and in relation to students' support for, and beliefs about peer feedback.

Chapter six builds upon the discussion and implications of chapter five, and describes the development of a questionnaire to assess students' peer feedback beliefs. Students' beliefs about peer feedback are likely to be shaped by their cumulative experience of multiple peer feedback occurrences over time and, in turn, are likely to influence their perceptions and behavior. However, a comprehensive measure of students' peer feedback beliefs appears to be missing, and prior studies vary in terms of both approaches and outcomes. This makes it difficult to align research findings and to draw meaningful inferences upon these findings. To this end, the concise though comprehensive Beliefs about Peer Feedback Questionnaire (BPFQ) was developed. Based on the different themes covered in the literature, four scales were conceptualized: 1) valuation

of peer feedback as an instructional method, 2) valuation of peer feedback as an important skill, 3) confidence in own peer feedback quality and 4) confidence in quality of received peer feedback. The construct validity of the questionnaire was separately tested in an exploratory phase, based on a cohort of 219 second-year students in Biopharmaceutical Sciences, and a confirmatory phase, based on a first-year cohort of students in Education and Child Studies (N=121).

Taken together, these chapters aim to enhance our current understanding of the effects that formative peer feedback may have on higher education students' academic writing performance, as well as our understanding of the role that specific aspects of peer feedback task-design have in explaining differences in students' writing performance. These chapters address different aspects of peer feedback task-design as embedded in the framework of Gielen, Dochy, & Onghena (2011), including the nature of peer feedback (chapters two, three and five), the number of peers with whom a student engages (chapter two), the ability match between students (chapters three and four) and students' feedback role (chapter five). Here, an important theoretical contribution is their relatively controlled design combined with their focus on students' academic writing performance. An important practical contribution of these chapters follows from the relatively controllable nature of these design aspects of peer feedback tasks. Generally speaking, this should allow higher education teaching staff to incorporate these findings into their teaching practice, which could contribute to the development of students' academic writing skills.

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

2

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 ($g = 0.46\ [\text{-}0.44,\ 1.36]$). 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 g	SE_g	95% CI 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: * = <u>selected</u> 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
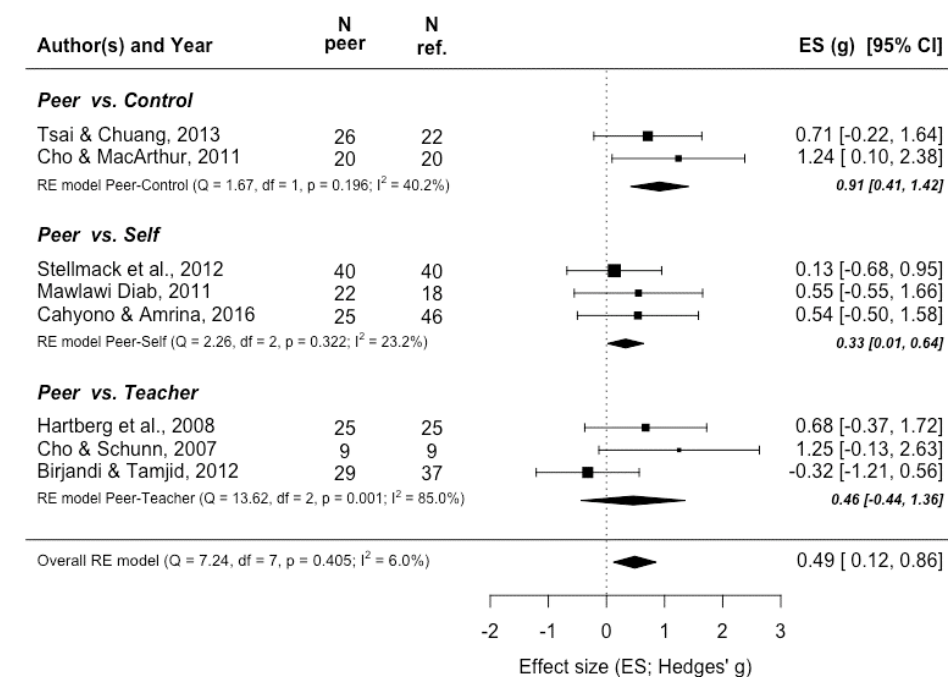

Figure 1. Peer feedback effect sizes for varying reference-group comparisons

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

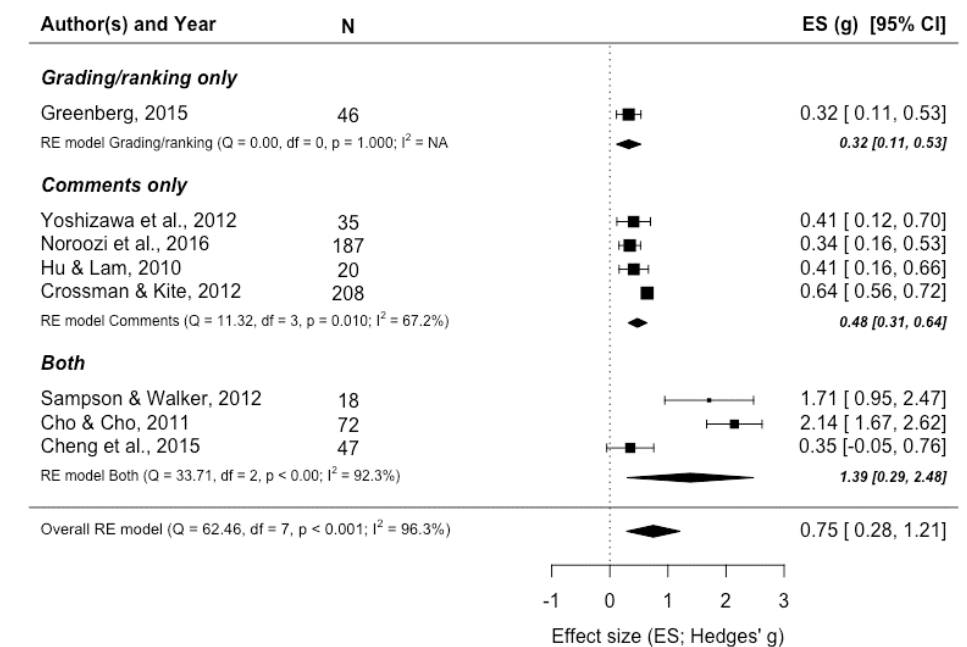| Author(s) and Year	N		ES (g) [95% CI]
Grading/ranking only			
Greenberg, 2015	46		0.32 [0.11, 0.53]
RE model Grading/ranking (Q = 0.00, df = 0, p = 1.000; I² = NA)			*0.32 [0.11, 0.53]*
Comments only			
Yoshizawa et al., 2012	35		0.41 [0.12, 0.70]
Noroozi et al., 2016	187		0.34 [0.16, 0.53]
Hu & Lam, 2010	20		0.41 [0.16, 0.66]
Crossman & Kite, 2012	208		0.64 [0.56, 0.72]
RE model Comments (Q = 11.32, df = 3, p = 0.010; I² = 67.2%)			*0.48 [0.31, 0.64]*
Both			
Sampson & Walker, 2012	18		1.71 [0.95, 2.47]
Cho & Cho, 2011	72		2.14 [1.67, 2.62]
Cheng et al., 2015	47		0.35 [-0.05, 0.76]
RE model Both (Q = 33.71, df = 2, p < 0.00; I² = 92.3%)			*1.39 [0.29, 2.48]*
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 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
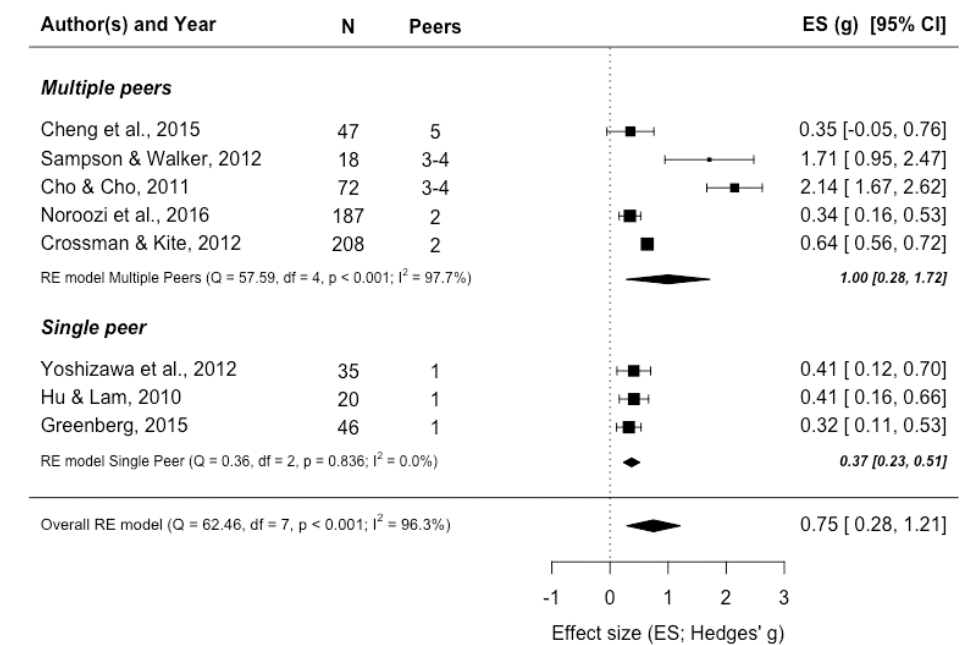

Figure 3. Peer feedback effect sizes for one-group pretest-posttest studies by number of peers engaged with

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.

2

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

Statement on Open Data

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

Exploring the relation between students' ability match, the nature of peer feedback and essay performance

3

Abstract

There does not appear to be consensus on how to optimally match students during the peer feedback process: with same-ability peers (homogeneously) or different-ability peers (heterogeneously). In fact, there appears to be no empirical evidence that either homogeneous or heterogeneous student matching has any direct effect on writing performance. The current study addressed this issue in the context of an academic writing task. Adopting a quasi-experimental design, 94 undergraduate students were matched in 47 homogeneous or heterogeneous reciprocal dyads, and provided anonymous, formative peer feedback on each other's draft essays. The relations between students' individual ability or dyad composition, the nature of the peer feedback, and writing performance were investigated. Neither individual ability nor dyad composition directly related to writing performance. Also, the nature of the feedback did not depend on students' individual ability or dyad composition, although trends in the data suggest that high ability reviewers provided more content-related feedback. Finally, the nature of the peer feedback was not related to writing performance, and authors of varying ability levels benefited to a similar extent from peer feedback on different aspects of the text. The results are discussed in relation to their implications for the instructional design of academic writing assignments that incorporate peer feedback.

Keywords: peer feedback; academic writing; higher education; feedback quality; student ability match.

Introduction

Research on peer feedback in education has expanded in the last two decades. This has increased our knowledge on the reliability and validity of peer feedback in primary, secondary and post-secondary education (Cho, Schunn, & Wilson, 2006; Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Topping, 2009), and with respect to the variables that are important for the design and implementation of peer feedback (e.g., Topping, 1998; van Zundert, Sluijsmans, & van Merriënboer, 2010). However, regarding structural features such as feedback group composition (see van Gennip, Segers, & Tillema, 2009), there does not yet appear to be consensus on how to optimally match students in terms of ability.

This study focuses on the ability match between students during peer feedback on academic writing in a higher education context. There are three reasons for this focus. First, it seems fair to conclude from the literature that peer feedback can be beneficial to higher education students' learning, and that students can perceive these benefits (Hanrahan & Isaacs, 2001). Students can expect reliable and valid assessments from each other regarding the quality of their work (Cho et al., 2006; Falchikov & Goldfinch, 2000). Also, the process of providing feedback can help students improve their writing performance (Lundstrom & Baker, 2009). Providing peer feedback prompts a reviewer to go beyond mere problem detection, engaging him or her in problem diagnosis and in suggesting solutions (Patchan & Schunn, 2015). Second, being able to provide feedback to peers and utilize feedback from peers can be considered important skills in students' subsequent academic or professional career. Importantly, both students' peer assessment skills and their attitudes towards it can be positively influenced through preparation and practice (van Zundert et al., 2010). Third, academic writing skills are considered important across disciplines and are an integral part of higher education curricula. Given the sometimes large student-to-staff ratios in higher education institutes, however, adequate instructor feedback on academic writing tasks can be a challenge. One aid comes from (web-based) applications that facilitate the peer feedback process (see Luxton-

Reilly, 2009, for an overview). With the increasing availability and usability of such applications, the peer feedback process becomes easier to design and implement for academic teaching staff. This may increase the extent to which peer feedback is implemented within academic writing tasks.

Student Ability Matching

Another benefit of applications that facilitate the implementation of peer feedback is the potential array of possibilities in terms of instructional design. For example, it should be possible to automatically match students on certain criteria, such as ability. Although the potential benefits of student matching have already been discussed in 1998 by Topping, there does not appear to be a clear consensus on whether students should be matched with similar ability peers (homogeneously) or with peers or different ability (heterogeneously).

Regarding the homogeneous matching of students, Topping (2009) prescribes matching students with same-ability peers. In addition, King (1997) argues that beneficial socio-cognitive conflict is more likely between equal peers and that higher level learning is more likely to be accomplished when ideas are exchanged on an equal basis. Also, a mindful, critical appraisal of received feedback may be critical to its effectiveness, which could be stimulated by the uncertainty that the peer's status induces (Gielen et al., 2010). An experimental study by Strijbos, Narciss, and Dünnebier (2010) investigated the relation between peer feedback content and sender's (perceived) competence on the one hand, and feedback perceptions and revision on the other hand. Their results suggest that status differences between peers may have negative effects; receiving elaborate, specific feedback from high ability peers was related to more negative affect and less effective text revision. One possible explanation suggested by the authors is that elaborate, specific feedback from high competence peers rendered students to become passive and overly reliant on the feedback they received. These theoretical arguments and empirical findings support the suggestion to match students in a homogeneous manner. However, they do not provide empirical evidence for a direct relation between homogeneous matching with peer feedback and writing performance.

Regarding the heterogeneous matching of students, it has been found that higher ability authors tend to focus more on global issues, detect more problems, and are more likely to use effective strategies for revision than lower ability authors (e.g., Patchan & Schunn, 2015). As a result, they may provide more critical peer feedback than lower ability authors do . Patchan, Hawk, Stevens, and Schunn (2013) differentiated between feedback comments that focused on 'high prose' (flow, logic or insight), 'low prose' (lower-level writing issues such as grammar), or 'substance' (issues fixable only with content knowledge). They found that the feedback received by low ability authors was qualitatively different when they were matched with a high ability reviewer (heterogeneous match), compared to a low ability reviewer (homogeneous match). Specifically, low ability authors received and implemented more 'low prose' and 'substance' feedback from high ability reviewers. High ability authors received similar types of feedback, irrespective of reviewer ability. A similar trend was reported for provided solutions. Because feedback containing explicit criticism and suggestions for improvement is likely to contribute to feedback implementation and performance (Nelson & Schunn, 2009), these arguments support matching students heterogeneously. Here too, however, they do not provide empirical evidence for a direct relation between heterogeneous matching with peer feedback and students' writing performance.

Defining Student Ability and Nature of Peer Feedback

Student ability. Student ability has been defined in different ways in prior research. Generally, a distinction can be made between students' task-related ability (e.g., writing skills) and students' ability to provide peer feedback and/or assess others' work (e.g., use of criteria, see for example van Zundert et al., 2010). The current study matched students in terms of task-related ability, i.e. their scores on a preceding essay assignment, for four reasons; availability, similarity, proximity, and validity. First, this ability indicator was both available and practically applicable. Moreover, comparable ability indices are likely to be available in other higher education institutes. Second, this preceding academic writing assignment was similar to the academic writing assignment central to

the current study. Third, the assignment was part of an immediately preceding course in the same curriculum, making it an up-to-date indicator of students' academic writing ability. Fourth and finally, although it is not self-evident that an able writer also is an able reviewer, it is plausible that writing and reviewing ability are interrelated. A rationale for this is provided by Patchan and Schunn (2015), who identified conceptually identical elements between writing and providing feedback on writing: task definition, problem detection and diagnosis, and selection or revision strategy. This overlap in cognitive processes supports the notion that students' ability to write and students' ability to review each other and provide feedback indeed are interrelated, and that high ability writers can be expected also to be high ability reviewers.

Peer feedback nature. As is the case with student ability, the nature of peer feedback has been defined in multiple ways in the literature. Definitions range from relatively simple categorizations such as holistic feedback versus specific feedback (Lin, Liu, & Yuan, 2001) to more elaborate categorizations such as proposed by Nelson and Schunn (2009) differentiating between summarization, specificity, explanation and scope. The current study adopts the definition of peer feedback nature as used by van den Berg, Admiraal, and Pilot (2006c), which includes the aspect of the text to which the feedback relates (content, structure, style) and the function the feedback comments serve in relation to the text (analysis, evaluation, explanation, revision). An important reason for this choice is that the feedback aspects were aligned with both task instructions and assessment criteria.

In summary, theoretical accounts and empirical findings on how to optimally match students in terms of ability vary and sometimes appear contradictory. To our knowledge, there are no studies that address the direct link between student matching, the nature of peer feedback, and writing performance. Specifically, there appears to be no empirical evidence that either homogeneous or heterogeneous student matching has any effect on writing performance. Moreover, there is a need for (quasi-)experimental studies investigating the effects of peer assessment (Strijbos & Sluijsmans, 2010). Adopting an exploratory approach, the current quasi-experimental study specifically focuses

on the relation between the students' ability match, the nature of the peer feedback, and their performance on an academic writing task.

Research Questions

This issue was investigated in the context of an essay assignment within a first-year introductory course Education and Child Studies. In this context, student matching was reciprocal, meaning that the students within a particular dyad provided feedback to their dyad member and received feedback from that dyad member. Three main research questions are formulated (See Figure 1). Research question 1 is: 'to what extent is student ability in, and dyad composition of reciprocal dyads related to authors' increase in essay performance?' Research questions 2 and 3 explore this relation in more detail. Specifically, research question 2 is: 'to what extent is student ability in reciprocal dyads related to the nature of the peer feedback?' Here, two sub-questions are formulated. The first focuses on reviewers' individual ability; a) 'what is the relation between reviewer ability and the nature of the peer feedback they provide?' The second sub-question takes into account the interdependence of authors and reviewers within the dyads; b) 'to what extent does the nature of the provided peer feedback vary between differently composed dyads?' Finally, research question 3 focuses on the relation between the nature of the peer feedback and essay
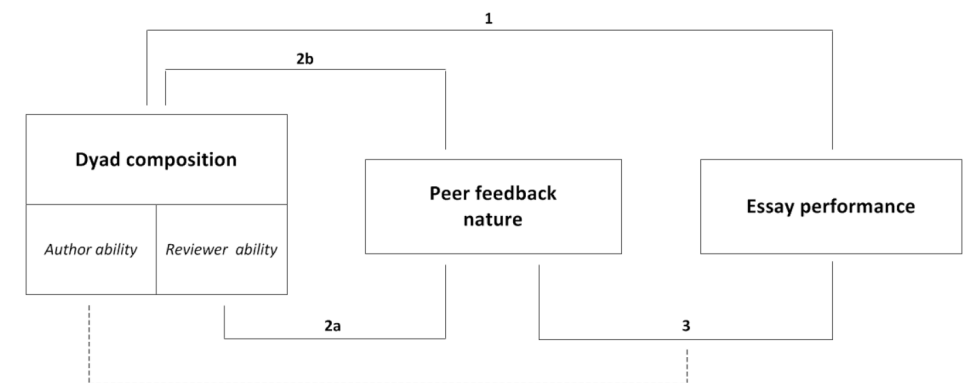

Figure 1. Graphical representation of research questions.

performance: 'to what extent is the nature of the received peer feedback related to essay performance, and to what extent is this relation moderated by author ability?'

Method

Participants and Procedure

Participants were undergraduate students in a first-year introductory course Education and Child Studies ($N = 220$) at a large research-intensive university in The Netherlands. In total, 121 students both agreed to participate and submitted all assignments. 94 Students were included in the study as they were part of a dyad in which both students participated. The mean age was 19.8 years ($SD = 1.67$), with 88 females and 6 males. Students had three weeks to work on a draft essay assignment, followed by one week for peer feedback and another week to produce a final version based on the draft and received peer feedback. Peer feedback was formative, and was provided anonymously and reciprocally within dyads through a virtual learning environment (Turnitin; e.g. Buckley, & Cowap, 2013).

Essay Assignment, Criteria, and Grading

The essay assignment was instructed to be about 500-750 words excluding references. Students were free to choose one of two essay topics: one in the field of Family Pedagogy ('FP') or one in the field of Educational Sciences ('ES'). For each topic, two scientific articles were provided. The submission of a (serious) draft essay, a final essay, and the provision of adequate peer feedback were mandatory course requirements.

Peer feedback guidelines and criteria were provided through a plenary meeting and digital handouts. Essay grades were assigned by the teaching staff according to the following assessment criteria: Content (30%), Structure (20%), Writing style (20%), Referencing (20%), and Presentation and spelling (10%). Within the context of this study, writing style, referencing and spelling

were taken together and categorized as elements of Style. Grades ranged from 1 (lowest possible score) to 10 (highest possible score), and grades on the final essay versions were communicated with the students, whereas draft essays were graded for research purposes only.

Essay grading. Draft essays were graded by one trained research assistant, whereas final essays were graded by four teaching assistants. The research assistant was trained by one of the teaching assistants and the first author. Inter-rater agreement between the trained research assistant and the teaching assistant was calculated, based on a subset of 44 draft essays, resulting in high inter-rater agreement ($r = .77$, $p < .001$). Moreover, average scores were similar ($t(43) = 0.07$, $p = .946$). Thus, grades assigned by these two raters provided comparable measures of essay quality. Both graders were blind to the matching condition of the students, but were aware of the manuscripts being drafts or final versions. This was not considered problematic, however, because all graders were instructed to grade the manuscripts using the same standards, and the analyses focused on relative improvement across students (cf. Cho & MacArthur, 2010).

Participant Grouping

Dyads were formed by matching students in terms of their ability, defined as students' performance on a similar essay assignment from a directly preceding introductory course in the same curriculum. In the remainder of this study, this ability indicator will be used both in relation to students' role as author and as reviewer.

Within each topic group, students were first rank-ordered on ability, after which they were alternately assigned to one of two conditions (Matching Type): a homogeneous condition (to be matched with a similar ability peer) or a heterogeneous condition (to be matched with a peer of different ability). Following this procedure, students in the two Matching Type conditions were optimally comparable in terms of ability, both containing high- and low ability students across the entire range of ability. Next, within the homogeneous condition, dyads were formed by pairing students adjacent on ability. Within the heterogeneous condition, a split-half procedure was conducted to differentiate

between higher and lower ability students. A 'moving window' procedure was applied to pair students from the top and bottom half, thereby keeping the ability difference within heterogeneous dyads as constant as possible.

Between topic groups, irrespective of Matching Type, higher ability students in the FP group ($N = 32$, $M = 7.75$, $SD = 0.75$) and the ES group ($N = 15$, $M = 7.64$, $SD = 0.73$) scored similarly on the preceding essay ($t(45) = 0.47$, $p = .639$). For the lower ability students, mean scores for those in the FP group ($N = 30$, $M = 5.54$, $SD = 0.90$) and those in the ES group ($N = 17$, $M = 5.30$, $SD = 1.08$) were similar as well ($t(45) = 0.47$, $p = .639$). Within topic groups, higher and lower ability students significantly differed in both the FP group ($t(60) = 10.47$, $p < .001$) and the ES group ($t(30) = 7.25$, $p < .001$).

Measures and Instrumentation

Peer feedback nature was defined in terms of feedback aspects and feedback functions, in line with van den Berg et al. (2006c). *Feedback aspects* concerned the aspects of the text to which the feedback related, distinguishing between content, structure, and style. Here, 'Content' referred to clarity of the problem, argumentation, and the relevance of the presented information. 'Structure' referred to the internal consistency of the text, such as that between the problem statement, the presented arguments, and the discussion. 'Style' referred to grammar, spelling, language use and referencing. *Feedback functions* concerned the function that feedback comments served in relation to the essay in question, distinguishing between 'Analysis', 'Evaluation', 'Revision', and 'Explanation' (Flower, Hayes, Carey, Schriver, & Stratman, 1986; van den Berg et al., 2006c). Feedback comments were coded 'Analysis' if they concerned the meaning of the text or the reviewer's perceived understanding thereof. These reviewer comments were regularly phrased as questions, such as 'What do you mean with […]?'. Further, 'Evaluation' referred to feedback comments that included explicit or implicit quality statements. 'Revision' referred to explicit suggestions for improvement of the text, or implicit suggestions for improvement that included at least a direction for a solution (e.g. 'these references are not adhering to APA guidelines'). Finally, 'Explanation' referred to arguments that supported evaluative comments or suggestions for improvement.

Coding procedure. The nature of the peer feedback was coded in two steps. First, the peer feedback was coded in terms of feedback aspects. Second, every aspect-segment was also coded as having one or more feedback functions (thus allowing for multiple feedback functions per feedback aspect). Inter-rater agreement for both feedback aspects and functions was determined based on the judgment of two coders. Randomly chosen draft essays on which peer feedback was provided were independently coded for feedback aspects, and agreement was calculated. Having reached acceptable agreement, the remaining peer feedback was coded for feedback aspects by one coder. This procedure was repeated for feedback functions.

Inter-rater agreement for feedback aspects was calculated based on a random sample of 17 essays. Agreement was moderate for Structure ($k = .59$, 95% CI $[.38, .80]$) and substantial for Content ($k = .64$, 95% CI $[.50, .78]$) and Style ($k = .78$, 95% CI $[.69, .87]$). Using the coded feedback aspects as units of analysis, inter-rater agreement for feedback functions was calculated on another random sample of 10 essays. Agreement was moderate for Explanation ($k = .57$, 95% CI $[.33, .81]$), substantial for Analysis ($k = .70$, 95% CI $[.51, .90]$) and Evaluation ($k = .73$, 95% CI $[.61, .84]$), and almost perfect for Revision ($k = .85$, 95% CI $[.76, .93]$) (Landis & Koch, 1977).

Analyses

Research question 1: To what extent is student ability in, and dyad composition of reciprocal dyads related to authors' increase in essay performance? First, the direct relation between Performance Increase on the one hand and Author Ability or Reviewer Ability on the other hand was explored. Performance Increase was defined as the difference between an author's score on the draft essay and the final version of the essay. Two linear regressions were performed with Performance Increase as dependent variable and either Author Ability or Reviewer Ability as independent variable. In terms of the ability match between authors and reviewers, an analysis of variance (ANOVA) was performed with Performance Increase as dependent variable and Dyad Composition as independent variable. Dyad Composition was defined as one of four types of

ability matches between an author and a reviewer. With homogeneously matched students, this refers either to a match between two higher ability students or to two lower ability students. With heterogeneously matched students, this refers either to a low ability author matched with a high ability reviewer or vice versa. In case a significant relation with Dyad Composition was found, post hoc comparisons were performed to identify differences in Performance Increase for the differently composed dyads.

Research question 2: To what extent is student ability in reciprocal dyads related to the nature of the peer feedback? To test the effect of Reviewer Ability on the nature of the provided feedback, a multivariate analysis of variance (MANOVA) was performed with Reviewer Ability (high versus low) as independent variable and Feedback Nature as dependent variable. Feedback Nature was defined as the frequency in which the 12 combinations of feedback aspects (Content, Structure, Style) and functions (Analysis, Evaluation, Explanation, Revision) occurred. See Table 1 for an overview. Subsequent ANOVAs on specific Aspect-Function combinations were performed where appropriate. To test the effect of the ability match between authors and reviewers on the nature of the peer feedback, a MANOVA was performed with Dyad Composition as independent variable and Feedback Nature as dependent variable. Again, subsequent ANOVAs on specific Aspect-Function combinations were performed where appropriate.

Research question 3: To what extent is the nature of the received feedback related to essay performance, and to what extent is this relation moderated by author ability? To test the effect of feedback nature on essay performance, a hierarchical linear regression analysis was performed with Final Essay Performance as the dependent variable. Author Ability and Draft Essay Performance were included as independent variables in step 1, followed by received feedback comments on aspects of Content, Structure, and Style in step 2. In the third and final step, the interaction terms between, on the one hand, Author Ability and, on the other hand, the received feedback comments on aspects of Content, Structure, and

Style were added to assess the extent to which the relation between feedback nature and essay performance is moderated by author ability[1].

Results

Manipulation Check
Overall, the preceding essay assignment appeared to be significantly related to the quality of students' draft essays before the peer feedback phase (r_s = .24, p = .020). With respect to student ability matching, the intention was to create homogeneous and heterogeneous dyads. The average ability difference between students in homogeneous dyads (N = 25, M = 0.12, SD = 0.17) was significantly smaller than that between students in heterogeneous dyads (N = 22, M = 2.27, SD = 0.39), $t(45)$ = 24.82, p < .001. However, the difference within homogeneous dyads did not equal zero ($t(24)$ = 3.57, p = .002). Thus, the two Matching Type conditions differed from each other as intended, although, on average, there still was a minimal difference in ability within homogeneous dyads.

Feedback Nature and Quantity
For the 94 included draft essays, 1580 peer feedback segments were coded as distinct feedback aspects, averaging 16.81 segments per essay (see Table 1).

[1] In our definition of the nature of peer feedback, feedback aspects related to either the content, structure, or style of the text. This was aligned with the components of the rubric used to assess the final essays (e.g. Content was weighted for 30%, Structure for 20%, and Style aspects weighted for 50% in calculating overall essay grades). We are aware that, given these differences between weights, and given that only a single composite final essay grade was available, caution is in place when comparing the impact of these various feedback aspects on writing performance (effect sizes are restrained proportionately to the relative weights attributed to these three aspects). Therefore, an additional residualizing procedure was conducted in which three separate dependent variables were created: content-related, structure-related, and style-related Final Essay Performance. This allowed for a comparison of the separate effects that the feedback comments on aspects of Content, Structure and Style had on students' Final Essay Performance on those particular aspects of the texts. Because no significant relations were found, the different weights of these aspects were not further attended to in the principal analyses.

Table 1. *Provided* feedback aspects and feedback functions

		(1) Analysis			(2) Evaluation			(3) Explanation			(4) Revision			(5) Total		
	Reviewer ability¹	Low	High	Total	Low	High	Total	Low	High	Total	Low	High	Total	Low	High	Total
Aspects	**(A) Content**	43	64	107	96	135	231	66	104	170	60	118	178	265	421	686
	(B) Structure	3	5	8	63	75	138	46	44	90	35	38	73	147	162	309
	(C) Style	8	9	17	59	75	134	108	119	227	379	375	754	554	578	1132
	Total	54	78	132	218	285	503	220	267	487	474	531	1005	966	1161	2127

¹ = Low ability reviewer N = 46, High ability reviewer N = 48
Frequencies based on 1580 feedback Aspects, with multiple feedback functions allowed per aspect.

In terms of the average number of provided feedback segments, higher ability students ($N = 48$, $M = 17.15$, $SD = 11.79$) and lower ability students ($N = 46$, $M = 16.46$, $SD = 8.87$) did not differ ($t(92) = -0.32$, $p = .749$). In general, analytical feedback comments were rare, whereas suggestions for improvement occurred frequently. Students predominantly made such suggestions for improvement about aspects of writing style, however, and to a much lower extent about content-related or structural aspects of the essays. Whereas feedback comments about the content or structure of the text were generally evaluative, feedback comments about stylistic aspects predominantly were suggestions for improvement.

Student Ability, Dyad Composition and Performance Increase

In general, there was improvement between scores for draft versions ($M = 6.51$, $SD = 1.70$) and scores for final essays ($M = 7.04$, $SD = 0.94$), $t(93) = 2.91$, $p = .005$. Table 2 appears to indicate that the academic writing performance of lower ability students may increase more than that of higher ability students, irrespective of dyad composition. However, performance increase did not depend directly on author ability ($\beta = -0.16$, $p = .117$, $\Delta R^2 = .03$) or reviewer ability ($\beta = -0.02$, $p = .837$, $\Delta R^2 = .00$). Most importantly, dyad composition appeared unrelated to students' essay performance increase ($F(3, 90) = 0.850$, $p = .470$, $\eta_p^2 = .03$). Thus, performance increase was neither related to authors' or reviewers' individual ability, nor to the composition of the dyad.

Table 2. Essay performance and dyad composition

Dyad Composition	N	Draft essay grade		Final essay grade		Performance increase	
		Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
Low ability author & Low ability reviewer	24	6.16	1.46	7.05	0.95	0.89	1.69
Low ability author & High ability reviewer	22	6.15	1.95	6.90	1.05	0.75	1.87
High ability author & Low ability reviewer	22	7.00	1.45	7.33	1.00	0.32	1.73
High ability author & High ability reviewer	26	6.72	1.84	6.91	0.75	0.19	1.79
Average	94	6.51	1.70	7.04	0.94	0.53	1.77

Grades range from 1 (lowest) to 10 (highest)

Student Ability in Reciprocal Dyads and the Nature of Peer Feedback

In general, reviewer ability was not directly related to the nature of the provided peer feedback ($V = 0.10$, $F(12, 81) = 0.77$, $p = .672$, $\eta_p^2 = .10$). However, visual inspection of Table 1 suggests that higher ability reviewers provide more content-related feedback. Specifically, univariate tests suggested that higher ability reviewers provide more content-related suggestions for improvement ($F(1, 92) = 6.23$, $p = .014$, $\eta_p^2 = .06$) and more content-related explanatory feedback ($F(1, 92) = 4.19$, $p = .043$, $\eta_p^2 = .04$). Given the exploratory nature of our research question, though, the risk for Type I errors needed to be addressed. Hence, a Bonferroni correction was applied, after which these results no longer remained significant.

In general, the nature of the peer feedback also was not related to dyad composition ($V = 0.28$, $F(36, 243) = 0.69$, $p = .908$, $\eta_p^2 = .09$). Only the univariate analysis regarding content-related suggestions for improvement suggested a potential difference between differently composed dyads ($F(3, 90) = 3.44$, $p = .002$, $\eta_p^2 = .10$). On average, 3.00 ($SD = 3.27$) content-related suggestions for improvement were provided within high ability homogeneous dyads. In contrast, such feedback comments appeared to be less common in low ability homogeneous dyads ($M = 1.54$, $SD = 1.50$), heterogeneous dyads with high ability reviewers ($M = 1.82$, $SD = 2.19$), and heterogeneous dyads with low ability reviewers ($M = 1.05$, $SD = 0.99$). As with the relation between student ability and peer feedback nature, however, a Bonferroni correction rendered this univariate effect non-significant.

Thus, at first glance, the nature of peer feedback appears unrelated to either individual reviewer ability and dyad composition. However, a closer look reveals trends suggesting that high ability reviewers may provide more content-related explanations and suggestions for improvement, and that such suggestions for improvement occur more frequently in homogeneous, high ability dyads than in dyads of other compositions.

The Nature of Received Peer Feedback, Author Ability and Essay Performance

Final Essay Performance did not depend on the number of feedback comments that students received on either content-related aspects ($\beta = -0.01$, $t(87) = -0.11$, $p = 0.911$), structure-related aspects ($\beta = -0.12$, $t(87) = -1.03$, $p = 0.304$), or style-related aspects ($\beta = 0.00$, $t(87) = 0.00$, $p = 0.997$) of their draft essay. Moreover, author ability did not significantly interact with feedback comments on these content-related aspects ($\beta = 0.08$, $t(84) = 0.56$, $p = 0.579$), structure-related aspects ($\beta = -0.15$, $t(84) = -1.08$, $p = 0.282$), and style-related aspects ($\beta = -0.08$, $t(84) = 0.63$, $p = 0.529$). See Table 3 for an overview.

Hence, the nature of the peer feedback did not relate to Final Essay Performance, and no significant moderating (interaction) effect of author ability was found, suggesting this is the case for all authors irrespective of their individual ability.

Discussion

The central aim of this study was to assess whether homogeneous or heterogeneous student matching during the peer feedback phase has an effect on the nature of the peer feedback and students' performance on an academic writing task. In the following section, we discuss our findings in the order of the three main research questions.

Student Ability, Dyad Composition and Performance Increase

Research question 1 addressed the direct relation between students' ability in reciprocal dyads and authors' essay performance increase. Authors' essay performance increase neither was directly related to their own ability, nor was it directly related to the ability of their reviewing peer. Most importantly, no relation was found between dyad composition and students' essay performance increase. Based on these data, it apparently does not matter how students are matched on ability during the peer feedback phase of an academic writing assignment.

Table 3. Hierarchical regression analysis (research question 3)

Step	Variables	B	SE	β	t	sig.
1	(Constant)	6.15	0.53		11.54	0.000
	Author Ability	0.03	0.07	0.04	0.41	0.682
	Draft Essay Performance	0.11	0.06	0.19	1.85	0.067
2	(Constant)	6.35	0.56		11.24	0.000
	Author Ability	0.03	0.07	0.04	0.42	0.677
	Draft Essay Performance	0.10	0.06	0.18	1.74	0.085
	FB_content	0.00	0.03	-0.01	-0.11	0.911
	FB_structure	-0.07	0.07	-0.12	-1.03	0.304
	FB_style	0.00	0.02	0.00	0.00	0.997
3	(Constant)	6.23	0.58		10.73	0.000
	Author Ability	0.04	0.07	0.06	0.49	0.624
	Draft Essay Performance	0.11	0.06	0.20	1.86	0.067
	FB_content	-0.01	0.03	-0.05	-0.37	0.713
	FB_structure	-0.07	0.07	-0.13	-1.01	0.315
	FB_style	0.01	0.02	0.04	0.33	0.746
	Author Ability*FB_content	0.01	0.02	0.08	0.56	0.579
	Author Ability*FB_structure	-0.05	0.05	-0.15	-1.08	0.282
	Author Ability*FB_style	-0.01	0.01	-0.08	-0.63	0.529

Step 1: R^2 = .043 (p = .133)   *Dependent variable: Final Essay Performance*
Step 2: R^2 = .059, ΔR^2 = .016 (p = .687)
Step 3: R^2 = .078, ΔR^2 = .019 (p = .631)

These findings contradict prior research that advocates matching students in any particular way, be it homogeneous or heterogeneous matching. Possibly, the anonymous distribution of essays provided a sufficient degree of uncertainty regarding their peer's status to induce a mindful and critical appraisal of the received peer feedback (Gielen et al., 2010). This may suggests that, conditional on students' (perceived) anonymity, *how* students are matched becomes less relevant, emphasizing the role of student perceptions in the peer feedback process (Strijbos et al., 2010).

Student Ability in Reciprocal Dyads and the Nature of Peer Feedback
Research questions 2a and 2b addressed the relation between reviewer ability and the nature of the provided peer feedback, and the relation between dyad composition and provided feedback nature, respectively. In line with prior research (e.g., Snowball & Mostert, 2013), peer feedback primarily focused on issues relating to writing style. In general, however, reviewer ability was not related to the nature of the provided peer feedback. A closer look suggested that high ability reviewers may provide more content-related suggestions for improvement and content-related explanations, but this effect disappeared when a Bonferroni correction was applied to control for false positives (Type I errors). Similarly, dyad composition and provided peer feedback nature appeared unrelated, with a possible exception worth mentioning being homogeneous, high ability dyads: when high ability students were matched with each other, the number of content-related suggestions for improvement appeared to be higher compared to differently composed dyads. Here too, a Bonferroni correction rendered these differences insignificant. However, because of the rather conservative nature of the Bonferroni correction (it may increase the risk for false negatives, Type II errors), we think these trends deserve a closer look in future research

If future research would indicate that these trends are reliable, then they may reflect the possibility that high ability reviewers had a deeper understanding of the assigned theoretical content than the low ability reviewers. If higher ability reviewers indeed are better at diagnosing problems and selecting strategies

for revision (Flower et al., 1986; Patchan et al., 2013), this would explain why they provided somewhat more explanatory feedback and suggestions for improvement on content-related aspects of the texts. These trends could also represent ability differences within a restricted range: in the Dutch higher education context, students typically have completed secondary school at pre-university level, which makes them relatively similar in terms of educational background, age, and probably also in terms of writing ability. In terms of the interpretation of these trends, this similar background is important for at least two reasons. On the one hand, it may simultaneously explain why in the current study only non-significant trends were found and justify to consider these trends informative. After all, if these trends become apparent in a (homogeneous) sample that is fairly similar in terms of students' ability, they may become more salient as samples become more heterogeneous (i.e. open or online educational contexts such as MOOCs; see Huisman, Admiraal, Pilli, van de Ven, & Saab, 2018). On the other hand, it suggests that academic teaching staff may not have to worry too much about the ability matching of higher education students on campus, at least when these students are similar in terms of ability.

The Nature of Received Peer Feedback, Author Ability and Performance Increase

Research question 3 addressed the relation between the nature of the received peer feedback, essay performance, and authors' ability. Authors' essay performance was not related to the nature of the received peer feedback. Specifically, it did not matter whether peer feedback comments focused on content-related, structure-related, or stylistic aspects of authors' drafts. This was the case for all authors, irrespective of their individual ability level.

Whether a student benefits from received (peer) feedback is contingent on his or her mindful reception of, engagement with, and utilization of the feedback (Handley, Price, & Millar, 2011). This study focused directly on authors' summative essay performance, and not on the preceding step of feedback utilization. If we would assume that making revisions based on received peer feedback would generally increase writing performance (although this

assumption is debatable, see Flower et al., 1986), our results appear to contradict those of Patchan et al. (2013). Among others, these authors found that, compared to high ability authors, low ability authors received and implemented more feedback on 'substance' (issues fixable with content knowledge) from high ability reviewers. We did not find such a significant relation between dyad composition, content-related feedback, and content-related essay performance increase. If anything, a contradicting trend was found in which high ability authors received more content-related suggestions for improvement than low ability authors when matched with high ability reviewers. Possibly, the drafts of high ability authors were already perceived somewhat better in terms of structure and style, allowing the high ability reviewers to focus more on content-related aspects.

Implications and Limitations

Some remarks are in place. First, we did not take into account students' perceptions regarding the adequacy of the received peer feedback. As such, it remains an open question how the peer feedback was perceived, and how this is related to student ability matching, peer feedback nature, and essay performance. Future research may focus on these relations by incorporating students' responses on questionnaires (e.g., Strijbos et al., 2010) or interviews (e.g., Hanrahan & Isaacs, 2001). Second, as is the case in many studies on peer feedback, the students in this study were simultaneously feedback provider and receiver. Hence, it was not possible to disentangle what the effects on providing versus receiving peer feedback were on students' essay performance were. Because the act of providing peer feedback may be as effective as receiving peer feedback (Lundstrom & Baker, 2009), investigating these separate effects in relation to student ability seems an interesting topic for future research.

No differences were found in terms of writing performance for homogeneously and heterogeneously matched students. This suggests that ability matching is not related to students' essay performance and that students may very well be matched randomly. Because random student-matching is a feature of many web-based peer feedback applications, this may simplify at least one decision

that academic teaching staff have to make when designing anonymous feedback processes.

Statement on Open Data
The anonymized data and syntaxes are accessible via the following link: https://osf.io/3b48u

3

Peer assessment in MOOCs:
The relationship between peer reviewers' ability
and authors' essay performance

4

4

4

Abstract

In a relatively short period of time, massive open online courses (MOOCs) have become a considerable topic of research and debate, and the number of available MOOCs is rapidly growing. Along with issues of formal recognition and accreditation, this growth in the number of MOOCs being developed increases the relevance of assessment quality. Within the context of a typical xMOOC, the current study focuses on peer assessment of essay assignments. In the literature, two contradicting theoretical arguments can be found: that learners should be matched with same-ability peers (homogeneously) versus that students should be matched with different-ability peers (heterogeneously). Considering these arguments, the relationship between peer reviewers' ability and authors' essay performance is explored. Results indicate that peer reviewers' ability is positively related to authors' essay performance. Moreover, this relationship is only established for intermediate and high ability authors; essay performance of lower ability authors appeared not to be related to the ability of their reviewing peers. Results are discussed in relation to the matching of learners, and instructional design of peer assessment in MOOCs.

Keywords: Peer assessment; Massive Open Online Course; essay performance; ability match

Introduction

Despite their relatively recent introduction, massive open online courses (MOOCs) have become a topic of research in the field of higher education (Raffaghelli, Cucchiara, & Persico, 2015), as well as a topic of scientific and public debate (Kovanović, Joksimović, Gašević, Siemens, & Hatala, 2015). Since the launch of the "Connectivism and Connective Knowledge" MOOC (Downes, 2008), MOOCs became a trend reaching thousands of participants at a time (Evans, Baker, & Dee, 2016). Such large numbers are perhaps not surprising, considering the unrestricted access to university courses for a global audience. The most influential categorization of MOOC pedagogies distinguishes between more connectivist cMOOCs, on the one hand, and more institutionally oriented xMOOCs, on the other hand (e.g., Admiraal, Huisman, & Pilli, 2015; Terras & Ramsay, 2015). Generally speaking, autonomy, interaction, and a construction-oriented teaching approach are central in cMOOCs (Kop, 2011; Toven-Lindsey, Rhoads, & Berdan Lozano, 2015). In contrast, the more institutionally oriented xMOOCs are often characterized by step-by-step learning paths and an emphasis knowledge transfer (Ebben & Murphy, 2014; Rhoads, Sayil Camacho, Toven-Lindsey, & Berdan Lozano, 2015).

As a new form of distance education, MOOCs are in many ways different from traditional university courses. From a research perspective, DeBoer, Ho, Stump, and Breslow (2014) argue that educational variables need to be reconceptualized altogether. For participants, there is usually limited supervision from or direct contact with the teaching staff. Also, assessment procedures are characterized by automated assessment and peer assessment instead of assessment by the teaching staff. Self- and peer assessment - which have been historically used for logistical, pedagogical, metacognitive, and affective benefits - offer promising solutions that can scale the grading of complex assignments in courses with thousands of participants. How to design self- and peer assessment is a challenge in itself as MOOCs have massive, diverse participant enrollment. Within the context of a typical xMOOC, this study focuses on peer assessment of such relatively complex, open-ended assignments, i.e. essay assignments.

Assessment in MOOCs

With the number of available MOOCs rapidly rising, issues of formal recognition and accreditation become increasingly relevant (Lawton & Lunt, 2013). Indeed, several platforms, such as Coursera and EdX, have started to integrate forms of digital 'badges'. This raises important issues such as the reliability of participant identification and the quality of assessment. Regarding the former, several verification methods are being used in a complementary fashion, such as verification via webcams and individual typing-pattern recognition. Such verification methods will undoubtedly continue to develop in the near future. With respect to assessment quality, reliable and valid assessment of participants' learning is required. A practical limitation of having these large numbers of enrolled participants is that alternatives to assessment by teaching staff need to be considered. Not surprisingly, often-occurring forms of assessments in MOOCs are automatic assessment of quizzes and short answer questions, next to self- and peer assessment of more complex, open-ended assignments such as essays. The value of including assessments of participant-generated, open-ended products seems self-explanatory. However, the question which scalable assessment form or process is optimal for such open-ended assignments is not. Different approaches are possible, such as automated essay scoring (AES; e.g., Chauhan, 2014), which come in both supervised and unsupervised variations (Reich, Tingley, Leder-Luis, Roberts, & Stewart, 2015), and human based assessment such as self- and peer assessment. Arguments for the use of peer assessment are twofold. First, peer assessment can be a valid and reliable way to assess student performance (e.g., Cho, Schunn, & Wilson, 2006; Falchikov & Goldfinch, 2000). Second, peer assessment may not only benefit the receiving individual, but may also be beneficial for the peer reviewer him- or herself (Lundstrom & Baker, 2009), since it exposes the peer reviewer to other examples and requires him- or her to actively consider the goals and criteria of the assignment (Flower, Hayes, Carey, Schriver, & Stratman, 1986). In short, both receiving and providing peer assessment can be expected to enhance learning and performance.

Peer Assessment of Essay Assignments in MOOCs

With essay assignments in MOOCs, participants can receive formative feedback from, as well as summative assessment (grading) by multiple peers. The weighted sum of these peer grades usually determines final essay grades, in which self-assessments are occasionally weighted as well. Compared to self-assessments, though, peer assessments might provide a more valid measure of performance. In a recent analysis of three MOOCs, Admiraal, Huisman, and van de Ven (2014) found that self-assessments were biased, and did not explain variance in final exam scores. In contrast, weekly quizzes and peer assessments significantly explained differences in participants' final exam scores. Moreover, research by Cho and colleagues (Cho & MacArthur, 2010; Cho & Schunn, 2007; Cho, Schunn, & Charney, 2006) indicates that assessment by multiple peers can compete with assessment by an expert in terms of reliability (summative), feedback quality (formative), and subsequent improvement by the receiver. Also, in order to get reliable and valid peer feedback and assessments, clear criteria and standards are essential for both authors and reviewers (e.g., Topping, 1998; van Gennip, Segers, & Tillema, 2009), as well as are clear instructions for the provision of feedback (e.g., Gielen & de Wever, 2015). This is an important reason for the inclusion of rubrics in the peer assessment procedure; they explicate the criteria and standards on which the assignment is to be assessed, aiming to simultaneously increase participants' awareness of these criteria and the quality of the provided peer feedback and assessment.

In addition to assessment by multiple peers, and clear standards and criteria, peer assessment might be improved by taking into account the ability of an author and his or her reviewing peers. However, there does not appear to be consensus on how to optimally match authors and reviewers in terms of ability. One the one hand, some authors (e.g., Topping, 2009) argue that learners should be matched with peers of similar ability (homogeneous matching). On the other hand, research by for example Patchan, Hawk, Stevens, and Schunn (2013) suggests that lower ability learners benefit more from assessment by higher ability peers (heterogeneous matching). However, these types of studies on ability matching are generally based on on-campus courses, or at least on

courses in which participants can be expected to be relatively similar in terms of educational background. In open online learning environments such as MOOCS, participants may differ substantially with respect to educational background, ability and motivations. Therefore, a first possible step towards a better understanding of ability matching in open online education is an exploration of how reviewer ability is related to authors' performance, and how author ability and reviewer ability interact in explaining learners' performance. The current study focuses on these questions in the context of a MOOC by Leiden University, launched in 2013.

Research Questions

The central aim of this study is to explore the extent to which peer reviewers' ability is related to authors' essay performance, and to what extent authors' and reviewers' ability interact. Two research questions are formulated. Research question 1 is: "to what extent is peer reviewers' ability related to authors' essay performance?" Research question 2 is: "to what extent does the ability of authors and peer reviewers interact in explaining authors' essay performance?"

Method

The MOOC central to this study is Terrorism and Counterterrorism: Comparing Theory and Practice, organized by Leiden University. It concerns the first run of this particular MOOC, offered in the fall of 2013 via the Coursera platform. The MOOC covered 5 weeks, with an intended workload of 5-8 hours per week.

Participants and Procedure

In total, 26889 participants enrolled for this MOOC. Assessment consisted of five weekly quizzes, two peer reviewed assignments with accompanying self-assessments, and a final exam in the form of a quiz. All five consecutive weeks contained a quiz, and the peer reviewed assignments were scheduled in weeks two and four. The final exam took place in week five (see Table 1 for an overview).

Determination of final grades depended on the track participants choose to follow. In the 'Basic' track, the five quiz scores accumulated to 50%, with the final exam counting as the remaining 50%. In this study, we focus on participants in the 'Advanced' track, which includes the peer reviewed assignments. Here, the quiz scores accumulated to 30%, the two peer reviewed assignments counted for 15% each, and the final exam for the remaining 40%. Within this 'Advanced' track, participants were instructed to review the essays of at least four peers and to perform a self-assessment. A failure to review at least four essays produced by peers and/or submit the self-assessment resulted in a penalty of -20% on the average peer assessment score. This administrative correction is not taken into account in our analysis for two reasons: First, because earlier research showed such self-assessments tend to be biased (Admiraal et al., 2014), and second, because participants' assessment scores then optimally reflect the quality of their submitted work. Self-assessments were done in 94.7% and 95.8% of the cases for assignments 1 and 2 respectively.

Table 1. Chronological overview of assessments (total enrollment = 26889)

Week	Assessment	$N_{(included)}$*	$N_{(total\ submissions)}$
1	Quiz 1	565	5399
2	Quiz 2	564	4077
2	Peer assignment 1	565	842
3	Quiz 3	561	3593
4	Quiz 4	553	3230
4	Peer assignment 2	565	593
5	Quiz 5	544	3014
5	Final exam	540	2988

* = Participants were included when both peer reviewed assignments, *and* at least one of the quizzes was made.

Variables

Quizzes and final exam. The five weekly quizzes were automatically graded, and final scores were based on the best of three possible attempts. Quizzes generally consisted of ten to fifteen multiple choice (MC) questions. For example, one

MC question read "What phrase best explains why terrorism is a contested concept?", with answer alternatives varying from "The enemy of my enemy is my friend" to "One man's terrorist is another man's freedom fighter". Quizzes 1 and 2 slightly deviated from the standard MC question format, both consisting of 9 MC questions plus one open-ended question. These open-ended questions required short answers such as the name of an author, allowing automatic assessment. The final exam consisted of 25 (varying types of) automatically assessed MC questions.

Essay assignments. The two peer reviewed assignments were essay assignments of 600-800 words, excluding references. Each participant was instructed to review at least four peers. A rubric was provided, which allowed for open-ended, freely constructed feedback in addition to every predefined criterion. The predefined criteria of the rubrics slightly differed across the two assignments. Assignment 1 focused on designated terrorist organizations, for which participants were instructed to choose a (in their view) terrorist organization currently not listed as such. The weighted rubric for this assignment included argumentation on chosen examples (40%), argumentation on context (20%), use of sources (30%), and presentation of the essay (10%). Assignment 2 concerned the theoretical assumptions underlying debates on terrorism or counterterrorism, for which participants could choose one of four assumptions to test. The weighted rubric for this assignment included "origin of the claim" (10%), importance of the claim (20%), use of sources (30%), conclusion (30%), and presentation of the essay (10%). Based on these criteria, participants' essay performance was defined as the average score provided by the group of peer reviewers. Within the current study, essay scores were rescaled for interpretation purposes to range between 1 (lowest possible score) and 10 (highest possible score).

Inclusion and Participant Grouping

Participant ability is defined as their average performance on the quizzes made. As such, participants are included in the analysis when they completed both peer reviewed assignments and at least one of the five quizzes. Based on these inclusion criteria, 565 participants are included in this study. In their role as author, participants are grouped post hoc based on ability, defined as average quiz performance (Avg Q1-Q5). Because of the skewed distribution of scores, a visual binning procedure is used to identify three different ability groups: high ($M = 9.94$, $SD = 0.07$, $N = 237$), intermediate ($M = 9.31$, $SD = 0.34$, $N = 257$), and low ($M = 7.67$, $SD = 0.88$, $N = 71$).

Analyses

To answer the two research questions, hierarchical linear regressions are performed with authors' performance on the second peer reviewed essay assignments (PA2) as dependent variable. For the research question 1, authors' performance on the first peer reviewed essay assignment (PA1) is included as an independent variable in step 1 to control for prior essay performance. Average peer reviewer ability (Avg Q1-Q5) is included as an independent variable in step 2. For research question 2, a similar hierarchical regression analysis is performed while differentiating for the three subgroups of author ability (high, intermediate, and low).

Results

In Table 2, the average quiz score and essay scores are presented, both for the total group of authors and for the three ability subgroups. Scores on peer reviewed essay assignments 1 and 2 are significantly correlated, $r(565) = 0.429$, $p < .001$. However, the mean score for essay assignment 2 ($M = 8.24$, $SE = 2.06$) is lower than the one for essay assignment 1 ($M = 8.75$, $SE = 1.60$), $t(564) = 6.13$, $p < .001$. Apparently, the second essay assignment was more difficult than the first. Further, average quiz performance (ability measure; Avg Q1-Q5) correlates significantly with performance on their first essay assignment: $r = 0.301$, $p < .001$. Thus, authors' ability is moderately correlated to their initial essay performance, before the peer assessment phase.

The central aim of this study is to explore the extent to which peer reviewers' ability is related to authors' essay performance, and to what extent authors' and

reviewers' ability interact. Two research questions were formulated, which are addressed one by one below.

Table 2. Assessment descriptives for author ability subgroups

Assessment	Author ability group							
	Lowest (1)		Intermediate (2)		Highest (3)		Total	
	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N	Mean (SD)	N
Avg Q1-Q5	7.67 (0.88)	71	9.31 (0.34)	257	9.94 (0.07)	237	9.37 (0.81)	565
PA1	7.82 (1.87)	71	8.64 (1.66)	257	9.15 (1.28)	237	8.75 (1.60)	565
PA2	7.58 (2.13)	71	8.01 (2.26)	257	8.68 (1.69)	237	8.24 (2.06)	565
Final exam	6.26 (1.93)	67	7.49 (1.73)	244	8.55 (1.13)	232	7.79 (1.71)	543

Peer Reviewers' Ability and Authors' Essay Performance

The ability of peer reviewers appears to be positively related to authors' essay performance (β = 0.13, $t(563)$ = 3.37, p < .001, R^2 = .016), see Table 3. Thus, while correcting for prior essay performance, the ability of peer reviewers is positively related to authors' performance on a subsequent essay assignment. This effect appears to be small, however (Cohen, 1988).

Interaction Between Authors' and Peer Reviewers' Ability

To assess whether the positive influence of peer reviewers' ability on essay performance varies for authors of different ability levels, regression analyses were performed with the three subgroups of authors: relatively low, intermediate, and high ability, as indicated by their average quiz scores. Indeed, there appears to be an interaction between authors' ability and peer reviewers' ability. Specifically, peer reviewers' ability is positively related to the essay performance of the intermediate ability authors (β = 0.11, $t(255)$ = 2.03, p = .044, R^2 = .013) and high ability authors (β = 0.22, $t(235)$ = 3.56, p < .001, R^2 = .046), see Table 4. Here too, however, these effects appear to be small (Cohen, 1988).

Table 3. Regression coefficients for essay performance

Author ability group	Step	Variables included	B	SE	β	t	sig.
Total	1	Constant	3.40	0.44			
		PA1 score	0.55	0.05	0.43	11.27	.000
	2	Constant	-1.68	1.57			
		PA1 score	0.55	0.05	0.43	11.39	.000
		Peer reviewers' ability	0.55	0.16	0.13	3.37	.001
Low	1	Constant	3.94	1.00			
		PA1 score	0.47	0.13	0.41	3.74	.000
	2	Constant	4.81	3.63			
		PA1 score	0.47	0.13	0.41	3.72	.000
		Peer reviewers' ability	-0.09	0.38	-0.03	-0.25	.802
Intermediate	1	Constant	2.80	0.67			
		PA1 score	0.60	0.08	0.44	7.88	.000
	2	Constant	2.40	2.65			
		PA1 score	0.59	0.08	0.43	7.76	.000
		Peer reviewers' ability	0.57	0.28	0.11	2.03	.044
High	1	Constant	4.80	0.75			
		PA1 score	0.42	0.08	0.32	5.21	.000
	2	Constant	-2.92	2.29			
		PA1 score	0.46	0.08	0.35	5.74	.000
		Peer reviewers' ability	0.79	0.22	0.22	3.56	.000

Note:
R^2_{Total} = .184 for step 1, ΔR^2 = .200 for step 2 (p = .001) dependent variable: PA2 score
R^2_{Low} = .169 for step 1, ΔR^2 = .001 for step 2 (p = .802)
$R^2_{Intermediate}$ = .196 for step 1, ΔR^2 = .013 for step 2 (p = .044)
R^2_{High} = .104 for step 1, ΔR^2 = .046 for step 2 (p < .001)

Discussion

In this study, we explored how the average ability of peer reviewers relates to authors' essay performance, and to what extent authors' and peer reviewers' ability interact in explaining differences in essay performance. In general, the ability of the reviewing peers was significantly related to authors' essay performance: the higher the ability of peer reviewers, the more authors' essay performance increased. However, this is not the case for all authors: only the essay performance of the (relatively) intermediate and high ability authors is related to peer reviewers' ability, whereas that of lower ability authors is not. Except for this group of relatively low ability participants, this finding supports the idea of matching of MOOC participants with high ability reviewers during peer reviewed essay assignments.

Different explanations are conceivable, which do not necessarily exclude each other. For example, the very ability to utilize received feedback could have an effect on authors' essay performance. This may imply that participants, perhaps especially those of low ability, may benefit from training or guidance on utilizing feedback. Alternatively, and possibly complementary to the former, these findings may indicate that the quality of the provided feedback could be improved. One possible approach here could be to enhance feedback quality by increasing the awareness of different task aspects such as content, structure, style, and to stimulate the provision of more concrete suggestions for revision (e.g., Nelson & Schunn, 2009; van den Berg, Admiraal, & Pilot, 2006a, 2006b). Another approach could be to provide more structured guidance during the feedback process of peer assessment, for example through detailed feedback templates (Gielen & de Wever, 2015).

Implications and Limitations

Certain limitations regarding the current study need to be addressed, and some cautions are in place when interpreting the results of this study. First, the exact mechanism through which peer reviewers' ability is related to the essay performance of intermediate and high ability authors' remains unclear. It is

possible that peer reviewers' ability is related to the quantity or quality of the feedback, and that higher ability authors are better at utilizing this feedback from high ability peers. Since this study does not assess the quantity or quality of peer feedback comments, or the degree to which revisions are done based on received peer feedback (e.g., Patchan & Schunn, 2015), it remains an open question what the exact role of peer feedback has been. Second, this study focuses on received peer assessments. It is possible that the very act of providing peer assessments contributes to participants' learning too (cf. Lundstrom & Baker, 2009), and that providing peer assessment is particularly beneficial for higher ability participants because they tend to more actively consider the assignment goals and criteria (e.g., Flower et al., 1986; Patchan et al., 2013). For future research in online and on-campus education, research on the relation between author and reviewer ability, feedback quality and essay performance seems a fruitful endeavor.

Finally, this study aimed to provide a first exploratory step towards a better understanding of ability matching in open online education. With such first steps however, the degree to which results can be generalized is limited. For one, the available information on the MOOC participants in this study is limited; we have no information with respect to participants' national or educational background, age, or professional occupation. In addition, and potentially related to these variables, it remains unknown whether participants' preference for particular topics, learning activities (i.e. peer assessment), and assignment types (i.e. argumentative texts) influences how they perform peer assessments. In the current study, participants were grouped randomly and not based on such variables. As such, they could be presumed to be relatively evenly distributed over the different ability groups, making them unlikely to confound with the variables used in the analyses of this study. Either way, with respect to future MOOC design and MOOC research, more information on participants could prove valuable. Especially if MOOC platforms would facilitate (quasi-)experimental interventions within MOOC iterations (e.g., A/B testing) or between cohorts of participants, variables such as participants' national or educational background could be interesting matching criteria. This

information on participants should ideally be available a priori, for example through pre-course surveys, in order to purposefully match participants during the peer assessment phase in a MOOC. Another limitation of this explorative study is that only one MOOC was studied, and that the topic of terrorism may be sensitive to participant characteristics such as national background. Hence, future research on peer matching should include MOOCs with different course designs, on different topics and from different platforms, in order to validate the current findings.

Despite these limitations, this empirical study contributes to our knowledge regarding peer assessment in MOOCs. The study provides a first insight into the relationship between the ability of authors and peer reviewers in peer assessment with essay assignments, and gives directions for future research on online peer assessment practices. We believe these findings to be informative for educational developers involved in the instructional design of MOOCs, and hope to instigate future research on peer matching in both open online and on-campus education.

Statement on Open Data

The anonymized data and syntaxes are accessible via the following link: https://osf.io/fv4mw

Students' peer feedback role, peer feedback perceptions and essay performance

5

Abstract

Within the higher education context, peer feedback is frequently applied as an instructional method. Research on the learning mechanisms involved in the peer feedback process has covered aspects of both providing and receiving feedback. However, a direct comparison of the impact that providing and receiving peer feedback has on students' writing performance is still lacking. The current study compared the writing performance of undergraduate students (N=83) that either provided or received anonymous written peer feedback in the context of an authentic academic writing task. In addition, we investigated whether students' peer feedback perceptions were related to the nature of the peer feedback they received and to writing performance. Results showed that both providing and receiving feedback led to similar improvements of writing performance. The presence of explanatory comments positively related both to how adequate students perceived the peer feedback to be, as well as to students' willingness to improve based upon it. However, no direct relation was found between these peer feedback perceptions and students' writing performance increase.

Keywords: peer feedback; academic writing; perceptions; performance

The reader as evaluator imposes additional goals or criteria on the text […]. In a sense then, the process of evaluation simply turns up the power on the reading process: It enlarges the set of constraints that the mental representation one is building must meet and turns reading into testing. (Flower et al. 1986, p. 23)

Introduction

Peer feedback is frequently applied within the higher education context. As an instructional method, it can be beneficial to students' learning of domain-specific skills (van Zundert, Sluijsmans, & van Merriënboer, 2010). With respect to the learning mechanisms involved in the peer feedback process, some prior studies have differentiated between providing and receiving peer feedback on academic writing (e.g., Cho & MacArthur 2011; Nicol, Thomson, & Breslin, 2014; Greenberg 2015; McConlogue, 2015). To our knowledge, however, a direct experimental or quasi-experimental comparison of the impact that providing versus receiving peer feedback has on students' learning gains is lacking. As a consequence, it remains an open the question how these compare in terms of their relative impact on students' writing performance.

The current study has two central aims. First, it compares the effects of providing versus receiving peer feedback on students' performance in the context of an authentic academic writing assignment. Second, to gain more insight into the peer feedback process, it investigates the relations between the nature of the received peer feedback, students' perceptions thereof, and their subsequent writing performance.

Providing versus Receiving Peer Feedback

Providing peer feedback is considered beneficial to students' writing as it stimulates them to actively consider the task-specific processes and criteria. According to Flower et al. (1986), three specific processes come into play when a student reviews a text. First, there is problem detection. Second, there is problem diagnosis, which helps to improve writing when potential revision strategies are not obvious, i.e. do not involve relatively straightforward

corrections or rewriting. Third, strategies for revision concern actions that follow problem detection and diagnosis. The act of providing peer feedback triggers students to engage in problem detection, and can stimulate them to engage in problem diagnosis and subsequently contemplate solutions and suggest revisions. As a result, students who provide peer feedback gain experience in problem detection, may become more aware of (types of) writing problems, and may discover different revision strategies (Patchan & Schunn, 2015). These feedback processes include students taking different perspectives, comparing others' work to their own and the assimilation of new knowledge, which can be coherently referred to as reflective knowledge building (e.g., Tsui & Ng, 2000; van Popta et al., 2017).

Two quantitative empirical studies have provided support for such learning-by-reviewing with academic writing (Cho & MacArthur, 2011; Greenberg, 2015). Cho and MacArthur (2011) found that students who reviewed three example papers performed better on a subsequent writing task compared to students that only read these same example papers and compared to controls reading papers on an entirely different subject. Greenberg (2015) also found that students improved their research reports after providing peer feedback, and this improvement was evident across both simple and more complex sections of their reports. Yet, neither of these studies directly compared the impact of providing versus receiving peer feedback on students' final writing performance. To our knowledge, such a comparison has only been reported by Lundstrom and Baker (2009). They found that lower proficiency 'givers' outperformed lower proficiency students in a 'receiver' condition. In this particular study, however, students' experience of providing versus receiving (utilizing) peer feedback was restricted to a controlled *training* intervention, without them actually providing or receiving peer feedback on each other's writing.

In summary, none of these studies directly compared the impact of providing versus receiving peer feedback in the context of an authentic writing task. As authentic writing tasks concern self-generated texts and may weigh into students' grade within a course, students may be inclined and incentivized to provide peer feedback and respond to received feedback more seriously (McDowell,

2012). Qualitative inquiries in authentic contexts indicate that students can perceive the benefits of providing peer feedback (Chen, 2010), and that they may even consider this more beneficial to their learning than receiving feedback from peers (Ludemann & McMakin, 2014; Nicol, Thomson, & Breslin, 2014; McConlogue, 2015). The current study's first central aim is to compare the impact that providing versus receiving peer feedback has on students' academic writing performance.

Research question 1:

To what extent do students that provide peer feedback improve their writing compared to students that receive peer feedback?

Following the aforementioned theoretical rationale and empirical findings, students providing peer feedback are expected to improve their writing at least as much as students receiving peer feedback. If this expectation is confirmed, this would support the learning-by-reviewing rationale. In contrast, if students receiving peer feedback outperform those providing it, this would indicate that the learning mechanisms involved in the act of providing peer feedback are not as strong as the learning mechanisms involved in receiving and utilizing peer feedback (e.g., receiving information on the gap between current performance and goal performance; Nicol & Macfarlane-Dick, 2006; Hattie & Timperley, 2007).

Student Perceptions of Received Peer Feedback

The second central aim of the current study is to investigate the relation between the nature of the received peer feedback and students' perceptions thereof, and the relation between these perceptions and subsequent writing performance. The following section first covers the relation between the nature of peer feedback and students' perceptions, culminating in research questions two and three. Thereafter, we address the relation between students' peer feedback perceptions and their subsequent writing performance, leading up to research question four.

The nature of the peer feedback message. The current study focused on task-level peer feedback, adopting the operationalization proposed by van den Berg, Admiraal and Pilot (2006c). This operationalization differentiates between the aspects of the text on which the feedback focuses (including *content*, *structure* and *style*) and the functions of the feedback (including *analysis*, *evaluation*, *explanation* and *revision*). There were three reasons for adopting this operationalization. First, the four feedback functions by van den Berg, Admiraal and Pilot (2006c) are largely consistent with the different feedback functions and components described in prior review studies. For example, evaluations, explanations and suggestions for revision mirror a conceptual resemblance with 'correcting' and 'guiding' (Narciss, 2008) and with the question how a student is doing in relation to the standard and how to proceed towards that goal (Hattie & Timperley, 2007). Second, we considered the inclusion of van den Berg, Admiraal and Pilot's (2006c) feedback aspects content, structure and style as a valuable addition to the feedback functions, as we expected these feedback aspects to be relatively salient to students. For example, we expected that students will differentiate between the value of relatively superficial peer feedback on writing style or grammar versus more content- or structure related peer feedback. Third, the operationalization of feedback aspects closely aligned with the criteria of the essay assignment that was the subject in this study.

Student perceptions of peer feedback aspects and functions. A relation between the nature of the peer feedback and subsequent writing performance is likely to be mediated by students' perceptions of the received peer feedback (Strijbos, Narciss, & Dünnebier, 2010). However, empirical inquiries into students' perceptions tend to focus on students' general experience of the peer feedback process (e.g., Mostert & Snowball, 2013). This study contributes to the existing literature by investigating the relations between the nature of the received peer feedback, students' perceptions thereof, and subsequent writing performance. To this end, we used the feedback-perception questionnaire developed by Strijbos, Narciss and Dünnebier (2010). Among others, this questionnaire measures students' perceptions regarding the adequacy of the received peer feedback and their willingness to improve based upon it. In particular, we wish to assess the extent to which peer feedback on particular aspects of the text (content, structure or style) and with particular functions (analytical, evaluative, explanatory or suggesting revisions) relates to students' perceptions of adequacy and their willingness to improve.

Regarding the peer feedback aspects, comments on content and structure are more likely to go beyond straightforward corrections or rewriting than comments on style and, therefore, are expected to stimulate more substantial revisions. Prior research indicates that complex revisions predict subsequent writing quality (Cho & MacArthur, 2010). If students can recognize the different peer feedback aspects and, at least to some extent, differentially value the potential contributions of these aspects in making substantial revisions, then it seems plausible to expect that peer feedback on content and structure will be perceived as more adequate than peer feedback on style.

Regarding the peer feedback functions, these - implicitly or explicitly- indicate discrepancies between students' current performance and the performance goal of the task (analysis, evaluation), provide suggestions on how to advance towards that goal (revision), and provide explanatory information on either the gap between current and goal performance or the suggested revision (explanation) (Hattie & Timperley, 2007; Lizzio & Wilson, 2008). As a result, we expect these peer feedback functions to positively relate to students' perceptions of adequacy and their willingness to improve.

Research question 2:

To what extent do students perceive peer feedback on aspects of content and structure as adequate compared to peer feedback on aspects of style?

Research question 3:

To what extent are perceived peer feedback adequacy and students' willingness to improve related to the degree in which the peer feedback is analytical, evaluative, explanatory, or suggesting revisions?

Peer Feedback Perceptions and Writing Performance

Students' perceptions may mediate between the received peer feedback and subsequent performance (e.g., Van der Pol et al., 2008; Strijbos, Narciss & Dünnebier, 2010). It clearly is important to understand how such peer feedback perceptions relate to students' subsequent writing performance in authentic learning contexts. It is to be expected that students' perceptions of adequacy and their willingness to improve based upon the received peer feedback positively relate to their subsequent writing performance. However, empirical evidence for such perceptions-performance relations is mixed. Van der Pol et al. (2008) found that students were more inclined to use peer feedback for revising their work when they regarded the peer feedback as important. In contrast, Strijbos, Narciss, and Dünnebier (2010) did not find a relation between students' peer feedback perceptions and revision efficiency (including error detection, error diagnosis and correctly suggested revisions) in a controlled experimental setting. In the context of a more authentic online peer assessment task, Kaufman and Schunn (2011) also found no relation between student perceptions and the frequency of revisions made. Focusing on students' writing *performance* instead of revision, the current study investigates the relation between peer feedback perceptions and writing performance within an authentic academic writing assignment.

Research question 4:

For students receiving peer feedback, to what extent do perceived adequacy and willingness to improve relate to their subsequent writing performance increase?

A positive relation between on the one hand perceived peer feedback adequacy and/or students' willingness to improve and, on the other hand students' subsequent writing performance would support the findings by van der Pol et al. (2008). Moreover, if peer feedback in relation to certain aspects of the text or serving a particular function relates to these peer feedback perceptions (research questions two and three), this would shed light on *how* the nature of peer feedback influences students' writing performance. In contrast, if students' peer feedback perceptions do not relate to their subsequent writing

performance, that would be in line with prior studies by Strijbos, Narciss, and Dünnebier (2010) and Kaufman and Schunn (2011). This would suggest alternative pathways through which the reception of peer feedback may influence subsequent writing performance, such as through inducing reflection (cf. Kaufman & Schunn, 2011).
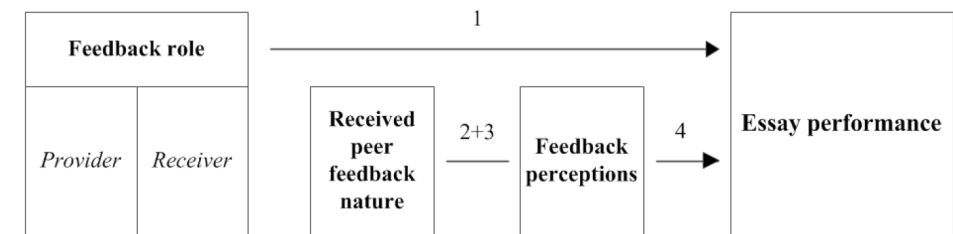
Figure 1. Graphical representation of research questions

Method

Participants and Procedure

Participants were students of a research-intensive university in The Netherlands that partook in an introductory course on education and child development studies. Of the 136 students majoring in Education and Child Studies, 91 students fully participated by providing informed consent and filling in both pre-test and post-test questionnaires. Out of these 91 students, data for 8 students were removed after the peer feedback phase because they did not adhere to their assigned role of either provider or receiver, resulting in a final sample of 83 students. Their mean age was 19.46 years (*SD* = 1.83), with 77 students (92.8%) being female, which was not uncommon for this and prior cohorts. In eight weekly lectures, the course covered topics from two different fields: family pedagogy and educational sciences. Between weeks 3 and 6 students were required to write and submit a draft essay on one of these two topics. The peer feedback phase took place in week 7, after which students were provided the opportunity to revise their drafts and submit a final version of their essay during the 8th and final week.

Experimental Manipulation

Within the virtual learning environment (Turnitin) and within each of the two essay topics, the researcher matched students with a similar-ability peer based on their performance on a comparable essay assignment from a preceding course. Students were assigned the role of either feedback provider or receiver. The online peer feedback was provided and received anonymously to control for the potential effects of (perceived) status differences between students (e.g. Strijbos, Narciss, & Dünnebier, 2010). Among the included students, 46 students were assigned the role of feedback provider, whereas 37 were assigned the role of feedback receiver. For ethical considerations, these roles were reversed in a subsequent course.

Peer Feedback Guidelines, Assignment Criteria and Grading

In the first week of the course students were informed on the course structure, essay assignment and peer feedback process. It was mandatory for students to submit a draft essay, to provide serious peer feedback, and to submit a final version of their essay within the pre-set deadlines. The essay was to be about one of two preassigned topics, within the fields family pedagogy or educational sciences, with two scientific articles being provided for each topic. Students had at least one prior experience with peer feedback through the same virtual learning environment. Verbal instructions during the meeting in week 1 therefore focused on how to provide constructive feedback and on the assignment criteria. These instructions were made available online.

Final essays were graded by the teaching staff on a scale of 1 (lowest) to 10 (highest), with overall grades being the weighted sum of the following criteria: content (30%), structure (20%), writing style (20%), referencing (20%), and presentation and spelling (10%). Based on the same criteria, an experienced research assistant graded the draft essays, the grades for which were not communicated to students. For the purpose of this study, the elements of writing style, referencing, and presentation were aggregated into a single style variable (weighing 50% into calculating final grades). Both the teaching staff and research assistant were unaware of students' assigned feedback role. To ascertain the comparability of grades as indicators of essay quality, inter-rater agreement

between the grades of the research assistant and the teaching staff was calculated based on nine (> 10%) random draft essays. Inter-rater agreement was high ($r(9) = .84$, $p = .005$) and absolute grades were similar ($t(8) = 0.57$, $p = .584$).

Measures and Instruments

Peer feedback perceptions. Students' peer feedback perceptions were measured post-test, that is, directly after the deadline for the revised final essays. An adapted and translated version (Agricola et al., 2016) of the feedback perception questionnaire by Strijbos, Narciss, and Dünnebier (2010) was used, with a Likert-scale ranging from 1 ('totally disagree') to 5 ('totally agree'). The subscales for perceived adequacy (9 items, $\alpha = .92$, $N = 37$) and willingness to improve (3 items, $\alpha = .87$, $N = 37$) both proved reliable.

The nature of peer feedback. The nature of peer feedback was operationalized based on van den Berg, Admiraal, and Pilot (2006c), distinguishing between aspects and functions. Regarding the peer feedback *aspects*, content referred to elements such as the argumentation within the text or the clarity or use of information and concepts. The structure aspect referred to issues concerning the internal consistency of the essay (e.g., between the problem statement and the discussion), and the style aspect referred to issues including grammar, language use, and referencing. Regarding the feedback *functions*, peer feedback segments were coded as analysis when they concerned the reader's understanding of the text. Analytical comments were often phrased as questions such as 'What did you mean by […]?'. Peer feedback segments were coded as evaluation when they reflected quality statement, including comments such as 'Well-structured paragraph' or 'This sentence is very hard to read'. Segments were coded as revision when they either directly or indirectly suggested revisions. These could include comments such as 'Rephrase your main question to incorporate […]' or 'See the APA manual for correct in-text referencing'. Finally, feedback segments were coded as explanation when they provided arguments supporting either evaluative comments or suggestions for revision. For example, explanatory peer feedback could follow up suggestions for revisions such as 'Rephrase your main question to incorporate […]' with '*because right now it does not align with your*

conclusion', or follow up evaluative comments such as 'This sentence is very hard to read' with '*it is too long and there are multiple spelling and grammar issues*'.

Coding procedure. Following the two-step procedure by Huisman et al. (2017), feedback segments were first coded as an aspect of content, structure or style, after which each aspect-coded segment was assigned one or more feedback functions. Hence, a feedback segment was attributed only one feedback aspect, which could include multiple functions. Feedback segments were independently coded by the first author and a trained research assistant, with initial agreement indices for the separate peer feedback aspects and functions ranging between $k =$ .73 and $k = .87$ (see Table 1). Disagreements were resolved through consultation between the coders.

Table 1. Inter-rater agreement for peer feedback coding

	Code	Inter-rater agreement		95% CI
		Kappa	SE_{kappa}	
Feedback aspect	Content	.80	.026	$.75 \leq .85$
	Structure	.77	.047	$.68 \leq .86$
	Style	.87	.019	$.83 \leq .91$
Feedback function	Analysis	.73	.048	$.64 \leq .82$
	Evaluation	.76	.027	$.70 \leq .81$
	Explanation	.75	.027	$.70 \leq .80$
	Revision	.87	.021	$.83 \leq .92$

Note: N=711 feedback segments (multiple feedback functions per feedback aspect possible)

Analyses

Research question 1: Peer feedback role and writing performance
To compare the impact of providing versus receiving peer feedback on students' overall writing performance increase (final grade minus draft grade), an independent t-test was conducted to compare overall performance increase between feedback providers and feedback receivers (feedback role). Subsequently, a multivariate analysis of variance (MANOVA) was conducted to more specifically investigate the relation between students' feedback role

(independent variable) and students' performance increase on the assignment criteria content, structure, and style (dependent variables). All standardized mean differences and standardized gains reported in this study were corrected for sample size (Hedges' *g*, see Borenstein et al., 2009).

Research question 2: Student perceptions in relation to peer feedback aspects
To assess the extent to which content- or structure-related peer feedback contributed to students' perceptions of adequacy in comparison to peer feedback on style, a multiple regression analysis was conducted. The independent variables were feedback comments on content, structure and style, with the dependent variable being perceived adequacy. Using the statistical software R (v3.4.1, R Core team 2017), the 'relaimpo' package (Groemping, 2006) was applied to compare the relative contribution of the independent variables. This procedure compares two independent variables with respect to the proportions (percentages) of the total explained variance (R^2) that each account for in explaining the dependent variable. In particular, it assesses the differences between these relative contributions and provides bootstrapped confidence intervals (here set at $\alpha = 0.05$) to test whether this interval includes zero. For the current research question, two specific comparisons were made: one comparing the relative contributions of content and style in predicting perceived adequacy and one comparing the relative contributions of structure and style in predicting perceived adequacy.

Research question 3: Student perceptions in relation to peer feedback functions
Research question three assessed the extent to which analytical, evaluative or explanatory peer feedback comments or peers' suggestions for revisions were related to two components of students' peer feedback perceptions: perceived adequacy and willing to improve. Two separate multiple regression analyses were conducted to facilitate an exploration into the relative contribution of the independent variables (analogous to research question 2). Independent variables were the received feedback functions analysis, evaluation, explanation and revision. Dependent variables were either perceived adequacy or willingness to improve.

Research question 4: Peer feedback perceptions and students' writing performance

To explore the relation between students' perceptions of the peer feedback they received and their subsequent increase in writing performance, a multiple regression analysis was conducted. Independent variables were students' willingness to improve and the perceived adequacy of the received peer feedback, the dependent variable was students' performance increase.

Results

Feedback Role and Writing Performance

Overall writing performance between drafts ($M = 6.56$, $SD = 1.38$) and final essays ($M = 6.99$, $SD = 0.99$) improved significantly ($t(82) = 2.62$, $p = .010$, $g = 0.35$; see Table 2). Confirming our expectations, students providing peer feedback ($N = 46$, $M = 0.45$, $SD = 1.43$) improved their essays to a similar degree as students receiving peer feedback ($N = 37$, $M = 0.42$, $SD = 1.61$; $t(81) = 0.09$, $p = .928$, $g = 0.02$). In addition, the performance increase for feedback providers and –receivers was similar on the separate assignment criteria content, structure and style ($V = .01$, $F(3, 79) = 0.18$, $p = .912$, $\eta^2_p = .01$). In summary, students generally improved from draft to final essay, and peer feedback providers and –receivers similarly improved their writing performance after the feedback phase across all aspects of the assignment.

Student Perceptions in Relation to Peer Feedback Aspects

Peer feedback was generally provided on aspects of style (62.4%) or content (26.9%), whereas peer feedback on essay structure (8.9%) did not occur frequently (see Table 3). Taken together, peer feedback on aspects of content, structure and style explained 21.6% of the total variance in students' perceived peer feedback adequacy ($F(3,33) = 3.04$, $p = .043$). Separately however, peer feedback on content ($\beta = 0.29$, $p = .100$, $R^2_{content} = 0.11$), structure ($\beta = 0.19$,

$p = .277$, $R^2_{structure} = 0.07$), or style ($\beta = 0.12$, $p = .471$, $R^2_{style} = 0.04$) did not significantly predict the extent to which students perceived the peer feedback as adequate. Peer feedback on content and style (relative contribution difference = 0.04, [-0.15, 0.27]) and peer feedback on structure and style (relative contribution difference = 0.03, [-0.14, 0.20]) contributed similarly in explaining students' perceptions of adequacy (see Figure 2). In summary, all peer feedback comments on content, structure and style combined significantly explained 21.6% of the variance in students' peer feedback perceptions, and their relative contribution to perceived adequacy was similar.

Table 2. Draft and final essay performance by feedback role.

Feedback role	Assignment criterion	Draft essay		Final essay		Performance increase	
		Mean	SD	Mean	SD	Mean	SD
Provider (N=46)	Total weighted grade	6.55	1.27	7.00	0.85	0.45	1.43
	- *Content*	6.22	1.59	6.27	1.26	0.05	1.85
	- *Structure*	6.20	1.75	6.68	1.39	0.49	1.90
	- *Style*	6.89	1.36	7.56	0.94	0.67	1.53
Receiver (N=37)	Total weighted grade	6.56	1.52	6.98	1.16	0.42	1.61
	- *Content*	6.46	1.52	6.31	1.53	-0.15	2.04
	- *Structure*	6.27	1.81	6.82	1.67	0.55	2.40
	- *Style*	6.75	1.72	7.45	1.32	0.70	1.62
Total group (N=83)	Total weighted grade	6.56	1.38	6.99	0.99	0.43	1.51
	- *Content*	6.33	1.55	6.29	1.38	-0.04	1.93
	- *Structure*	6.23	1.76	6.75	1.51	0.52	2.03
	- *Style*	6.83	1.52	7.51	1.12	0.68	1.56

Note: Grades range from 1 (lowest) to 10 (highest)

Table 3. Received peer feedback aspects and functions

		Function				
		Analysis	*Evaluation*	*Explanation*	*Revision*	***Total***
Aspect	*Content*	46	88	59	81	274
	Structure	1	35	27	27	90
	Style	0	72	160	421	653
	Total	47	195	246	529	1017

Note: N=37 receivers; 1017 segments (97.60%) coded as an Aspect with ≥ 1Function(s); 25 segments (2.40%) coded as N/A
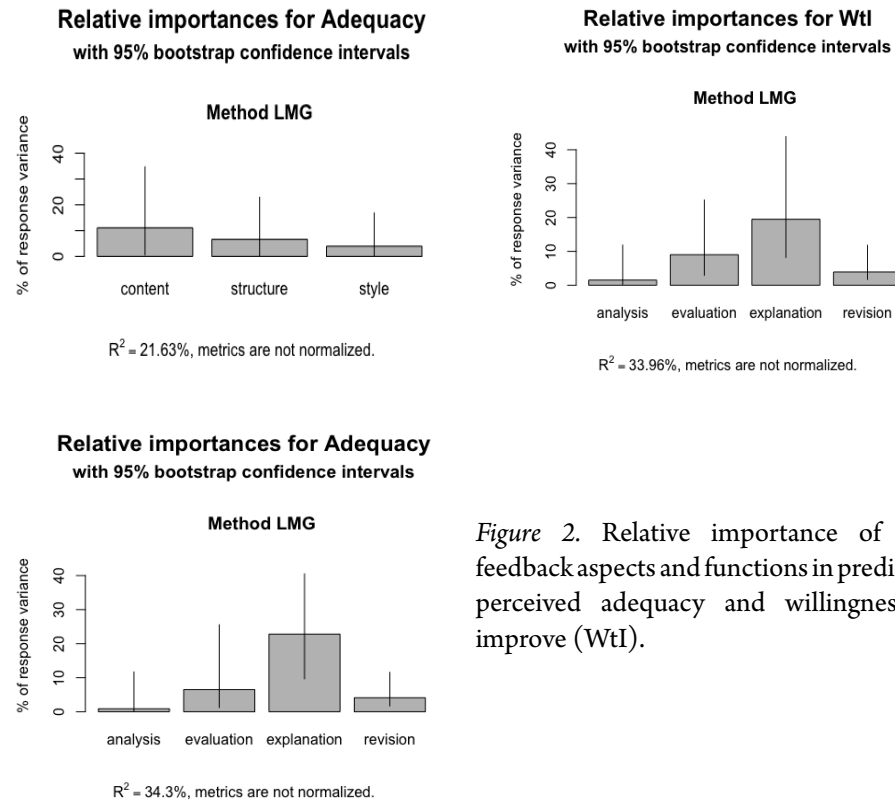
Figure 2. Relative importance of peer feedback aspects and functions in predicting perceived adequacy and willingness to improve (WtI).

Student Perceptions in Relation to Peer Feedback Functions

Overall, peer feedback functions predominantly concerned (suggestions for) revisions (52.0%), evaluations (19.2%) or explanations thereof (24.2%). Taken together, these four peer feedback functions explained 34.3% of the variance in students' perceived adequacy ($F(4,32) = 4.18$, $p = .008$) and 34.0% of the variance in students' willingness to improve ($F(4,32) = 4.11$, $p = .008$). Regarding the extent to which students perceived the peer feedback as adequate (research question 3), we only found a significant positive relation with explanatory peer feedback ($\beta = 0.69$, $p = .004$, $\eta_p^2 = .45$). This relative contribution of explanatory peer feedback was significantly higher than that of analytical peer feedback (relative contribution difference = 0.22, [0.03, 0.38]) and that of peers' suggestions for revisions (relative contribution difference = 0.19, [0.04, 0.35]) in explaining perceived peer feedback adequacy. Regarding

the extent to which students were willing to improve their writing based on the received peer feedback, we again found explanatory peer feedback to be the only significant predictor ($\beta = 0.57$, $p = .016$, $\eta_p^2 = .37$). Moreover, the relative contribution of explanatory peer feedback was again significantly higher than that of analytical peer feedback (relative contribution difference = 0.18, [0.02, 0.41]) and that of peers' suggestions for improvement (relative contribution difference = 0.16, [0.05, 0.38]). In predicting both perceived adequacy and willingness to improve, the relative contributions of evaluative peer feedback did not differ from any of the other three feedback functions.

In summary, peer feedback in the form of analytical, evaluative and explanatory comments and suggestions for revision taken together explained over a third of the variance in both students' perceived peer feedback adequacy and their willingness to improve. However, a closer look revealed that only explanatory peer feedback comments significantly predict these peer feedback perceptions. In predicting these perceptions of adequacy and willingness to improve, explanatory peer feedback was more important than analytical peer feedback comments and suggestions for revisions[3].

Peer Feedback Perceptions and Students' Writing Performance

We investigated the extent to which students' perceived peer feedback adequacy and their willingness to improve related to their writing performance increase (research question 4). Neither perceived peer feedback adequacy ($\beta = .17$, $p = .617$, $\eta_p^2 = .085$) nor students' willingness to improve based upon the received peer feedback ($\beta = -.45$, $p = .088$, $\eta_p^2 = -.295$) significantly related to students' subsequent increase in writing performance.

Discussion

The current study had two central aims: (1) to compare the impact of providing versus receiving peer feedback on students' performance on an authentic academic writing assignment and (2) to explore the relations between the

nature of the received peer feedback, students' perceptions thereof and their subsequent writing performance. In the following sections, we discuss the main outcomes with regards to these central aims and the corresponding research questions.

Feedback Role and Writing Performance

Students in this study either provided or received anonymous written peer feedback in the context of an authentic academic writing assignment. As expected, feedback providers and receivers were found to improve to a similar extent from draft to final essay, both in terms of their overall grades as in terms of the separate assignment criteria relating to content, structure, and style. These results suggest that the learning mechanisms involved in the act of providing peer feedback (e.g., triggering problem detection, stimulating problem diagnosis and revision strategies; Flower et al., 1986; Patchan & Schunn, 2015) and those involved in receiving peer feedback (e.g., receiving information on the discrepancies between current performance, goal performance, and how to close this gap; Nicol & Macfarlane-Dick, 2006; Hattie & Timperley, 2007) have a similar impact on students' subsequent writing performance. This finding corroborates those of prior studies reporting the positive effects providing peer feedback on students' own writing performance (e.g., Cho & MacArthur, 2011; Greenberg, 2015). Also, this study's findings corroborate the survey data of Nicol, Thomson and Breslin (2014), who reported on students' *perceptions* of the benefits of providing versus receiving peer feedback. Specifically, similar proportions of students reported to have modified their initial writing assignment as a result of providing peer feedback versus receiving it. In addition to exploring the relation between the specific nature of the peer feedback and students' perceptions thereof, the current study builds on Nicol, Thomson and Breslin's (2014) findings is by investigating how students' feedback role impacts their actual writing *performance*.

We know of only one prior study that related students' feedback role to their writing performance (Lundstrom & Baker, 2009). This study reported higher writing performance for students that were trained in providing peer feedback,

particularly for relatively low performing students. Contextual differences may explain the different findings of this study and ours. In particular, Lundstrom and Baker's (2009) peer feedback intervention was restricted to a training phase, and as such related less directly to the measured writing assignments, potentially limiting students' engagement. The current study was conducted within an authentic writing assignment and incorporated both a carrot (grade) and a stick (mandatory participation) to incentivize students' engagement. Given these differences, it may not be surprising that the feedback providers in the Lundstrom and Baker study outperformed the feedback receivers; the 'providing' training may have sufficiently activated learning mechanisms such as problem detection, -diagnosis and the contemplation of strategies for revision (Flower et al., 1986; Patchan & Schunn, 2015), whereas the 'receiving' training may not have been perceived as sufficiently relevant to the students' own writing assignment and performance. In the current study, students' task-engagement and these learning mechanisms are stimulated to more similar degrees for peer feedback providers and receivers.

Research on the training of students *before* the peer feedback phase and research conducting interventions during the peer feedback process both have their merits, and may even be complimentary. In both cases, however, the authenticity of the learning context may be crucial in determining the practical value of the research findings. In order to more confidently pin down the effects that providing and receiving peer feedback have on students' own subsequent performance, we would therefore like to make a case for further empirical research in authentic writing contexts, ideally including a control- or comparison-group.

Student Perceptions and the Nature of Peer Feedback Comments

The current study adopted van den Berg, Admiraal and Pilot's (2006c) operationalization of the nature of peer feedback, which distinguishes between the aspects of the text to which the peer feedback refers (content, structure, style) and the function of the peer feedback (analysis, evaluation, explanation, suggestion for revision). We expected that students would perceive peer feedback

on essay content and structure as more adequate compared to peer feedback on style. Contrary to our expectation, students perceived peer feedback on aspects of content, structure and style as equally adequate. An explanation could be that aspects of style weighted for fifty percent into calculating students' final grades. Students may have perceived peer feedback on style as relatively important as a result of the 'backwash effect' (Biggs, 1996), meaning that the assessment criteria could have driven students' perceptions of what is adequate peer feedback and performance. Hence, future research applying differently weighted assignment criteria may clarify to what extent perceptions of peer feedback adequacy are driven by such characteristics of the task.

Given that the feedback functions can provide information on the discrepancies between current performance, goal performance, and how to close this gap (e.g. Hattie & Timperley, 2007), we expected that each feedback function could contribute to students' perceptions of peer feedback adequacy and their willingness to improve based upon it. However, only explanatory peer feedback positively related to these peer feedback perceptions. In particular, explanatory peer feedback comments were relatively important in comparison to analytical peer feedback and peer feedback containing suggestions for revision. These findings are largely aligned with those of Strijbos, Narciss, and Dünnebier (2010) who found that elaborate and specific peer feedback can be perceived as more adequate in comparison to concise and general peer feedback, and with Bolzer, Strijbos and Fischer (2015) whose findings suggest that justifications influence mindful cognitive processing. They also resonate with the findings of Lizzio and Wilson (2008), who reported a relation between explanatory ('justifying') peer feedback comments and perceptions of fairness.

Knowing what students perceive as adequate peer feedback and what drives these perceptions can be applied in the training of students in the peer feedback process (see for example Sluijsmans 2002; Gielen, Dochy, & Onghena, 2011). Contingent on the extent to which the assessment criteria may have driven students' perceptions of peer feedback adequacy, the current findings indicate that the role of explanations should be emphasized when training students for the peer feedback process.

Peer Feedback Perceptions and Students' Writing Performance

In addition to investigating the effects of providing versus receiving peer feedback, this study explored the extent to which students' perceptions of the received peer feedback related to an increase in their subsequent writing performance. We found that students' perceived adequacy of the peer feedback and their willingness to improve based upon it were unrelated to their writing performance increase. Apparently, these perceptions do not mediate between the nature of the peer feedback and subsequent writing performance.

Considering the mixed findings in prior research, the current study's results do not provide a conclusive answer regarding the role of students' peer feedback perceptions in relation to their performance. On the one hand, Gielen et al. (2010) found that the presence of explanations ('justifications') in peer feedback could raise subsequent performance of assessees in secondary education. In that light, and as perceptions are likely to influence information processing (Pajares, 1992), it may be considered somewhat surprising that students in the current study did perceive peer feedback as adequate when it included such explanatory comments, but that these perceptions – in turn – did not relate to their subsequent writing performance. On the other hand, the current study's findings are in line with prior research by Schunn and colleagues (e.g., Cho, Schunn, & Wilson, 2006). Among others, these studies indicate that students' peer feedback perceptions are unrelated to revision behavior, and that what students perceive to be helpful peer feedback may not always be linked to subsequent writing performance. Possibly, the peer feedback process may have induced students' reflection about their writing, which may act independent from how they perceive their fellow students' peer feedback (cf. Kaufman & Schunn, 2011). Alternatively, an explanation could be that perceptions of a single peer feedback experience do not weigh enough into affecting students' attitudes, beliefs and/ or performance. If the influence of students' peer feedback perceptions indeed depends on their cumulative experience of multiple peer feedback occurrences over time (e.g., van Zundert, Sluijsmans, & van Merriënboer, 2010), this would suggest a more longitudinal approach for investigating the relation between peer feedback perceptions and writing performance.

Limitations

The finding that peer feedback providers and –receivers improve their writing to similar degrees suggests that the learning mechanisms involved in these different activities are similarly beneficial. Some caution is in place, however. Although we consider it plausible to regard the performance increase of providers and receivers as a result of the peer feedback phase (serious participation was both mandatory and incentivized), the authentic context of the current study made it practically and ethically unfeasible to incorporate a no-feedback control group. Hence, we can only refer to the *relative* writing performance increase of feedback providers versus –receivers, and not to their absolute performance increase as compared to a true baseline measure. Clearly, future studies that are able to combine the inclusion of such a control group with an authentic context could provide meaningful information regarding the absolute effects of providing and receiving peer feedback on students' academic writing.

With respect to the nature of the peer feedback and students' perceptions thereof, the empirical findings in this study – as in most empirical studies – are inherently limited as they represent one measure in time for one particular group of students. Although the participating students did have at least one prior experience with peer feedback on writing, a specific peer feedback training was not integrated in their curriculum. Therefore, and in addition to the potential backwash effect resulting from the weighting of the assignment criteria, this lack of training should be considered when comparing the nature of the reported peer feedback in this study with that reported in other studies.

Implications

The current study's findings are informative for higher education professionals who contemplate the design and implementation of peer feedback training within their course or curriculum. Specifically, in designing such peer feedback training, we believe that our findings regarding students' feedback role provide higher education professionals with a degree of flexibility. In addition, the importance of explanatory peer comments indicates what should be included in such a training for students.

We would argue for a more longitudinal approach (e.g., van Zundert, Sluijsmans, & van Merriënboer, 2010), both for training students for the peer feedback process as for researching the relations between the nature of the peer feedback students produce, their perceptions thereof and subsequent learning outcomes. Within such a longitudinal context, the current study's findings suggest that students could confidently, at first, only be involved in the provision of peer feedback in order to avoid issues such as distrust in each other's quality as assessor (for example by initially withholding the feedback provided by students' peers). After all, this study suggests that students' learning gains are similarly affected by providing peer feedback and receiving it. In addition, students may also *perceive* the act of providing peer feedback as the most beneficial part of the peer feedback process (cf. McConlogue, 2015). When students gain experience and follow training, among others with respect to the importance of explanatory peer feedback, students may perceive the peer feedback as increasingly positive as a result of increasing peer feedback quality. Through such iterative experiences, a classroom-culture can be developed in which peer feedback is accepted or even is the norm (see McConlogue, 2015, for a similar rationale). It should be mentioned here that we are currently conducting such a longitudinal inquiry with a large group of biopharmaceutical science students. Specifically, these students are followed during their first three semesters of their undergraduate program with respect to their peer feedback quality and their perceptions thereof, and with respect to how these measures relate to their academic writing. In addition, students' more general attitudes towards peer feedback are investigated to gauge their support for peer feedback as an instructional method across this period of time. With the conception in mind that peer feedback is an important academic and professional skill in itself (e.g., Liu and Carless 2006; Nicol, Thomson, & Breslin, 2014) which should be trained as such, we believe that such a longitudinal approach is a promising way to address the development of students' peer feedback skills and their attitudes towards it.

Acknowledgements

We would like to thank Kim Stroet for facilitating the data collection of this study. Further, we would like to thank Vivian van der Werf for her assistance in coding the peer feedback, and Marleen Danel and Suzanne Andeweg for their assistance in grading the essays. Finally, thanks go out to Benjamin Telkamp for inspiring confidence to using the R language and to Kirsten Ran for her help with the questionnaires.

Statement on Open Data

The anonymized data and syntaxes are accessible via the following link: https://osf.io/awkd9

5

A questionnaire to assess students' beliefs about peer feedback

6

Abstract

Research into students' peer feedback beliefs varies both thematically and in terms of approaches and outcomes. The current study describes the development of the concise though comprehensive Beliefs about Peer Feedback Questionnaire (BPFQ). Based on the different themes in the literature four scales were conceptualized. In separate exploratory (N=219) and confirmatory (N=121) studies, the structure of the questionnaire was explored and tested. The analyses confirmed the a priori conceptualized four scales: (1) students' valuation of peer feedback as an instructional method, (2) students' confidence in the quality and helpfulness of the feedback they provide to a peer, (3) students' confidence in the quality and helpfulness of the feedback they receive from their peers and (4) the extent to which students regard peer feedback as an important skill. The value of this practically applicable BPFQ is discussed with regard to future research into students' peer feedback beliefs and with regard to the insights it may offer higher education teaching staff.

Keywords: peer feedback; peer assessment; student beliefs; questionnaire

Introduction

Belief systems help a person to define and understand the world and one's place within that world, functioning as a lens through which new information is interpreted. Not surprisingly therefore, most definitions of 'beliefs' emphasize how these guide attitudes, perceptions and behavior (Pajares, 1992). Considering beliefs as a direct precursor to attitudes and behavior, (Ajzen, 1991; Ajzen & Fishbein, 2005), the current study describes the need for, and development of a questionnaire to assess higher education students' beliefs about peer feedback.

Given this interpretation of beliefs, students' educational beliefs are likely to influence both their perceptions and behavior during learning processes. For example, students' beliefs regarding the utility of a task may relate to their effort, time-on-task and performance (see Hulleman et al., 2008). In the context of peer feedback, this could mean that students' active engagement in the peer feedback process is contingent upon the degree to which they believe that peer feedback contributes to their learning and/or is an important skill to acquire. At the same time, students' peer feedback beliefs can also be regarded as an *outcome* of the peer feedback process (van Gennip et al., 2009). Here, a relevant overview is provided by van Zundert et al. (2010) Among others, their review focused on how student training and experience in peer feedback influence students' attitudes towards peer feedback. Van Zundert et al. found that twelve out of the fifteen studies reported positive attitudes towards peer feedback. At the same time, they also concluded that 'It is notable that, whereas the procedures varied tremendously, there was also an enormous variety in the instruments used to measure student attitudes' (p.277). In other words, a single comprehensive measure of students' peer feedback beliefs appears to be missing. Such a comprehensive measure seems imperative as peer feedback is frequently applied within higher education, and as the availability and user-friendliness of (often web-based) instruments increases. From an academic perspective, such a measure could facilitate the alignment of research findings. The resulting comparability of research findings across different contexts could allow for more generalizable conclusions with regard to students' beliefs about peer feedback

and the factors that influence those beliefs. From a practical perspective, such a measure could assist higher education teaching staff in understanding how their peer feedback practice affects students' experience of, and support for peer feedback as an instructional method. Therefore, the purpose of the current study is to develop and test such a practical, comprehensive instrument for investigating students' beliefs about peer feedback.

Themes of Student Beliefs in Prior Research

Across prior studies investigating students' beliefs in relation to peer feedback, different approaches have been adopted to address a variety of themes. Nevertheless, three broader themes can be distinguished in the literature. The first concerns students' beliefs about the value of peer feedback as an instructional method. The second and third theme concern students' confidence in either themselves or their peers as reliable assessors of quality. Within the concise, comprehensive instrument that is developed and tested in the current study, these three themes are conceptualized and integrated as separate constructs. To illustrate how these themes are derived from the literature, the following sections describe prior research approaches and –findings on these different aspects of students' peer feedback beliefs.

Peer feedback as an instructional method. Regarding students' valuation of peer feedback as an instructional method within their educational context, prior research tends to ask students questions such as how they value the peer feedback activity, whether they believe that students should be involved in assessing their peers and to the extent they believe that peer feedback contributes to their learning.

With respect to the involvement of students in formal feedback and the valuation of peer feedback activities, students generally appear to be positive. For example, McGarr and Clifford (2013) explicitly asked both undergraduate and postgraduate students how they valued peer assessment within their educational program. They found that both groups of students regarded peer assessment as valuable, although the postgraduate students valued it to a larger extent. Cheng and Warren (1997) found that 63.5% of the students believed

that students should take part in assessing their peers. Additionally, Li and Steckelberg (2004) asked students whether they believed peer assessment to be a worthwhile activity. On a scale of 1 (*strongly disagree*) to 5 (*strongly agree*), the 22 students scored a 4.18 on average, with all students scoring a 3 or higher. Also, Nicol et al. (2014) found students to hold positive beliefs with respect to peer feedback. After engaging in a peer feedback activity, which was the first such experience for most students, 86% reported to have a positive experience and 79% reported that they would definitely choose to participate again on future occasions. McCarthy (2017) also found that a majority of students was willing to receive peer feedback on future occasions, although here students were more positive towards future peer feedback in an online context (92% in favor) than in-class context (67% in favor). Other studies differentiated between students' beliefs regarding the provision and reception of peer feedback. For example, Palmer and Major (2008) found that students valued both aspects of the peer feedback process (scores ranging between 3.5 and 4.1 on a 5-point scale). In contrast to these generally positive findings, Liu and Carless (2006) findings were more ambiguous. These authors reported on a survey in which 1740 students were asked for their views on the purpose of assessment. Only 35% agreed with the notion that the development of 'students' ability to assess their classmates' should be a purpose of assessment, whereas 40% was neutral and 25% disagreed. Also, the study by Mulder et al. (2014) shows that, although students were relatively positive on forehand, the experience of the peer feedback process did lead to a small downward shift in their appreciation of peer feedback.

With respect to the impact of peer feedback, students generally appear to believe that it can contribute to their own learning. For example, Saito and Fujita (2004) asked 45 students how helpful they considered the comments and marks to be that they both *received from* and *provided to* peers. Scores ranged between 3.12 and 3.26 on a scale ranging from 1 (most negative) to 4 (most positive), suggesting that students regarded both aspects of the peer feedback process as contributing to their own learning. Similarly, 55% of the surveyed students in the study by Nicol et al. (2014) reported that they learned from

both the provision and reception of peer feedback. In the focus group data of the same study, however, students' beliefs with respect to the benefits of providing peer feedback appeared more salient, a finding that is corroborated by the in-depth case study by McConlogue (2015). Wen and Tsai (2006) also reported on the extent to which students believe peer feedback to contribute to their own learning. In their study, 280 students responded to statements such as 'peer assessment is helpful to my learning' and 'peer assessment activities can improve my skills' using a Likert scale that ranged from 1 (*strongly disagree*) to 5 (*strongly agree*). Students responded moderately positive to these statements, although there was a notable variation in responses (*M* = 3.36, *SD* = 0.89 and *M* = 3.55, *SD* = 0.92, respectively). Taken together, students appear to hold at least moderately positive beliefs about the value of peer feedback as an instructional method.

Confidence. Across prior studies, issues revolving around students' confidence tended to focus on either the confidence they had in their own competence or that of their peers. More specifically, both themes generally addressed the extent to which students consider themselves or their peers as eligible assessors of quality and to what extent they believed their own or their peers' comments or ratings to be reliable and helpful.

With respect to students' beliefs about the eligibility of their peers as assessors of quality and the reliability and helpfulness of their peers' feedback, Wen and Tsai and colleagues (e.g., Wen & Tsai, 2006; Wen et al., 2006) asked students to respond to statements such as 'I think students are eligible to assess their classmate's performance' and 'I think students should not be responsible for making assessments'. Unfortunately, only mean scores were reported for the latter item. Students' average score on this item was 2.63 (*SD* = 0.91) on a 5-point scale, indicating a more or less even split with respect to students' general belief about the role and responsibility of students in formal feedback. Focusing more on the notion of reliability, Saito and Fujita (2004) directly asked students to what extent they considered their peers to be reliable raters. Students' average response to this question was 2.96 (SD = 0.60) on a 4-point scale, suggesting that students hold moderately positive beliefs about the reliability of their peers' ratings.

Students' confidence in their own competence as an assessor could de facto be considered as a context-specific self-efficacy beliefs (cf. Pajares, 1992). Sluijsmans et al. (2004) investigated such self-efficacy beliefs that students hold. In particular, they addressed students' self-perceived assessment skills through an eight-item scale, which included items such as 'I am able to analyze a product of a peer'. Across conditions and timing of measurements mean scores on this scale ranged between 3.69 and 3.89 on a 5-point Likert scale, suggesting that these students were fairly confident in their own competence. McGarr and Clifford (2013) also asked students whether they regarded themselves as having the knowledge and skills to assess their peers. On a scale ranging from 0 (*strongly agree*) to 11 (*strongly disagree*), the response of both undergraduate and postgraduate students indicated that these groups were relatively confident in this respect (on the reversely phrased item, both median scores = 8). In contrast, students in the study by Cheng and Warren (1997) were less confident in their own competence as assessor. In response to the question 'Do you think that you will make a fair and responsible assessment of your peers?', 36.5% said yes, 23.1% said no, and 40.4% was unsure. Possibly, the findings in these studies may differ as a result of differences in participant samples. In the Sluijsmans et al. (2004) study, participants were student-teachers, who are likely to have encountered peer feedback tasks to a larger extent than the first-year undergraduate students in the study by Cheng and Warren (1997). However, a multitude of contextual differences makes it difficult to directly compare these studies and interpret these varying outcomes.

Peer feedback skills as an important learning goal. In addition to the first three themes, we argue there is a fourth important aspect of students' peer feedback beliefs. This concerns the extent to which they regard peer feedback skills as being an important learning goal in themselves. Although we did not encounter empirical research that explicitly addressed this aspect of students' peer feedback beliefs, we do believe that the theoretical relevance of this factor warrants its inclusion. After all, students' engagement in the peer feedback process may be contingent on the extent to which they regard peer feedback skills as important to acquire or develop. According to expectancy-value

theory, for example, subjective task value influences the achievement-related choices students make (e.g., Wigfield & Eccles, 2000). In particular, the valued utility of a task appears to positively relate to students' effort, time-on-task and performance (e.g., Hulleman et al., 2008). In addition, higher education students are the future members of academic or other professional organizations. Consequently, being able to provide, receive and utilize feedback from peers could – or indeed *should* – in themselves be considered as important learning goals in higher education curricula (see also Cowan, 2010; Liu & Carless, 2006; Sadler, 2010; Sluijsmans et al., 2004; Topping, 2009).

Research Aims

The current study describes the development of the Beliefs about Peer Feedback Questionnaire (BPFQ) and has two central aims. Our first aim is to construct and validate a concise, comprehensive questionnaire that addresses the four following themes: students' valuation of peer feedback as an instructional method within their educational context, students' confidence in the quality and helpfulness of the feedback they provide to a peer, students' confidence in the quality and helpfulness of the feedback they receive from their peers and the extent to which students regard peer feedback skills in themselves as an important learning goal. In doing so, our second aim is to provide a practically applicable instrument to both academic researchers and higher education teaching staff that comprehensively assesses students' peer feedback beliefs.

Method

The BPFQ was constructed in three steps. In step one, a concise questionnaire was developed to address the four above mentioned themes, which were conceptualised in four scales: 'Valuation of peer feedback as an instructional method' (VIM; 4 items), 'Confidence in own peer feedback quality' (CO; 2 items), 'Confidence in quality of received peer feedback' (CR; 2 items) and 'Valuation of peer feedback as an important skill' (VPS; 3 items). Items of

the VIM scale related to, for example, Cheng and Warren (1997), Li and Steckelberg (2004) and Palmer and Major (2008). Items of the CO scale related to Sluijsmans et al. (2004) and Cheng and Warren (1997), whereas items of the CR scale were based on Wen and Tsai and colleagues (e.g., Wen & Tsai, 2006; Wen et al., 2006) and Saito and Fujita (2004). Finally, the VPS scale was designed to assess how important students regarded three different skills within the peer feedback process: providing peer feedback, dealing with critical peer feedback and utilizing it for improving one's work. These three were conceived as applicable and generalizable to future contexts in which students were likely to arrive at some point in time, either within their studies or during their subsequent careers. In step two an *exploratory* study was conducted. Using the data from this study, principal component analyses were performed to assess the initial component structure of the BPFQ. The first principal component analysis indicated that one item of the initial VIM scale ('Involving students in feedback through the use of peer feedback is instructive') did not uniformly load on one single component. This item was therefore omitted in all subsequent analyses. A second and third principal component analysis were performed on the remaining ten items to compare the proposed model with four scales to a model without a predefined number of components (cf. Visser-Wijnveen et al., 2016). In the third and final step, a *confirmatory* study was conducted. In particular, two confirmatory factor analyses were performed to compare the proposed model and the non-fixed model in terms of their fit on the data.

Participants and Data Collection Procedure

In the *exploratory* study, the questionnaire was completed by 220 second-year Biopharmaceutical Science students from a large research-intensive university in The Netherlands. The data for one student was dropped as cases with missing data were deleted list-wise. The mean age of the 219 included students was 19.51 years (SD = 1.39) with 140 students (63.9%) being female. From the very start of their undergraduate program, these students were introduced to peer feedback as an instructional method through explanation, instruction, exercises, and formative peer feedback activities. Over the course of the first

three semesters, the role of peer feedback gradually expanded, with the ultimate aim of the teaching staff being that students would perceive peer feedback as a normal and integral part of formal feedback. In the *confirmatory* study, the questionnaire was administered to a group of first-year students in Education and Child Studies (N=121). Their mean age was 19.48 years (*SD* = 1.62) with 114 students (94.2%) being female. These students had at least one prior experience with anonymous online peer feedback in the context of an academic writing assignment, and were expected to engage in two similar peer feedback activities in the semester that had just started. All students received the questionnaire during the starting lecture of a course. Questionnaires were administered in paper-and-pencil format and responses on all items were provided on a 5-point Likert scale. For the VIM and VPS scales, responses could range between 1 (*completely disagree*) and 5 (*completely agree*), whereas responses could range between 1 (*completely not applicable to me*) and 5 (*completely applicable to me*) for the CO and CR scales.

Analyses

Principal component analyses were performed using SPSS (v23) to empirically explore the underlying structure of the questionnaire. As we anticipated the conceptualized scales to be correlated, oblique (oblimin) rotation was applied. In the second study, confirmatory factor analyses were conducted to test the construct validity of the questionnaire. To this end, the 'lavaan' package (v0.5-23.1097; Rosseel, 2012) in R (v3.4.2; R Core team, 2017) was used. For the final scales emerging from the confirmatory analyses, internal reliability was computed as Cronbach's alpha and the relations between these scales were assessed in terms of Pearson correlations. The anonymized data and analyses (syntaxes) are available as supplemental online materials.

Results

Exploratory Analyses

In the exploratory study, two principal component analyses were conducted on the retained ten items. The a priori proposed model consisting of four fixed components (see Table 1) was compared to a model without a pre-fixed number of scales, effectively allowing a 'bottom-up' structure to emerge from the data. For both models, sampling appeared adequate (*Kaiser-Meyer-Olkin* measure of sampling adequacy = .69, individual items values ranging from .57 to .87) and inter-item correlations appeared sufficiently large (Bartlett's test of sphericity $\chi^2(45) = 630.97, p < .001$. The non-fixed principal component analyses provided a three-component structure, aggregating the VIM and CR scales (see Table 2). This bottom-up model suggests that the beliefs about the value of peer feedback as an instructional method (VIM) on the one hand and, on the other hand students' confidence in the quality of the feedback they receive from their peers (CR) tapped into the same conceptions that these first-year Biopharmaceutical Science students held. Comparing these two exploratory models, the total common variance was higher for the items in the proposed model with four fixed components (average of communalities being 0.718) than for the items in the non-fixed model with three components (average of communalities being 0.624).

Confirmatory Analyses and Scale Reliability

To test which of the models best fitted students' response patterns in the second student sample – the a priori proposed four component structure that was based on the themes in the literature, or the bottom-up three-component structure – confirmatory factor analyses were conducted on the sample of Education & Child Studies students to compare the two. As the principal component analyses generally indicated relatively low correlations between components, neither of the two confirmatory models included between-factor correlations. The proposed four factor model ($\chi^2(29) = 56.78$, *p* = .002, *TLI* = .91, *CFI* =

.94, *RMSEA* = .089 [.05, .12], *SRMR* = .06) appeared to fit the data better than the bottom-up 3 factor model that emerged in the exploratory phase (χ^2(32) = 117.69, *p* < .001, *TLI* = .75, *CFI* = .82, *RMSEA* = .15 [.12, .18], *SRMR* = .11). Although the difference in fit between these two models cannot be directly tested given that these models are not nested, the fit indices do seem to suggest that the construct validity of the a priori proposed four-factor model is superior to that of the three-factor model. In fact, the four-factor model's fit indices either adhere to or approach the cut-off values suggested by Hu and Bentler (1999), which may be considered relatively stringent (e.g., Perry et al., 2015). Therefore, the final BPFQ was considered to be best described in terms of the four scales that were conceptualized on forehand. The respective scale-reliabilities were acceptable (see Table 3), especially given the concise nature of the individual scales (cf. Cohen, 1988; Cortina, 1993; see Tables 3 and 4).

Discussion

The current study aimed to develop and test a concise and comprehensive questionnaire to assess students' peer feedback beliefs. A priori, the BPFQ was conceptualized to include four scales. Three scales were constructed to cover as many themes emerging from the literature: students' valuation of peer feedback as an instructional method, students' confidence in the quality and helpfulness of the feedback they *provide to* their peers, and students' confidence in the quality and helpfulness of the peer feedback they *receive*. In addition, the extent to which students regarded peer feedback skills as an important learning goal was considered an important aspect of their peer feedback beliefs. Hence, this was conceptualized as the fourth scale within the BPFQ. An exploratory and a confirmatory study were conducted on two separate groups of students. After one ambiguous item was omitted, both the exploratory and the confirmatory analyses confirmed these a priori conceptualized four scales. Hence, the final BPFQ consisted of ten items within four scales: students' valuation of peer feedback as an instructional method (VIM; 3 items), students' valuation of peer

Table 1. PCA component loadings (*N* = 219; four factors fixed a priori).

Scale	Items	Components			
		I	II	III	IV
VIM	Involving students in feedback through the use of peer feedback is meaningful	**.67**	-.21	.06	.11
	Peer feedback within [course] is useful	**.65**	-.13	.00	.22
	Feedback should only be provided by the teaching staff [*reversed*]	**.86**	.12	.00	-.11
Removed	*Involving students in feedback through the use of peer feedback is instructive*				
VPS	Being capable of giving constructive peer feedback is an important skill	.19	**-.70**	.01	-.08
	Being capable of dealing with critical peer feedback is an important skill	-.08	**-.87**	.04	-.03
	Being capable of improving one's work based on received peer feedback is an important skill	-.04	**-.86**	-.05	.03
CO	In general, I am confident that the peer feedback I provide to other students is of good quality	-.03	.05	**.92**	-.04
	In general, I am confident that the peer feedback I provide to other students helps them to improve their work	.03	-.04	**.86**	.05
CR	In general, I am confident that the peer feedback I receive from other students is of good quality	.02	.07	.04	**.87**
	In general, I am confident that the peer feedback I receive from other students helps me to improve my work	-.01	.00	.00	**.91**
Eigenvalue		3.08	2.09	1.07	0.94
% Variance explained		30.82	20.86	10.72	9.42
Scale reliability (Cronbach's α)		.67	.76	.73	.78
Component correlations	I		-.32	.19	.25
	II			.19	-.06
	III			-.03	.35

Table 2. PCA component loadings (*N* = 219; no fixed factors).

Scale (*initial*)	Items	I	II	III
		Components		
VIM	Involving students in feedback through the use of peer feedback is meaningful	**.62**	-.34	.01
	Peer feedback within [course] is useful	**.69**	-.23	-.03
	Feedback should only be provided by the teaching staff [*reversed*]	**.58**	-.13	-.16
Removed	*Involving students in feedback through the use of peer feedback is instructive*			
VIM (CR)	In general, I am confident that the peer feedback I receive from other students is of good quality	**.68**	.30	.24
	In general, I am confident that the peer feedback I receive from other students helps me to improve my work	**.69**	.26	.22
VPS	Being capable of giving constructive peer feedback is an important skill	.12	**-.74**	.02
	Being capable of dealing with critical peer feedback is an important skill	-.05	**-.83**	.11
	Being capable of improving one's work based on received peer feedback is an important skill	.04	**-.80**	.02
CO	In general, I am confident that the peer feedback I provide to other students is of good quality	-.06	-.06	**.90**
	In general, I am confident that the peer feedback I provide to other students helps them to improve their work	.06	-.13	**.85**
Eigenvalue		3.08	2.09	1.07
% Variance explained		30.82	20.86	10.72
Scale reliability (Cronbach's α)		.70	.76	.73
Component correlations	I		-.21	.27
	II			.11

Note: VIM = Valuation of peer feedback as instructional method; CR = Confidence in quality of received peer feedback; CO = Confidence in own peer feedback quality; VPS = Valuation of peer feedback as an important skill.

feedback as an important skill (VPS; 3 items), students' confidence in the quality of the peer feedback they provide (CO; 2 items) and students' confidence in the quality of received peer feedback (CR; 2 items).

We believe the BPFQ is valuable both to academic researchers and higher education teaching staff. With respect to research into students' peer feedback beliefs, the availability of a comprehensive questionnaire could facilitate the alignment of research findings across contexts and disciplines, contributing to more coherent knowledge building in this area. For example, the consistent use of one instrument in multiple research contexts may shed light on how varying aspects of the design of peer feedback tasks (see Gielen et al., 2011, for an overview) influence students' peer feedback beliefs. In addition, the concise nature of the BPFQ could facilitate longitudinal research into students' peer feedback beliefs. In the higher education literature, peer feedback is increasingly recognized as important learning goal in itself (e.g., Cowan, 2010; Liu & Carless, 2006; Sadler, 2010; Sluijsmans et al., 2004). As students' peer feedback beliefs are likely to be influence through cumulative experiences over time, such longitudinal approaches are likely to involve multiple measurements. In such contexts, minimizing the burden on students' time seems highly desirable – if not pivotal – in making such repeated measures practically feasible. For the same reasons, the concise nature of the BPFQ may also assist higher education teaching staff in understanding how their peer feedback practice affects students' experience of, and support for peer feedback. In terms of students' experiences for example, the BPFQ could function as an evaluative measure that informs higher education staff on how to improve peer feedback within a course or curriculum. In terms of students' support for peer feedback, the BPFQ could for example be administered at the start of a course or semester. Having a priori information about students' peer feedback beliefs could provide teaching staff with the opportunity to address issues around students' confidence or their awareness of the importance of peer feedback skills. Especially in the case of student beliefs, it may be critical to act upon such information in a timely fashion given that students' early experiences can strongly influence judgments, which in turn become beliefs that may be relatively resistant to change (Pajares, 1992).

Table 3. BPFQ descriptive statistics, reliability indices and scale correlations

Scale	Items	Biopharmaceutical Science (N=219)						Education & Child Studies (N=121)					
		Descriptives			*Scale correlations*			*Descriptives*			*Scale correlations*		
		Mean	SD	α	VPS	CO	CR	Mean	SD	α	VPS	CO	CR
VIM	3	3.72	0.68	.67	.39**	.23**	.32**	3.84	0.76	.81	.32**	.23*	.35**
VPS	3	4.28	0.54	.76	-	.02	.02	4.23	0.51	.73	-	.29**	.29**
CO	2	3.49	0.68	.73		-	.37**	3.71	0.62	.82		-	.43**
CR	2	3.41	0.65	.78			-	3.64	0.67	.75			-

Note: VIM = Valuation of peer feedback as instructional method; CR = Confidence in quality of received peer feedback; CO = Confidence in own peer feedback quality; VPS = Valuation of peer feedback as an important skill; * = p < .05 (two-tailed); ** = p ≤ .01 (two-tailed)

Table 4. Final scales and items for the Beliefs about Peer Feedback Questionnaire

Scale	#	Item text
I		*Valuation of peer feedback as an instructional method ('VIM')*
	1	Involving students in feedback through the use of peer feedback is meaningful
	2	Peer feedback within [course] is useful
	3	Feedback should only be provided by the teaching staff [*reversed*]
II		*Valuation of peer feedback as an important skill ('VPS')*
	4	Being capable of giving constructive peer feedback is an important skill
	5	Being capable of dealing with critical peer feedback is an important skill
	6	Being capable of improving one's work based on received peer feedback is an important skill
III		*Confidence in own peer feedback quality ('CO')*
	7	In general, I am confident that the peer feedback I provide to other students is of good quality
	8	In general, I am confident that the peer feedback I provide to other students helps them to improve their work
IV		*Confidence in quality of received peer feedback ('CR')*
	9	In general, I am confident that the peer feedback I receive from other students is of good quality
	10	In general, I am confident that the peer feedback I receive from other students helps me to improve my work

Note: All items answered through 5-point Likert scale; For scales I and II, the labels range from 1 (completely disagree) to 5 (completely agree); For scales III and IV the labels range from 1 (completely not applicable to me) to 5 (completely applicable to me).

6

Limitations and Future Research

Some limitations need to be addressed. For one, additional sampling is required to confirm the external validity of the BPFQ. Although we purposefully sampled different groups of students for the exploratory and the confirmatory analyses, all participants in the current study were undergraduate students within the same university. As a result, their beliefs about peer feedback may be influenced by some common denominator, such as the general likelihood of being involved in peer feedback or the (digital) tools used to organize peer feedback. Hence, future applications within other higher education institutes and disciplines are needed to assess the extent to which the BPFQ continues to function consistently across contexts. Second, the BPFQ may not be exhaustive with respect to the potential variety of peer feedback beliefs that students' may hold. The four BPFQ scales were constructed based on the themes covered in prior research. Nevertheless, there may be aspects of students' peer feedback beliefs that the current BPFQ does not cover, for example because some of those aspects may currently be underrepresented in the literature. One way to address this could be through systematic, in-depth interviews with both graduate and undergraduate students from varying institutes and disciplines. Despite these inherent limitations, we are confident that this study provides a practical (concise) and comprehensive questionnaire to address students' beliefs about peer feedback. In particular, we demonstrated that the construct validity of the BPFQ is acceptable and that individual scale reliabilities are sufficient. We therefore believe that this questionnaire can contribute to higher education research by facilitating the comparability of research findings, and believe that it can help higher education teaching staff in understanding how their peer feedback practice affects students' experience of, and support for peer feedback.

Statement on Open Data

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

General discussion

7

Within a broader context of institutional initiatives to increase study success and decrease student attrition across bachelor programs, this thesis focused on the design of formative peer feedback tasks and the effects thereof on students' academic writing performance. In particular, this thesis investigated to what extent formative peer feedback impacts higher education students' academic writing performance and how particular aspects of peer feedback task-design affect students' writing performance. An important aim was to amalgamate both scientific and practical value. In terms of its practical value for higher education teaching staff, this thesis focused on relatively controllable aspects of peer feedback task-design. In terms of its contribution to the peer feedback literature, this thesis aimed to contribute to the limited number of studies adopting relatively controlled research designs. To reiterate, formative peer feedback was 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 for the purpose of learning'. Hence, formative peer feedback in this thesis could include information ranging from simple scores to elaborate comments, generally encompassing both 'peer assessment' and 'peer feedback' insofar as these reflect different practices in the literature.

Main Findings per Chapter

Chapter 2. The impact of formative peer feedback on writing performance: A meta-analysis

This study described a meta-analysis of quantitative studies reporting on the impact of formative peer feedback on higher education students' academic writing performance. Two sets of research questions were addressed. First, the synthesized effects of peer feedback on academic writing were compared both to a no-feedback baseline and to two oftentimes feasible alternatives: self-assessment and feedback from teaching staff. Second, we explored the moderating role of two variables that are simultaneously important in the design of peer feedback tasks (Gielen et al., 2011) and that were perceived as controllable by higher education teaching staff. Specifically, the number of peers that a student engaged with during peer feedback and the nature of the peer feedback (scores,

comments, or both) were assessed with respect to their moderating role in explaining students' writing improvement after peer feedback. Results indicated that the effects of peer feedback on academic writing performance tend to be larger than that of either no feedback or self-assessment, whereas the effects of peer- and teacher feedback do not appear to differ significantly. Results of the moderation analyses indicated that a combination of both peer comments and –scores tends to have a larger effect on writing performance than either peer comments or –scores alone. In contrast, the number of peers with whom a student engaged during peer feedback did not moderate subsequent writing performance, although there appeared to be a trend suggesting that engagement with multiple peers is more beneficial than engagement with a single peer. These results suggested that higher education teaching staff design their peer feedback tasks in such a way that students provide both comments and scores to each other's writing. A notable limitation of this meta-analysis concerned the limited number of studies that proved eligible for inclusion. We argued that this limitation signals an important observation regarding the current body of literature on the topic, namely that there is a need for more well-controlled, (quasi-)experimental studies into peer feedback on writing in HE.

Chapter 3. Students' ability match, the nature of peer feedback and writing performance

The literature remains inconclusive with regard to how students are optimally matched during peer feedback. Meanwhile, the increasing availability and user-friendliness of (often web-based) applications increases the options and opportunity for higher education teaching staff with respect to the design and implementation of peer feedback tasks. The third chapter described a quasi-experimental study that explored the relation between students' ability match, the nature of the peer feedback and subsequent performance in the context of an authentic academic writing task. Contrasting two different perspectives on student ability matching from the literature, students were assigned to a dyad together with either a similar-ability or different-ability peer. Results indicated that students in similar-ability and different-ability dyads do not differ

in the extent to which they improve their writing. In addition, neither dyad composition nor students' individual ability significantly related to the nature of the peer feedback they provided, although a trend suggested that content-related peer feedback may be more prevalent within similar-ability dyads consisting of two high ability students. Finally, the nature of the received peer feedback did not relate to the subsequent writing performance of either lower or higher ability authors. This last finding suggested that relatively high or low ability authors do not benefit differently from peer feedback on either content-, structure- or style-related aspects of their writing. In summary, students' ability match, the nature of the peer feedback and academic writing performance were not significantly related. Some remarks were in place, however. For one, this study did not take into account students' perceptions with respect to the peer feedback they received (cf. Hanrahan & Isaacs, 2001; Strijbos et al., 2010). Also, this study did not disentangle the separate effects that providing and receiving peer feedback could have on students writing performance (e.g., Lundstrom & Baker, 2009; Cho & MacArthur, 2011). Consequently, how students' peer feedback perceptions related to their subsequent writing performance remained a question for future research, and the reported effects of peer feedback on students' writing reflected the combined effects of both providing and receiving peer feedback. Finally, the students in this study were relatively similar in terms of gender, age and educational background, which could mean that the differentiation between higher and lower ability students was one within a restricted range. This may explain why only non-significant trends were found and, at the same time, suggests that these trends could be considered informative; if these patterns emerged within a sample that is relatively homogeneous in terms of students' ability, they may become more salient as heterogeneity increases, such as in open online educational contexts.

Chapter 4. Students' ability match and academic writing performance in a MOOC

The number of massive open online courses (MOOCs) and participants therein has rapidly expanded over recent years. This has increased the relevance of issues such as formal recognition of learning and accreditation, and raises the question how to optimally scale up the assessment of complex, open-ended assignments. In MOOCs, the answer to this oftentimes involves peer assessment in the form of both comments and grades between multiple peers. Following up on the on-campus study described in chapter three, this study conducted a first exploration of how participants' ability match during peer feedback relates to their writing performance. The subjects in this study were 565 participants in a Leiden University MOOC on Terrorism and Counterterrorism. They were categorized as either relatively high, intermediate or low performers based on available performance metrics prior to the first essay assignment. Post hoc analyses were conducted to assess the relation between participants' own performance level, that of their reviewing peers', and participants' subsequent increase in writing performance between essays. Results indicated that the average performance level of the assessing peers positively relates to participants' subsequent increase in writing performance. A closer look revealed an interaction between participants' own performance level and that of their peer assessors. Specifically, peer assessors' average performance level only related to the writing performance of intermediate and higher performing participants, not to that of relatively low performing participants. Effect sizes were small, however. Possible interpretations of these results could be that participants' relatively low, intermediate and high performance level also reflects their ability to utilize peer feedback. Alternatively, or complimentary, these findings indicated that peer feedback quality could be improved. Certain limitations needed to be considered when interpreting the results of this exploratory study. For example, the lack of background information on participants left open the question of what the exact mechanisms were through which the performance level of participants and that of their peer assessors interacted. Similarly, it also remained a question for future research what the role of peer feedback quality and quantity is in explaining MOOC participants' writing performance. Still, this study provided a first insight in how the match between MOOC participants and their peer assessors related to participants' writing performance, contributing to the knowledge base on how to optimally design peer assessment tasks in MOOCs.

Chapter 5. Students' peer feedback role, peer feedback perceptions and writing performance

Research on the mechanisms involved in the peer feedback process has focused on elements with respect to both the provision of peer feedback and the reception thereof. On the one hand, a learning-by-reviewing rationale suggests that providing peer feedback stimulates students' active problem detection, – diagnosis, and subsequent contemplation of revision strategies (e.g., Flower et al., 1986; Patchan & Schunn, 2015). On the other hand, receiving (peer) feedback provides students with information on the gap between current performance, goal performance, and how to close that gap (e.g., Nicol & Macfarlae-Dick, 2006; Hattie & Timperley, 2007). However, how these different aspects of the peer feedback process compare in terms of their impact on students' academic writing performance largely remains an open question. This study compared the impact of students' feedback role – being either the provider or receiver of peer feedback – on their subsequent performance increase in the context of an academic writing task. Additionally, the relations between the nature of the received peer feedback, students' perceptions thereof, and performance increase were investigated. Results indicated that feedback providers and receivers improved their writing to similar degrees, and that explanatory peer feedback mostly influenced the extent to which students perceived the received peer feedback as adequate. However, no direct relations were found between students' perceptions of the received peer feedback and their subsequent increase in writing performance. As a no-feedback control-group was not ethically feasible, one limitation to bear in mind was that this study only referred to the *relative* effect that providing or receiving peer feedback has on students' writing performance. Nevertheless, these results were believed to elucidate two findings that are informative for higher education researchers and –teaching staff alike. First, for students that are unfamiliar with the peer feedback process, students' role may at first be that of peer feedback provider only. Initially withholding received peer feedback could avoid issues such as students' distrust of their peers' feedback quality, while the exercise of providing peer feedback would still be likely to be beneficial to students' writing performance. Second,

there was the finding that explanatory peer feedback most strongly related to students' perceptions of adequacy, which was argued to be important with respect to students' more general support for – and engagement in – the peer feedback process. Recognizing the importance of students' support for peer feedback, which is likely to be shaped by multiple peer feedback experiences, this finding highlighted the importance of emphasizing the role of explanations in peer feedback training and instruction.

Chapter 6. A questionnaire to assess students' beliefs about peer feedback

Research on peer feedback may focus on different types of outcomes, including performance, perceptions of learning and beliefs about assessment (van Gennip, Segers, & Tillema, 2009). In case of students' beliefs about peer feedback, research appears to vary both in terms of the instruments used and in terms of reported outcomes (see van Zundert, Sluijsmans, & van Merriënboer, 2010, for an overview). Chapter six described the development of the Beliefs about Peer Feedback Questionnaire (PBFQ). This questionnaire served a dual aim. For one, it aimed to contribute to the alignment and, consequently, comparability of research findings. Simultaneously, it aimed to provide a practical instrument for higher education teaching staff to monitor how their teaching practice influences students' peer feedback beliefs. Four subscales were conceptualized based on an initial set of eleven items, tested in a separate exploratory study (N = 219) and a confirmatory study (N = 121). The a priori conceptualized model with four scales was consistently found to best fit the data. These four scales related to: (1) students' valuation of peer feedback as an instructional method, (2) students' confidence in the quality of the peer feedback they provide to a peer, (3) students' confidence in the quality of the peer feedback they receive from a peer and (4) the extent to which students regard peer feedback as an important skill. Individual scale reliabilities ranged between $\alpha = .67$ and $\alpha = .82$ across both samples. Although additional testing across other institutes and disciplines would be desirable to further confirm the external validity of the BPFQ, a first successful step in this respect was made by purposefully conducting the confirmatory analyses on a different group of students. Hence, we believe

that this study provides a comprehensive instrument to assess students' peer feedback beliefs. Additionally, we believe that the concise nature of the BPFQ makes it a practically applicable instrument, both for higher education teachers that want to conduct research within their own teaching practice as for researchers that aim to monitor the development of students' peer feedback beliefs (e.g., longitudinal designs).

Conclusion

The overarching research question in this thesis addressed two related issues: the available evidence for the impact of formative peer feedback on students' academic writing performance and the role of specific peer feedback task-design aspects in explaining writing performance. Regarding the available evidence, the findings presented in the meta-analysis (chapter two) may be considered as somewhat surprising. Despite the continuous development of a framework on the most relevant aspects of peer feedback task-design (cf. Topping, 1998; van den Berg, Admiraal, & Pilot, 2006b; van Gennip, Segers, & Tillema, 2009; Gielen, Dochy, & Onghena, 2011), and despite repeated calls by scholars for more systematic and controlled studies into the effects of peer feedback (e.g., Topping, 1998, 2010; Strijbos & Sluijsmans, 2010), only 16 studies with *some type of* reference group were eligible for inclusion. This clearly signals a gap in the current peer feedback literature that is worth noticing and that merits the attention of future research. Keeping in mind the limitations that this restricted number of studies imposes, the meta-analysis provided provisional evidence suggesting that engagement in peer feedback positively affects writing performance compared to a no-feedback baseline and compared to self-assessment, whereas no difference was found when compared to the effects of feedback from teaching staff.

Regarding the role that specific aspects of peer feedback task-design have in contributing to students' writing performance, chapters two through five covered four different task-design aspects that tend to be relatively controllable for HE teaching staff. First, the available evidence does not indicate that engaging in feedback with multiple peers contributes to larger writing improvements

than engaging in feedback with a single peer. However, this should probably be considered as a tentative finding because of the combination of a) the limited number of studies included in the meta-analysis, b) the non-significant pattern pointing towards larger writing improvements when multiple peers are involved and c) the theoretical arguments that would support a multiple-peer hypothesis. Second, the nature of peer feedback appears to matter in terms of students' perceptions of peer feedback adequacy and can relate to their writing performance depending on how the nature of peer feedback is defined. With respect to students' perceptions, the presence of explanatory peer feedback (i.e. peers' comments that explain or justify the peer feedback) best predicts how adequate the peer feedback is perceived to be. With respect to writing performance, a combination of peer scores and –comments appears to be more effective than either scores or comments alone. Focusing specifically on peer comments, however, the nature of those comments does not appear to be directly related to students' writing performance. Third, matching students based on prior performance appears to yield different results in different contexts. In particular, student matching within an on-campus course did not relate to subsequent writing performance, whereas the match between participants in a MOOC did positively relate to writing performance for the majority of them. Fourth, providing and receiving peer feedback appear to have a similar impact on students' own writing performance.

In addition to the role of these specific task-design aspects, over the course of the current thesis it became increasingly clear that students' support for peer feedback as an integral part of their education is pivotal. This insight developed based on various experiences during this thesis, including in-class observations and informal talks with students and with teaching staff. Recognizing the potential influence of students' beliefs about peer feedback on their related perceptions and behavior, and recognizing that these beliefs are likely to be shaped by multiple experiences over time, the Beliefs about Peer Feedback Questionnaire was developed. This questionnaire was developed to be both concise and thematically comprehensive, facilitating longitudinal inquiries into students' peer feedback beliefs.

Taken together, the current thesis furthers our knowledge on a) the available evidence for the impact of formative peer feedback on writing performance, b) how students' ability match and feedback role as either peer feedback provider or –receiver relate to writing performance, and c) the relations between the nature of the peer feedback and students' perceptions thereof. Provided that future research confirms these findings, several implications follow. For one, formative peer feedback positively affects higher education students' academic writing performance. In fact, it appears to be equally effective as formative feedback from teaching staff, which implies that higher education teaching staff can indeed regard peer feedback as an instructional method that simultaneously benefits students' learning and is logistically efficient in the case of large student-to-teacher ratios. However, it is far from self-evident that students think of teacher feedback as equally effective as peer feedback (cf. Liu & Carless, 2006; van Zundert et al., 2010; McConlogue, 2015), which could lead them to appraise the feedback they receive differently. There are understandable reasons for students' differing expectations. For example, teachers and students are not identical in terms of domain-specific knowledge and experience, which may result in feedback comments with a different focus (e.g., Cho & MacArthur, 2010). The current thesis found first-year students' peer feedback to be predominantly focused on aspects of writing style, and less on content– or structure-related aspects, although a trend suggested that higher performing dyads provide somewhat more content-related peer feedback than mixed or lower performing dyads. This could imply that peer feedback should be considered as a complimentary feedback source to teacher feedback during the first year(s) of higher education programs, and that it may be regarded as increasingly comparable to teacher feedback as students acquire more domain-specific knowledge and experience. The reported nature of first-year students' peer feedback also suggests that they should be trained and guided in providing good quality peer feedback. In particular, the findings presented in this thesis suggest that, from students' own perspective, such guidance training should emphasize the role of explanatory peer feedback comments.

These findings could be embedded in a broader framework of peer feedback training-design, which may simultaneously serve as a basis for comparing future studies that assess the effects of peer feedback training. Such a framework should include at least two complimentary aspects. One the one hand, it should include the knowledge and skills in which students are trained, including students' mastery of assessment criteria, students' ability to provide and utilize constructive peer feedback and their ability to make valid judgements of a peers' performance. On the other hand, it should include the extent to which the peer feedback training actively engages students in the process, which could for example range from simple instruction, to dialogue, to (guided) practice.

Naturally, there are strengths and weaknesses as a consequence of the methodological choices that were made and with respect to the practical implications and considerations.

Methodological reflections
The primary focus on students' academic writing performance naturally involves certain assumptions, such as that increasing grades on a writing task reflect the increase of a student's writing skills. Although it is not undebated what grades reflect (e.g., Knight, 2002), they generally remain the widest available proxy to students' learning in higher education. In addition, student performance as measured by grades aligns with Leiden University's initiatives focusing on study success. An alternative focus may be the critical appraisal of peer feedback by students (cf. Strijbos, Narciss, & Dünnebier, 2010; Gielen et al., 2011; Sadler, 2010). This is a potentially fruitful area for future research, the importance of which also became increasingly clear over the course of the current thesis. In both formal and informal conversations which are not included in this thesis, it became increasingly salient that students can have multiple motives for engaging with the received peer feedback or not. They could, for example, be skeptical with regard to their peers' legitimacy as assessor, or they were inclined to only engage in learning activities that were strictly required for passing a course. Such conversations highlighted the fact that feedback can only be expected to contribute to a student's learning when he or she mindfully engages with

it (e.g., Handley, Price, & Millar, 2011). From a more cognitive psychological perspective, this process of critical appraisal could be regarded as the bottleneck between on the one hand task- and learner characteristics and, on the other hand learning outcomes. In particular, it may be interesting to assess the interaction between specific learner characteristics (e.g., current proficiency, motivation, or task-related beliefs) and task characteristics (e.g., aspects of peer feedback task-design, the nature of peer feedback) in explaining critical appraisal. Some recent research has started to investigate such relations using, for example, eye-tracking (Bolzer, Strijbos, & Fischer, 2015), which may be a particularly promising approach to inferring a learner's mindful processing of (peer) feedback.

Another methodological reflection seems in place with respect to the focus of this thesis on the controllability of the task-design aspects. In this thesis, the notion of controllability was aligned with planned behavior theory (Ajzen, 1991; Ajzen & Fishbein, 2005). Six higher education teachers were interviewed and – using the variables from Gielen, Dochy, & Onghena's (2011) framework – performed a card-sorting task to assess which aspects of peer feedback task-design they perceived as controllable in their own teaching practice (chapter 2). Borrowing from planned behavior theory, in which perceived behavioral control is a key antecedent to intentions and behavior, teachers' perceived controllability was considered as a relevant indicator for the practical applicability of these variables. The interviewed teachers were asked to reason based upon their own specific teaching practice and experience and to weigh in practical constraints (e.g., available time and resources) when determining controllability. In other words, perceived controllability referred to teachers' perceptions of the extent to which they *could* adapt these variables. However, controllability did not directly incorporate the extent to which the teachers *would* adapt these design-aspects of peer feedback tasks. More elaborate research into the practicality (cf. Doyle & Ponder, 1977) of the different aspects of peer feedback task-design would therefore be valuable; this could provide more conclusive evidence with respect to what higher education teachers perceive they *could* and *would* change to the peer feedback tasks in their educational practice, and how these perceptions may vary between disciplines or institutes. For the six interviewed

teachers in chapter two, for example, somewhat differing patterns were found between teachers in research-intensive universities and universities of applied sciences. Possibly, these preliminary differences related to typical variations in teaching environments between these two – in the Dutch context commonly distinguished – types of universities.

A final methodological reflection concerns the absence of a no-feedback control-group in the empirical studies in this thesis. In general, studies including such a no-feedback control-group are scarce and therefore much needed to further our knowledge on *the extent to which* formative peer feedback impacts HE students' writing performance. Within the specific educational contexts of the empirical studies in this thesis, however, such control-groups were considered ethically unfeasible. Consequently, these empirical studies can only draw conclusions with regard to *relative* writing improvement for differently matched students or for students fulfilling different roles during peer feedback. The most salient objection to a no-feedback control-group that was encountered during informal conversations with both teaching staff and educational researchers, was that the omission of an existing beneficial learning element in a course is considered unethical. It seems that controlled interventions may be perceived as more acceptible in educational contexts where peer feedback was not implemented before, i.e. when peer feedback is an *additional* element instead of being *omitted as an existing* element within a course. Alternatively or complimentary, future studies into peer feedback may draw inspiration from research designs that are more common in other research domains, such as delayed treatment designs. Insofar as educational contexts would allow for such variations in research design, these may alleviate ethical concerns for quasi-experimental peer feedback studies with authentic academic writing tasks.

Practical implications and considerations

The findings of this thesis regarding the specific aspects of peer feedback task-design can be considered informative for both higher education teachers and educational advisors. Regarding the nature of the peer feedback, formative peer feedback had the most impact on students' writing performance if it included

both scores and comments. This suggests that teaching staff can optimize the impact of a peer feedback task by designing it accordingly, provided that they communicate the formative nature of the scores to their students upfront. Also, peer feedback containing explanations for accompanying value judgments or suggested revisions was particularly important for students to perceive the peer feedback as adequate. Hence, higher education teachers can optimize students' perceptions of adequacy by emphasizing the role of explanations in peer feedback. This can be achieved by systematically scaffolding the peer feedback that students provide on each other's work, for example through a combination of student training, task-instructions and rubrics to guide the peer feedback process (cf. Gielen and de Wever, 2015). In the long-term, this can positively influence students' beliefs about the importance of peer feedback as well as their confidence in themselves and their peers.

Regarding student ability matching, somewhat different conclusions were drawn in an on-campus and an online context. Irrespective of the reported significance values, however, effect sizes tended to be small in both contexts. In other words, if there is an effect of students' ability match on their subsequent increase in writing performance, this effect is probably small. This suggests that the ability-matching of higher education students should not be perceived as a top priority and that *if* ability-matching is contemplated when designing a peer feedback task, it should be weighed against the efforts required for implementation. In cases where time-constraints are salient, this renders random student matching an ethically defensible choice. When student matching becomes more or less effortless in the context of (web-based) platforms that facilitate the peer feedback process, student matching becomes a feasible option to attain these relatively small effects on students learning gains.

Regarding higher education students' feedback role, peer feedback providers and –receivers improved their academic writing to similar degrees. This knowledge should be comforting to higher education teaching staff in multiple ways. For example, when a student unintentionally does not receive feedback from a peer, the act of having provided peer feedback can still be expected to contribute to his or her own learning. This is especially relevant when group

sizes increase and, consequently, teachers' control over the implementation of the peer feedback process decreases. This knowledge is also helpful in situations where students are unfamiliar with the peer feedback process and/or are (still) skeptical with respect to their peers' feedback. In such situations, issues like students' distrust in each other's peer feedback quality may be circumvented by initially withholding the provided peer feedback from the intended receivers. As students' confidence in, and support for peer feedback as an instructional method increases with practice and experience, peer feedback may be made fully reciprocal by disclosing the peer feedback to the assessees.

Whenever the preparation of students for a future career is considered as a broad central aim of a higher education curriculum, peer feedback should be regarded as an important learning goal in itself (e.g., Cowan, 2010; Liu & Carless, 2006; Sadler, 2010; Sluijsmans et al., 2004). Students' support for peer feedback is pivotal with respect to their engagement in the peer feedback process and the learning gains that can be expected. It therefore seems particularly worthwhile to explore how to cultivate a classroom culture where peer feedback is the norm and to investigate how students' peer feedback beliefs and skills develop over time. Possibly, higher education teaching staff could set the norm early on by informing and preparing students with respect to the peer feedback process from the very beginning of their studies. Ideally, then, peer feedback would be formative, guided in terms of the nature of the peer feedback, and gradually expanded to more elaborate and complex tasks that students believe are within the range of their developing expertise. In this light, the current thesis facilitates evidence-based decisions with respect to the design aspects of peer feedback tasks, including the nature of the peer feedback, student matching, and students' feedback role. In addition, the Beliefs about Peer Feedback Questionnaire can be instrumental in the systematic, long-term monitoring of students' peer feedback beliefs and the aspects of task-design that influence those beliefs.

In the specific context of Leiden University, the findings of this thesis are also informative for educational advisors and policy makers. Educational advisors can incorporate these findings into their training of and advice to both beginning and senior teaching staff, helping them develop peer feedback

tasks that both positively affect students' performance and that are optimally supported by the students. Educational policy makers can weigh the findings and arguments in this thesis into their decisions that affect the teaching staff's capacity for designing peer feedback tasks. By stimulating teacher training programs and by supporting the availability of peer feedback software packages, for example, educational policy makers can facilitate higher education teachers in designing effective peer feedback tasks that help students to improve their academic writing skills.

References

References

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in Massive Open Online Courses. *The Electronic Journal of e-learning, 13*(4), 207-216.

Admiraal, W., Huisman, B., & van de Ven, M. (2014). Self- and peer assessment in Massive Open Online Courses. *International Journal of Higher Education, 3*(3), 119-128.

Agricola, B., Prins, F. J., van der Schaaf, M. F., & van Tartwijk, J. (2016). *Changes in students' self-regulated learning, feedback perception and Motivation during undergraduate research projects.* Paper presented at the European Association for Research on Learning and Instruction (EARLI), Munich, Germany.

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., & Fishbein, M. (2005). The influence of attitudes on behavior. In D. Albarracin, B. T. Johnson, & M. P. Zanna (Eds.), *The handbook of attitudes* (pp. 173-221). NJ: Lawrence Erlbaum Associates: Mahwah.

Bailey, R., & Garner, M. (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), 187-198. doi:10.1080/13562511003620019

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing Procedures for Implementing Peer Assessment in Large Classes Using an Action Research Process. *Assessment & Evaluation in Higher Education, 27*(5), 427-441. doi:10.1080/0260293022000009302

Biggs, J. (1996). Assessing Learning Quality: reconciling institutional, staff and educational demands. *Assessment & Evaluation in Higher Education, 21*(1), 5-16. doi:10.1080/0260293960210101

Birjandi, P., & Tamjid, N. H. (2012). The role of self-, peer and teacher assessment in promoting Iranian EFL learners' writing performance. *Assessment & Evaluation in Higher Education, 37*(5), 513-533. doi:10.1080/02602938.2010.549204

Bolzer, M., Strijbos, J. W., & Fischer, F. (2015). Inferring mindful cognitive-processing of peer-feedback via eye-tracking: role of feedback-characteristics, fixation-durations and transitions. *Journal of Computer Assisted Learning, 31*(5), 422-434. doi:10.1111/jcal.12091

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

Buckley, E., & Cowap, L. (2013). An evaluation of the use of Turnitin for electronic submission and marking and as a formative feedback tool from an educator's perspective. *British Journal of Educational Technology, 44*(4), 562-570. doi: 10.1111/bjet.12054

Cahyono, B. Y., & Amrina, R. (2016). Peer Feedback, Self-correction, and Writing Proficiency of Indonesian EFL Students. *Arab World English Journal, 7*(1), 178-193. doi:10.24093/awej/vol7no1.12

Chauhan, A. (2014). Massive Open Online Courses (MOOCs): Emerging trends in assessment and accreditation. *Digital Education Review, 25*, 7-18.

Chen, C. W.-Y. (2010). Graduate Students' Self-Reported Perspectives regarding Peer Feedback and Feedback from Writing Consultants. *Asia Pacific Education Review, 11*(2), 151-158. doi:10.1007/s12564-010-9081-5

Cheng, K. H., Liang, J. C., & Tsai, C. C. (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

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

Cho, Y. H., & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science, 39*(5), 629-643. doi:10.1007/s11251-010-9146-1

Cho, K., & MacArthur, C. (2010). Student revision with peer and expert reviewing. *Learning and Instruction, 20*(4), 328-338. doi:10.1016/j.learninstruc.2009.08.006

Cho, K., & MacArthur, C. (2011). Learning by reviewing. *Journal of Educational Psychology, 103*(1), 73-84. doi:10.1037/a0021950

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education, 48*(3), 409-426. doi:10.1016/j.compedu.2005.02.004

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing - Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260-294.

Cho, K., Schunn, C. D., & Wilson, R. W. (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

Ciftci, H., & Kocoglu, Z. (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 behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cortina, J. M. (1993). What Is Coefficient Alpha - an Examination of Theory and Applications. *Journal of Applied Psychology, 78*(1), 98-104. doi:10.1037/0021-9010.78.1.98

Cowan, J. (2010). Developing the ability for making evaluative judgements. *Teaching in Higher Education, 15*(3), 323-334. doi:10.1080/13562510903560036

Crossman, J. M., & Kite, S. L. (2012). Facilitating improved writing among students through directed peer review. *Active Learning in Higher Education, 13*(3), 219-229. doi:10.1177/1469787412452980

DeBoer, J., Ho, A. D., Stump, G. S., & Breslow, L. (2014). Changing "course": Reconceptualizing educational variables for massive open online courses. *Educational Researcher, 43*(2), 74-84.

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

Downes, S. (2008). Places to go: Connectivism & connected knowledge. *Innovate: Journal of Online Education, 5*(1), 1-6.

Doyle, W., & Ponder, G. A. (1977). The practicality ethic in teacher decision-making. *Interchange, 8*(3), 1-12. doi:10.1007/bf01189290

Ebben, M., & Murphy, J. S. (2014). Unpacking MOOC scholarly discourse: a review of nascent MOOC scholarship. [Article]. *Learning Media and Technology, 39*(3), 328-345.

Evans, B. J., Baker, R. B., & Dee, T. S. (2016). Persistence Patterns in Massive Open Online Courses (MOOCs). [Article]. *Journal of Higher Education, 87*(2), 206-242.

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

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287-322. doi:10.2307/1170785

Flower, L., Hayes, J. R., Carey, L., Schriver, K., & Stratman, J. (1986). Detection, diagnosis, and the strategies of revision + composition. *College Composition and Communication, 37*(1), 16-55. doi:10.2307/357381

Gielen, M., & de Wever, B. (2015). Structuring the peer assessment process: A multilevel approach for the impact on product improvement and peer feedback quality. *Journal of Computer Assisted Learning, 31*, 435-449.

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

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304-315. doi: http://dx.doi.org/10.1016/j.learninstruc.2009.08.007

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*(3), 445-476. 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), 211-217. doi:10.1177/0098628315587618

Groemping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software, 17*(1), 1-27. doi:10.18637/jss.v017.i01

Gunersel, A. B., Simpson, N. J., Aufderheide, K. J., & Wang, L. (2008). Effectiveness of Calibrated Peer Review[TM] for Improving Writing and Critical Thinking Skills in Biology Undergraduate Students. *Journal of the Scholarship of Teaching and Learning, 8*(2), 25-37. Retrieved from: https://josotl.indiana.edu/issue/view/154

Handley, K., Price, M., & Millar, J. (2011). Beyond 'doing time': investigating the concept of student engagement with feedback. *Oxford Review of Education, 37*(4), 543-560.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing Self- and Peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53-70. doi: 10.1080/07294360123776

Hartberg, Y., Gunersel, A. B., Simspon, N. J., & Balester, V. (2008). Development of Student Writing in Biochemistry Using Calibrated Peer Review. *Journal of the Scholarship of Teaching and Learning, 8*(1), 29-44. Retrieved from Retrieved from: https://josotl.indiana.edu/issue/view/153

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. doi:10.3102/003465430298487

Hayes, J. R., & Flower, L. S. (1987). On the structure of the writing process. *Topics in Language Disorders, 7*(4), 19-30. Retrieved from: http://journals.lww.com/topicsinlanguagedisorders/toc/1987/09000

Hu, G., & Lam, S. T. E. (2010). Issues of cultural appropriateness and pedagogical efficacy: Exploring peer review in a second language writing class. *Instructional Science, 38*(4), 371-394. doi:10.1007/s11251-008-9086-1

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure

analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101-110. doi:10.1111/bjet.12520

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development, 36*(7), 1433-1447. doi:10.1080/07294360.2017.1325854

Hulleman, C. S., Durik, A. M., Schweigert, S. B., & Harackiewicz, J. M. (2008). Task values, achievement goals, and interest: An integrative analysis. *Journal of Educational Psychology, 100*(2), 398-416. doi:10.1037/0022-0663.100.2.398

Kaufman, J. H., & Schunn, C. D. (2011). Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional Science, 39*(3), 387-406. doi:10.1007/s11251-010-9133-6

King, A. (1997). ASK to THINK-TEL WHY: A model of transactive peer tutoring for scaffolding higher level complex learning. *Educational Psychologist, 32*(4), 221-235. doi:10.1207/s15326985ep3204_3

Knight, P. T. (2002). Summative Assessment in Higher Education: Practices in disarray. *Studies in Higher Education, 27*(3), 275-286. doi:10.1080/03075070220000662

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Kop, R. (2011). The Challenges to Connectivist Learning on Open Online Networks: Learning Experiences during a Massive Open Online Course. *International Review of Research in Open and Distance Learning and Instruction, 12*(3).

Kovanović, V., Joksimović, S., Gašević, D., Siemens, G., & Hatala, M. (2015). What public media reveals about MOOCs: A systematic analysis of news reports. *British Journal of Educational Technology, 46*(3), 510-527.

Landis, J. R., & Koch, G. G. (1977). Measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. doi: 10.2307/2529310

Lawton, W., & Lunt, K. (2013). *Would you credit that? The trajectory of the MOOCs juggernaut*: The Observatory on Borderless Higher Education (OBHE). Retrieved from: http://www.obhe.ac.uk/documents/view_details?id=931.

Lazonder, A. W., & Harmsen, R. (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), 123-132. Retrieved from: http://www.tojet.net/articles/v14i1/14113.pdf

Li, L., & Steckelberg, A. L. (2004). *Using Peer Feedback to Enhance Student Meaningful Learning* (ED485111). Retrieved from https://eric.ed.gov/?id=ED485111

Lin, S. S. J., Liu, E. Z. F., & Yuan, S. M. (2001). Web-based peer assessment: feedback for students with various thinking-styles. *Journal of Computer Assisted Learning, 17*(4), 420-432. doi: 10.1046/j.0266-4909.2001.00198.x

Lindblom-Ylänne, S., Pihlajamäki, H., & Kotkas, T. (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., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*(12), 1181-1209. doi:10.1037/0003-066X.48.12.1181

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-analysis.* Thousand Oaks, Calif: SAGE Publications, Inc.

Liu, N., & Carless, D. (2006). Peer feedback: the learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290. doi:10.1080/13562510600680582

Lizzio, A., & Wilson, K. (2008). Feedback on assessment: students' perceptions of quality and effectiveness. *Assessment & Evaluation in Higher Education, 33*(3), 263-275. doi:10.1080/02602930701292548

Ludemann, P. M., & McMakin, D. (2014). Perceived helpfulness of peer editing activities: First-year students' views and writing performance outcomes. *Psychology Learning & Teaching, 13*(2), 129-136. doi:10.2304/plat.2014.13.2.129

Lundstrom, K., & Baker, W. (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

Luxton-Reilly, A. (2009). A systematic review of tools that support peer assessment. *Computer Science Education, 19*(4), 209-232. doi: 10.1080/08993400903384844

Mawlawi Diab, N. (2011). Assessing the relationship between different types of student feedback and the quality of revised writing. *Assessing Writing, 16*(4), 274-292. doi:10.1016/j.asw.2011.08.001

McCarthy, J. (2017). Enhancing feedback in higher education: Students' attitudes towards online and in-class formative assessment feedback models. *Active Learning in Higher Education, 18*(2), 127-141. doi:10.1177/1469787417707615

McConlogue, T. (2015). Making judgements: investigating the process of composing and receiving peer feedback. *Studies in Higher Education, 40*(9), 1495-1506. doi:10.1080/03075079.2013.868878

McDowell, L. (2012). Assessment for learning. In L. Clouder, C. Broughan, S. Jewell, & G. Steventon (Eds.), *Improving student engagement and development through assessment: Theory and practice in higher education.* Abingdon, UK: Routledge.

McGarr, O., & Clifford, A. M. (2013). 'Just enough to make you take it seriously': exploring students' attitudes towards peer assessment. *Higher Education, 65*(6), 677-693. doi:10.1007/s10734-012-9570-z

Mostert, M., & Snowball, J. D. (2013). Where angels fear to tread: online peer-assessment in a large first-year class. *Assessment & Evaluation in Higher Education, 38*(6), 674-686. doi:10.1080/02602938.2012.683770

Mulder, R. A., Pearce, J. M., & Baik, C. (2014). Peer review in higher education: Student perceptions before and after participation. *Active Learning in Higher Education, 15*(2), 157-171. doi:10.1177/1469787414527391

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. van Merrienboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125-143). Mahwah, NJ: Erlbaum.

Nelson, M. M., & Schunn, C. D. (2009). The nature of feedback: How different types of peer

feedback affect writing performance. *Instructional Science, 37*(4), 375-401. doi: 10.1007/s11251-008-9053-x

Nicol, D. J., & Macfarlane-Dick, D. (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., Thomson, A., & Breslin, C. (2014). Rethinking feedback practices in higher education: a peer review perspective. *Assessment & Evaluation in Higher Education, 39*(1), 102-122. doi:10.1080/02602938.2013.795518

Noroozi, O., Biemans, H., & Mulder, M. (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.* Retrieved from OECD Publishing, Paris: http://www.oecd.org/edu/education-at-a-glance-19991487.htm

Pajares, M. F. (1992). Teachers Beliefs and Educational-Research - Cleaning up a Messy Construct. *Review of Educational Research, 62*(3), 307-332. doi:10.3102/00346543062003307

Palmer, B., & Major, C. H. (2008). Using Reciprocal Peer Review to Help Graduate Students Develop Scholarly Writing Skills. *Journal of Faculty Development, 22*(3), 163-169.

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

Patchan, M. M., & Schunn, C. D. (2015). Understanding the benefits of providing peer feedback: How students respond to peers' texts of varying quality. *Instructional Science.* doi:10.1007/s11251-015-9353-x

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing model fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science, 19*(1), 12-21. doi:10.1080/1091367X.2014.952370

R Core team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.r-project.org

Raffaghelli, J. E., Cucchiara, S., & Persico, D. (2015). Methodological approaches in MOOC research: Retracing the myth of Proteus. *British Journal of Educational Technology, 46*(3), 488-509.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, J. C. Valentine, H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis, 2nd ed.* (pp. 295-315). New York, NY, US: Russell Sage Foundation.

Reich, J., Tingley, D., Leder-Luis, J., Roberts, M. E., & Stewart, B. M. (2015). Computer-assisted reading and discovery for student-generated text in massive open online courses. *Journal of Learning Analytics, 2*(1), 156-184.

Rhoads, R. A., Sayil Camacho, M., Toven-Lindsey, B., & Berdan Lozano, J. (2015). The massive open online course movement, xMOOCs, and faculty labor. *The Review of Higher Education, 38*(3), 397-424.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical

Software, 48(2), 36. doi:10.18637/jss.v048.i02

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional-Systems. *Instructional Science, 18*(2), 119-144. doi:10.1007/Bf00117714

Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education, 35*(5), 535-550. doi:10.1080/02602930903541015

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research, 8*(1), 31-54. doi:10.1191/1362168804lr133oa

Sampson, V., & Walker, J. P. (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), 1443-1485. 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), 238-255. doi:10.1109/47.44536

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. doi:10.3102/0034654307313795

Sluijsmans, D. M. A. (2002). *Student involvement in assessment: The training of peer assessment skills.* (Doctoral dissertation), Open University, Heerlen, The Netherlands. Retrieved from http://dspace.ou.nl/bitstream/1820/1034/1/dissertation%20Sluijsmans%20202002.pdf

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

Snowball, J. D., & Mostert, M. (2013). Dancing with the devil: formative peer assessment and academic performance. *Higher Education Research & Development, 32*(4), 646-659. doi: 10.1080/07294360.2012.705262

SoFoKleS. (2016). An exploration within in tertiary scientific education: The pressure to perform for scientific staff. [Verkenning in het WO: Prestatiedruk onder wetenschappelijk personeel]. Retrieved from https://www.sofokles.nl/wp-content/uploads/rapport_verkenning_prestatiedruk.pdf

Stellmack, M. A., Keenan, N. K., Sandidge, R. R., Sippl, A. L., & Konheim-Kalkstein, Y. L. (2012). Review, revise, and resubmit: The effects of self-critique, peer review, and instructor feedback on student writing. *Teaching of Psychology, 39*(4), 235-244. doi:10.1177/0098628312456589

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

Strijbos, J. W., Narciss, S., & Dünnebier, K. (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

Taskforce Study Success (2009). *Study success [Studiesucces].* Retrieved from Leiden University, Leiden, The Netherlands: http://media.leidenuniv.nl/legacy/rapport-eindadvies-studiesucces.pdf

Terras, M. M., & Ramsay, J. (2015). Massive open online courses (MOOCs): Insights and challenges from a psychological perspective. *British Journal of Educational Technology, 46*(3), 472-487.

Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249-276. doi:10.3102/00346543068003249

Topping, K. J. (2009). Peer Assessment. *Theory into Practice, 48*(1), 20-27. doi:10.1080/00405840802577569

Topping, K. J. (2010). Methodological quandaries in studying process and outcomes in peer assessment. *Learning and Instruction, 20*(4), 339-343. doi:10.1016/j.learninstruc.2009.08.003

Toven-Lindsey, B., Rhoads, R. A., & Berdan Lozano, J. (2015). Virtually unlimited classrooms: Pedagogical practices in massive open online courses. *Internet and Higher Education, 24*, 1-12.

Tsai, Y.-C., & Chuang, M.-T. (2013). Fostering revision of argumentative writing through structured peer assessment. *Perceptual and Motor Skills, 116*(1), 210-221. doi:10.2466/10.23.PMS.116.1.210-221

Tsui, A. B. M., & Ng, M. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing, 9*(2), 147-170. doi:10.1016/S1060-3743(00)00022-9

van den Berg, I., Admiraal, W., & Pilot, A. (2006a). Design principles and outcomes of peer assessment in higher education. *Studies in Higher Education, 31*(3), 341-356.

van den Berg, I., Admiraal, W., & Pilot, A. (2006b). Peer assessment in university teaching: evaluating seven course designs. *Assessment & Evaluation in Higher Education, 31*(1), 19-36.

van den Berg, I., Admiraal, W., & Pilot, A. (2006c). Designing student peer assessment in higher education: analysis of written and oral peer feedback. *Teaching in Higher Education, 11*(2), 135-147. doi: 10.1080/13562510500527685

van der Pol, J., van den Berg, B. A. M., Admiraal, W. F., & Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers & Education, 51*(4), 1804-1817. doi:10.1016/j.compedu.2008.06.001

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

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

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

VAWO. (2015). *Factsheet in response to Vision on Science 2025* [Factsheet naar aanleiding van Wetenschapsvisie 2025]. Retrieved from https://www.vawo.nl/wp-content/uploads/2015/02/Factsheet-nav-Wetenschapsvisie-2025.pdf https://www.vawo.nl/wp-content/uploads/2015/02/Factsheet-nav-Wetenschapsvisie-2025.pdf

VAWO. (2015). *Factsheet student-staff ratio.* Retrieved from https://www.vawo.nl/wp-content/uploads/2015/04/VAWO-factsheet-student_staf_ratio.pdf

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *2010, 36*(3), 48. doi:10.18637/jss.v036.i03

Visser-Wijnveen, G. J., van der Rijst, R. M., & van Driel, J. H. (2016). A questionnaire to capture

students' perceptions of research integration in their courses. *Higher Education, 71*(4), 473-488. doi:10.1007/s10734-015-9918-2

Walker, J. P., & Sampson, V. (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), 1269-1274. doi:10.1021/ed300656p

Wen, M. L., & Tsai, C. C. (2006). University students' perceptions of and attitudes toward (online) peer assessment. *Higher Education, 51*(1), 27-44. doi:10.1007/s10734-004-6375-8

Wen, M. L., Tsai, C. C., & Chang, C. Y. (2006). Attitudes towards peer assessment: a comparison of the perspectives of pre-service and in-service teachers. *Innovations in Education & Teaching International, 43*(1), 83-92. doi:10.1080/14703290500467640

Wigfield, A., & Eccles, J. S. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68-81. doi:10.1006/ceps.1999.1015

Wingate, U. (2010). The impact of formative feedback on the development of academic writing. *Assessment & Evaluation in Higher Education, 35*(5), 519-533. doi:10.1080/02602930903512909

Wong, H., & Storey, P. (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), 663-675. doi:10.1348/000709905X52210

Xiao, Y., & Lucking, R. (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), 186-193. doi:10.1016/j.iheduc.2008.06.005

Yang, M., Badger, R., & Yu, Z. (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., & Meng, W. T. (2013). The effects of online feedback training on students' text revision. *Language Learning & Technology, 17*(2), 220-238. Retrieved from: http://llt.msu.edu/issues/june2013/index.html

Yoshizawa, S., Terano, T., & Yoshikawa, A. (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

Summary

In 2008, Leiden University initiated the Taskforce Study Success, whose primary aim was to provide recommendations for increasing study success and decreasing student attrition across bachelor programs. Within the context of the Taskforce's (2009) broader set of recommendations, this thesis focused on the design of formative peer feedback on academic writing tasks and the effects thereof on students' performance. In particular, this thesis investigated to what extent formative peer feedback impacts higher education students' academic writing performance and how particular aspects of peer feedback task-design affect this performance. This thesis aimed to combine both theoretical and practical significance. To advance scientific knowledge, a quantitative focus on students' academic writing performance was combined with relatively well-controlled research designs in authentic educational contexts. To be of practical value for higher education teaching staff, this thesis focused on aspects of peer feedback task-design that were perceived as relatively controllable for higher education teachers.

Five studies were conducted. In chapter two, a meta-analysis was reported to assess the impact of formative peer feedback on higher education students' academic writing performance. Results indicated that the effects of formative peer feedback on academic writing performance tend to be larger than that of either no feedback or self-assessment, whereas the effects of peer- and teacher feedback appeared to be similar. In addition, two moderator analyses were conducted to investigate the role of two controllable aspects of peer feedback task-design: the nature of peer feedback and the number of peers engaged with during peer feedback. The results of these analyses indicated that a combination of both peer comments and –scores tends to have a larger effect on writing performance than either peer comments or –scores alone, whereas the number of peers with whom a student engaged during peer feedback did not moderate subsequent writing performance. These results suggest that higher education teaching staff can be confident that peer feedback positively affects their students'

learning gains, and indicate that they should design their peer feedback tasks in such a way that students provide both comments and scores to each other's writing. A notable limitation of this study concerned the small ($N = 24$) number of studies that proved eligible for inclusion. It was argued that this signals a need for more well-controlled, (quasi-)experimental studies.

In chapters three and four, two empirical studies were reported that focused on the effects of students' ability-match. Chapter three described a quasi-experimental study in which 94 students were anonymously matched into either same-ability (homogeneous) or different-ability (heterogeneous) dyads. Dyad composition appeared unrelated to the nature of the peer feedback or subsequent improvements in writing performance, although a trend was found suggesting that high-ability dyads focus more on content-related issues. Also, relatively high- or low ability authors did not differ in how they benefitted from peer feedback on aspects of essay content-, structure-, or writing style. If confirmed by future studies, these results suggest that higher education teaching staff should not worry too much about students' ability match during peer feedback on writing tasks. However, ability differences in this sample could have reflected between-student differences within a sample that – overall – was relatively similar in this respect. As a consequence, the reported effects of ability matching may reflect conservative estimates and therefore may be more profound in situations where ability differences are larger. Therefore, chapter four explored the effects of ability matching on writing performance in the context of a massive open online course (MOOC). A total of 565 participants were categorized as highest (42%), intermediate (45%) or lowest (13%) performers based on available performance metrics prior to the first of two essay assignments. Post hoc analyses were conducted to explore the relation between participants' own performance level, that of their reviewing peers', and participants' writing performance increase between the first and second essay. Overall, peer assessor's average performance level positively related to participants' increase in writing performance. More specifically, peer assessors' average performance level only related to the writing performance of the intermediate and higher performing participants, and not to that of the lowest performing participants. Effect sizes were small,

however. Different explanations were considered conceivable, including that performance level relates to participants' ability to utilize the received peer feedback or that performance level relates to variations in the peer feedback quality that assessors provide. Given the exploratory nature of this study, these findings and explanations are to be tested by future studies that, among others, include more background info on participants such as educational background and individual learning goals.

Chapter five reported on an empirical study that compared the impact of providing versus receiving peer feedback on students' academic writing performance. In addition, this study investigated how the nature of the peer feedback that students received related to their peer feedback perceptions and their writing performance. Results indicated that peer feedback providers and –receivers improve their writing to a similar extent, and that explanatory peer feedback comments are most influential with respect to students' perceptions of peer feedback adequacy. However, no direct relations were found between students' perceptions of the received peer feedback and their subsequent increase in writing performance. These results were believed to elucidate two findings that are informative for higher education researchers and –teaching staff. The finding that both providing and receiving peer feedback positively affect students' writing performance provides higher education teaching staff with a degree of flexibility in designing peer feedback tasks. For example, initially withholding received peer feedback could avoid issues such as students' distrust of their peers' feedback quality, while the exercise of providing peer feedback would still be likely to be beneficial to students' writing performance. Also, the finding that explanatory peer feedback most strongly related to students' perceptions of adequacy was argued to be important with respect to students' more general support for – and engagement in – the peer feedback process. This highlights the importance of emphasizing the role of explanations in peer feedback training and instruction.

Chapter six reported on the development of a questionnaire to assess students' peer feedback beliefs. This Beliefs about Peer Feedback Questionnaire (BPFQ) served a dual aim. For one, prior research into students' peer feedback

beliefs has adopted different approaches addressing a variety of themes. Therefore, the first central aim of the BPFQ was to contribute to the alignment and, consequently, comparability of research findings. Simultaneously, it aimed to provide a practical instrument for higher education teaching staff to monitor how their teaching practice influences students' peer feedback beliefs. Based on the variety of themes addressed in the literature, four scales were conceptualized: (1) students' valuation of peer feedback as an instructional method, (2) students' confidence in the quality of the peer feedback they provide to a peer, (3) students' confidence in the quality of the peer feedback they receive from a peer and (4) the extent to which students regard peer feedback as an important skill. These four scales, totaling ten items, were validated in a separate exploratory and confirmatory study, with scale reliabilities ranging between $\alpha = .67$ and $\alpha = .82$. Consequently, the BPFQ was considered a reliable, comprehensive instrument to assess students' peer feedback beliefs. The concise nature of the BPFQ was argued to make it an applicable instrument for both higher education teachers who want to conduct research within their own teaching practice as for researchers aiming to monitor the development of students' peer feedback beliefs over time.

In conclusion, the current thesis furthers our knowledge on a) the available evidence for the impact of formative peer feedback on writing performance, b) how students' ability match and feedback role as either peer feedback provider or –receiver relate to writing performance, and c) the relations between the nature of the peer feedback and students' perceptions thereof. Provided that future research confirms these findings, several implications follow. Formative peer feedback positively affects higher education students' academic writing performance, which it does to a similar extent as teacher feedback. This implies that higher education teaching staff can be confident that peer feedback contributes to students' writing performance. Students, however, could be skeptical to this notion as their peers may differ from their teaching staff in terms of domain-specific knowledge. Peer feedback may indeed have a different focus compared to teacher feedback. This thesis reported a trend where high-ability dyads focus more on content-related issues. This may imply that peer feedback

should as a complimentary feedback source to teacher feedback during the first years of higher education programs, and that it could be regarded as increasingly comparable to teacher feedback as students acquire more domain-specific knowledge and experience. The reported nature of first-year students' peer feedback also implies that they should be trained and guided in providing peer feedback that includes sufficient explanatory comments.

Some caution is in place as a result of the methodological choices that were made in this thesis. For example, the focus on writing performance assumed that increasing grades on a writing task reflect the increase of a student's writing skills. Also, what teachers perceived as controllable was aligned with planned behavior theory, referring to their perceptions of the extent to which they *could* adapt these variables. However, controllability did not directly incorporate the extent to which the teachers *would* adapt these design-aspects of peer feedback tasks. More elaborate research into the practicality of the different aspects of peer feedback task-design would therefore be valuable. A final methodological reflection concerns the absence of a no-feedback control group in the empirical studies of this thesis. A direct consequence is that this thesis can only draw conclusions with regard to the *relative* writing improvement for differently matched students or for students in different feedback roles.

In terms of practical implications and considerations, higher education teachers can optimize the impact of peer feedback by having students provide both formative comments and –scores to each other's written products. Also, higher education teachers can optimize students' perceptions of adequacy by emphasizing the role of explanations in peer feedback, as this positively influences students' beliefs about the importance of peer feedback as well as their confidence in themselves and their peers in the long run. Given the small effect sizes, matching students based on prior performance generally only appears to become viable when student-matching is relatively effortless as a result of automatization through ICT. Finally, providing and receiving peer feedback contributed similarly to students' writing performance. Knowing this, higher education teaching staff can provide students with opportunities to get used to the peer feedback process. For example, received peer comments

may be withheld during students' first experiences with peer feedback, thereby avoiding students' initial skepticism associated with, or triggered by critical peer feedback.

Whenever the preparation of students for a future career is considered as a broad central aim of a higher education curriculum, peer feedback should be regarded as an important learning goal in itself. Students' support for peer feedback is pivotal with respect to their engagement in the peer feedback process, and the learning gains that can be expected thereof. Hence, it seems particularly worthwhile to explore how to cultivate a classroom culture where peer feedback is the norm and to investigate how students' peer feedback beliefs and skills develop over time. In this light, the current thesis supports higher education teaching staff in optimally designing peer feedback tasks. Additionally, the Beliefs about Peer Feedback Questionnaire can be instrumental in the systematic, long-term monitoring of students' peer feedback beliefs and the aspects of task-design that influence those beliefs. In the specific context of Leiden University, the findings and arguments presented in this thesis can help educational advisors and teaching staff in developing effective peer feedback tasks that are optimally supported by the students.

Nederlandse samenvatting

In 2008 initieerde de Universiteit Leiden de Taskforce Studiesucces. Het primaire doel van deze Taskforce was het doen van aanbevelingen om studiesucces te verhogen en uitval te verminderen in de bachelor fase van de opleidingen. De Taskforce rapporteerde in 2009 een breed palet aan aanbevelingen. Deze betroffen onderwerpen als toetsbeleid en studentbetrokkenheid, waarbij actieve participatie, feedback en de vaardigheden voor het schrijven van een bachelor scriptie belangrijke aspecten waren. In de context van deze aanbevelingen focust dit proefschrift op formatieve peer feedback bij academisch schrijfopdrachten in het hoger onderwijs (HO). Meer specifiek wordt onderzocht in hoeverre formatieve peer feedback de schrijfprestaties van studenten in het hoger onderwijs verbeterd en hoe het ontwerp van peer feedback taken deze schrijfprestaties beïnvloedt. Een belangrijk doel hierbij is zowel een theoretische bijdrage te leveren alsook van praktische waarde te zijn. Met het oog op de bijdrage aan de huidige wetenschappelijke kennisbasis combineren de studies in dit proefschrift veelal een kwantitatieve focus op schrijfprestaties met relatief goed gecontroleerde onderzoeksopzetten binnen authentieke onderwijscontexten. Met het oog op de praktische waarde voor professionals in het hoger onderwijs focust dit proefschrift zoveel mogelijk op de ontwerp-aspecten van peer feedback taken die door docenten in het hoger onderwijs als controleerbaar percipiëren.

In totaal zijn vijf studies uitgevoerd. Hoofdstuk twee rapporteert een meta-analyse welke de impact van formatieve peer feedback op de schrijfprestaties van HO studenten onderzoekt. De resultaten van deze studie suggereren dat het effect van formatieve peer feedback op schrijfprestaties groter is ten opzichte van zelfbeoordeling alsook ten opzichte van een controle-conditie (helemaal geen feedback). Qua effect op schrijfprestaties blijken peer feedback en feedback vanuit de docent echter niet te verschillen. Tevens zijn twee aanvullende analyses uitgevoerd om de effecten te onderzoeken van twee relatief controleerbare aspecten m.b.t. het ontwerp van peer feedback taken:

de aard van de peer feedback en het aantal medestudenten (peers) waarmee een student interacteert. De resultaten van deze analyses wijzen erop dat peer feedback als *een combinatie* van schriftelijke scores en –opmerkingen een grotere impact heeft op de schrijfprestaties van studenten dan alleen scores *of* alleen opmerkingen. Het aantal peers waarmee een student interacteert blijkt niet significant gerelateerd aan daaropvolgende schrijfprestaties. Deze resultaten suggereren dat HO docenten erop kunnen vertrouwen dat peer feedback studenten helpt hun schrijfprestaties te verbeteren en dat zij peer feedback opdrachten zodanig moeten ontwerpen dat peer feedback wordt gegeven als een combinatie van scores en opmerkingen. Het beperkte aantal studies dat voor inclusie in aanmerking kwam (*N* = 24) maakt dat deze conclusies echter met enige voorzichtigheid dienen te worden geïnterpreteerd. Tegelijkertijd reflecteert dit beperkte aantal geïncludeerde studies *an sich* het beperkte aantal goed gecontroleerde studies in de literatuur, wat de noodzaak tot meer van dergelijke studies naar de relatie tussen peer feedback en schrijfprestaties onderschrijft.

Hoofdstukken drie en vier rapporteren empirische studies waarbij de match tussen studenten bij peer feedback op essays centraal staat. Hoofdstuk drie beschrijft een quasi-experimentele studie waarin 94 bachelor studenten Pedagogische Wetenschappen tijdens de peer feedback fase anoniem zijn gematched met een medestudent van gelijke schrijfvaardigheid (homogene koppels) of met een medestudent van meer of mindere schrijfvaardig (heterogene koppels). De wijze waarop studenten zijn gematched tijdens de peer feedback fase blijkt noch gerelateerd aan de aard van de peer feedback, noch aan de mate waarin zij hun essay verbeteren. Echter, een niet significante trend lijkt te suggereren dat homogene koppels met daarin twee schrijfvaardige studenten meer focussen op inhoud-gerelateerde criteria van de schrijfopdracht dan anders gematchte koppels. Verder blijken meer en minder schrijfvaardige studenten niet te verschillen in de mate waarin zij profiteren van peer feedback op de inhoud, structuur, of schrijfstijl van hun essay. Indien vervolgonderzoeken deze bevindingen bevestigen, dan suggereert dat dat HO docenten zich niet veel bezig hoeven te houden met hoe hun studenten zijn gematched tijdens de peer

feedback fase bij schrijfopdrachten. Echter, de onderzochte populatie studenten als geheel zou – gelet op hun doorgaans gelijke educatieve achtergrond – relatief homogeen kunnen zijn in hun schrijfprestaties. Het gevolg daarvan kan zijn dat de gerapporteerde effecten van student-matching conservatieve schattingen zijn, en dat dergelijke effecten derhalve nadrukkelijker tot uiting komen naarmate de verschillen in schrijfvaardigheid tussen studenten groter worden. Hoofdstuk vier verkent daarom het effect van student-matching in een 'massive open online course' (MOOC). In totaal worden, op basis van de beschikbare prestatie indicatoren vóór de eerste essay opdracht, 565 deelnemers gecategoriseerd als hoogst presterend (42%), laagst presterend (13%) of daar tussenin (45%). Middels post hoc analyses is vervolgens de relatie onderzocht tussen het prestatieniveau van de deelnemers, het gemiddelde prestatieniveau van de groep 'peer reviewers' van wie zij feedback krijgen, en de mate waarin de deelnemers zich verbeteren van de eerste essay-opdracht naar de tweede. Over het geheel genomen blijkt het gemiddelde prestatieniveau van de groep peer reviewers tijdens de eerste essay opdracht positief gerelateerd aan de score voor de daaropvolgende essay opdracht. Een meer gedetailleerde blik wijst echter uit dat dit niet het geval is voor de laagst presterende deelnemers. Van de gevonden relaties zijn de effecten echter klein. Ook is deze studie nadrukkelijk exploratief van aard. Met het oog op toekomstige studies op het gebied van student-matching in MOOCs zou het waardevol zijn om meer informatie over de deelnemers (zoals opleidingsniveau of leerdoelen) mee te kunnen nemen in de analyses.

Hoofdstuk vijf rapporteert een empirische studie waarin de impact van peer feedback *geven* versus *ontvangen* op de schrijfprestaties van HO studenten is onderzocht. Ook werd onderzocht hoe de aard van de ontvangen peer feedback door de studenten wordt gepercipieerd en hoe deze aard van de peer feedback relateert aan hun schrijfprestaties. De resultaten wijzen er op dat peer feedback geven en peer feedback ontvangen in gelijke mate bijdragen aan de schrijfprestaties van studenten, en dat studenten de ontvangen peer feedback voornamelijk als adequaat percipiëren wanneer deze verklaringen bevat (bijvoorbeeld voor gegeven waardeoordelen of suggesties voor revisie). Er is

echter geen directe relatie gevonden de peer feedback percepties van studenten en hun daaropvolgende schrijfprestaties. Deze resultaten leiden tot tenminste twee conclusies welke relevant worden geacht voor HO docenten en onderzoekers. Ten eerste verschaft de conclusie dat zowel peer feedback geven en ontvangen positief relateren aan de schrijfprestaties een bepaalde mate van flexibiliteit aan HO docenten bij het ontwerpen van peer feedback taken. Wanneer studenten (nog) weinig vertrouwen hebben in de feedback van hun medestudenten, zoals het geval zou kunnen zijn bij – of in aanloop naar – een eerste ervaring met peer feedback, zou een docent de ontvangen peer feedback in eerste instantie achter kunnen houden. Er mag dan immers alsnog een leereffect worden verwacht van het *geven* van peer feedback. Ten tweede is de conclusie dat de aanwezigheid van verklaringen, die bijvoorbeeld een waardeoordeel beargumenteren of een suggestie voor revisie toelichten, belangrijk is voor de mate waarin studenten de peer feedback percipiëren als adequaat. Dit is van belang met het oog op het draagvlak onder studenten voor het peer feedback proces. Immers, dit benadrukt de nadrukkelijke rol die verklaringen zouden moeten hebben in peer feedback training en instructie.

Hoofdstuk zes beschrijft de ontwikkeling van een vragenlijst om de opvattingen van studenten over peer feedback te onderzoeken. Deze *Opvattingen over Peer Feedback Vragenlijst* ('Beliefs about Peer Feedback Questionnaire'; BFPQ) heeft twee centrale doelen; enerzijds om een lijn te brengen in de versnipperde wijze waarop dergelijke opvattingen in eerdere studies worden onderzocht en zo bij te dragen aan de vergelijkbaarheid van toekomstige bevindingen, anderzijds om een praktisch (beknopt) instrument te verschaffen aan HO docenten waarmee zij kunnen monitoren wat de invloed van hun onderwijspraktijk is op de opvattingen van hun studenten met betrekking tot peer feedback. Op basis van de thema's die in de literatuur over de opvattingen van studenten over peer feedback naar voren komen zijn vier schalen geconceptualiseerd: (1) de waardering van studenten van peer feedback, (2) het vertrouwen van studenten in de kwaliteit van de peer feedback die zij geven aan andere studenten, (3) het vertrouwen van studenten in de kwaliteit van de peer feedback die zij ontvangen van andere studenten en (4) de mate waarin studenten peer feedback als

een belangrijke vaardigheid beschouwen. Deze vier schalen (10 items) zijn gevalideerd in een aparte exploratieve studie en een confirmatieve studie, waarbij de betrouwbaarheden van de separate schalen varieert tussen de $\alpha = .67$ en $\alpha = .82$. Door de beknopte aard van de BPFQ wordt deze beschouwd als een praktisch toepasbaar instrument voor zowel HO docenten als –onderzoekers.

Samengevat draagt dit proefschrift bij aan onze kennis over a) het beschikbare bewijs voor de leereffecten die peer feedback kan sorteren bij academisch schrijfopdrachten, b) hoe de match tussen (relatief) schrijfvaardige en minder schrijfvaardige studenten relateert aan daaropvolgende schrijfprestaties, c) hoe verschillende aspecten van het peer feedback proces (geven versus ontvangen) bijdragen aan schrijfprestaties en d) de relaties tussen de aard van ontvangen peer feedback en hoe studenten deze percipiëren. Hieruit volgen een aantal implicaties. Om te beginnen draagt formatieve peer feedback bij aan de schrijfprestaties van HO studenten, en doet dat in vergelijkbare mate als feedback vanuit docenten. Dit impliceert dat HO docenten er vertrouwen in kunnen hebben dat peer feedback een leereffect sorteert. Echter, studenten kunnen hier sceptisch over zijn omdat studenten en docenten niet over dezelfde domein-specifieke expertise beschikken. Peer feedback zou inderdaad een andere focus kunnen hebben dan docent feedback. Hoofdstuk 3 rapporteert een trend waarbij de peer feedback bij studentenkoppels van twee relatief schrijfvaardige studenten meer inhoudelijk van aard is. Dit zou kunnen impliceren dat peer feedback in eerste instantie vooral als complementair aan docent feedback moet worden beschouwd, en dat het meer vergelijkbaar wordt met docent feedback naarmate studenten meer domein-specifieke kennis en ervaring opdoen. De aard van de feedback die eerstejaars studenten elkaar geven (zie hoofdstukken 3 en 5) impliceert daarnaast dat studenten getraind en begeleid moeten worden in het geven van peer feedback die voldoende verklaringen bevat voor de gegeven waardeoordelen en suggesties voor verbeteringen.

De methodologische keuzes gemaakt in dit proefschrift zijn uiteraard niet zonder beperkingen. Zo wordt door de focus op schrijfprestaties bijvoorbeeld de aanname gemaakt dat hogere cijfers voor een schrijfopdracht daadwerkelijk een hogere schrijfvaardigheid van studenten reflecteren. Ook worden percepties

van docenten met betrekking tot de controleerbaarheid van bepaalde aspecten van peer feedback taak-ontwerp benaderd in lijn met de *planned behavior theory*. Daardoor refereren deze docent-percepties aan de mate waarin zij deze aspecten denken te *kunnen* controleren. Gepercipieerde controleerbaarheid omvat dus niet de mate waarin de docenten deze aspecten *zouden willen* controleren. Uitgebreider onderzoek naar een combinatie van gepercipieerde controleerbaarheid, praktische haalbaarheid en doelen bij HO docenten zou derhalve waardevolle toevoeging kunnen zijn aan de huidige literatuur. Een laatste methodologische reflectie betreft de afwezigheid van een geen-feedback controlegroep in de empirische studies. Dientengevolge kan dit proefschrift slechts conclusies trekken over de *relatieve* toename in schrijfprestaties voor studenten die op een bepaalde wijze zijn gematcht of een bepaalde feedback rol vervullen.

Een praktische implicatie is dat HO docenten de impact van formatieve peer feedback kunnen optimaliseren door studenten zowel scores als opmerkingen aan elkaar te laten geven. Ook kunnen docenten stimuleren dat de peer feedback als adequaat wordt ervaren door het belang van verklarende peer feedback te benadrukken. Het lijkt geen onwaarschijnlijke veronderstelling dat het frequent ontvangen van adequate peer feedback in positieve zin bijdraagt aan de opvattingen die studenten hebben over peer feedback en daarmee hun draagvlak voor peer feedback als geheel. Verder lijkt de beperkte grootte van de effecten, voor zover überhaupt aanwezig, het matchen van studenten alleen van praktisch voor docenten te zijn wanneer het matchen (nagenoeg) geen tijd en energie vergt qua organisatie. Dit zou bijvoorbeeld het geval kunnen zijn wanneer studenten automatisch gematcht kunnen worden binnen een digitale leeromgeving.

Elk hoger onderwijscurriculum dat als centraal leerdoel heeft om studenten voor te bereiden op een toekomstige carrière zou peer feedback moeten beschouwen als een belangrijke vaardigheid en als leerdoel op zichzelf. Daarbij is het draagvlak onder studenten voor peer feedback voorwaardelijk voor hun actieve betrokkenheid in het peer feedback proces en de daaruit voortvloeiende leereffecten. Derhalve lijkt het buitengewoon waardevol om binnen het HO

naar een cultuur te streven die dit draagvlak voor, en de actieve participatie in peer feedback faciliteert, alsook om te onderzoeken hoe peer feedback opvattingen en –vaardigheden zich ontwikkelen gedurende het curriculum. In dat licht ondersteund dit proefschrift HO docenten bij het optimaal ontwerpen van hun peer feedback taken. Daarbij kan de *Opvattingen over Peer Feedback Vragenlijst* waardevol zijn bij het systematisch (longitudinaal) monitoren van de opvattingen die studenten hebben met betrekking tot peer feedback. Samengevat zijn de bevindingen in dit proefschrift informatief voor docenten onderwijskundig adviseurs – zowel binnen het HBO als binnen universiteiten – om effectieve peer feedback taken te ontwerpen welke optimaal worden ondersteund en geaccepteerd door de studenten.

Dankwoord

Een promotietraject voelt als een hoofdstuk op zichzelf en behelst veel meer dan wat op deze pagina's gedrukt staat. Dank aan iedereen die hier inhoudelijk, praktisch en/of moreel aan heeft bijgedragen. Een aantal mensen noem ik graag in het bijzonder.

Nadira, Paul, Jan, dank voor jullie aangename en betrokken begeleiding. Nadira, bedankt voor alle brainstormsessies, positieve vibes, humor en support. En voor de nodige tegenspraak waar ik soms eigenwijs ben. Heel graag tot volgende kopjes koffie! Paul, dank voor je kritische reflecties en strategische blik, alsook voor je vaak snelle en uitgebreide feedback. Als ik twijfelde of vastliep gaven jouw heldere uitleg en creatieve oplossingen mij concrete handvatten en vertrouwen. Jan, bedankt voor je kritische vragen en no-nonsense mentaliteit. Door onze gesprekken kwam *the big picture* (weer) in zicht. De wereld van onderwijskundig onderzoek was bovendien nieuw voor mij en vanaf de eerste dag heeft jouw begeleiding mij geholpen hierin mijn weg te vinden.

Zonder de enthousiaste medewerking van, en gesprekken met docenten was dit onderzoek niet als zodanig tot stand gekomen. Marjo, Peter, Kim, Nicole, René, Beerend, Joanne, Memon, Tarik, Armin, dank jullie wel voor het delen van jullie kennis, ervaring en ideeën, en voor het meedenken in mogelijkheden. Benjamin, dank voor je hulp bij de analyses en het vertrouwen om te werken in R.

Het ICLON is een fantastische plek om te mogen werken. Everyone from the research group, thank you all for the constructive discussions, social support and *gezelligheid* over the years. Mayke en Indira, het was fijn om jullie als kamergenoten te hebben. Dank jullie wel voor het samen delen van vragen, uitdagingen, frustraties, overwinningen en feestelijkheden. Alle collega's bij HO en VO, bedankt voor de leerzame gesprekken en samenwerking. Tamara, Ton, Arthur, van onze samenwerking bij het verzorgen van onderwijs heb ik veel opgestoken. Ook wil ik graag Kirsten, Dianne en Gijs bedanken voor alle hulp bij het opmaken van de vragenlijsten.

Lieve vrienden en familie, bedankt voor jullie ondersteuning, afleiding en relativeringsvermogen. Pap en mam, Piya, bedankt voor de ruimte die jullie me

altijd hebben gegeven en jullie onvoorwaardelijke steun. Lieve Danitsja, Mare, in deze enerverende jaren waren jullie het rustpunt. Bedankt voor jullie liefde, geduld en energie. Jullie zijn mijn thuis.

Bart Huisman
Den Haag, 2018

About the author

Curriculum Vitae

Bart Huisman was born in The Hague on 7 October 1983. He grew up in The Hague, and finished his secondary education in Voorburg at College 't Loo in 2002. After some years of exploration and contemplation, Bart enrolled at Leiden University in 2006 to study Psychology. He obtained his bachelor degree in 2010, which he concluded with a thesis on the assessment of learning potential of primary school children. In 2013, Bart completed his research master Social & Organizational Psychology. The topic of his masters' thesis comprised the effects that different types of interventions have on negotiation outcomes during value conflicts and conflicts of interests. In September 2013, Bart started his PhD project at ICLON, Leiden University. Currently, Bart is working as a researcher at the Randstedelijke Rekenkamer.

Scientific publications

Huisman, B., Admiraal, W., Pilli, O., van de Ven, M., & Saab, N. (2018). Peer assessment in MOOCs: The relationship between peer reviewers' ability and authors' essay performance. *British Journal of Educational Technology, 49*(1), 101-110. doi:10.1111/bjet.12520

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2018). Peer feedback on academic writing: Undergraduate students' peer feedback role, peer feedback perceptions and essay performance. *Assessment & Evaluation in Higher Education (online first)*, 1-14. doi:10.1080/02602938.2018.1424318

Huisman, B., Saab, N., van Driel, J., & van den Broek, P. (2017). Peer feedback on college students' writing: Exploring the relation between students' ability match, feedback quality and essay performance. *Higher Education Research & Development, 36*(7), 1433-1447. doi:10.1080/07294360.2017.1325854

Admiraal, W., Huisman, B., & Pilli, O. (2015). Assessment in Massive Open Online Courses. *The Electronic Journal of e-learning, 13*(4), 207-216.

Admiraal, W., Huisman, B., & van de Ven, M. (2014). Self- and peer assessment in Massive Open Online Courses. *International Journal of Higher Education, 3*(3), 119-128.

Manuscripts under review/in preparation

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.

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2018). *A questionnaire to assess students' beliefs about peer feedback.* Manuscript submitted for publication.

Professional publications

Day, I. N. Z., Huisman, B. A., & Vereijken, M.W.C. (2017). Exploring freedom and control in global higher education: een conferentieverslag van de jaarlijkse bijeenkomst van de Society for Research into Higher Education [Exploring freedom and control in global higher education: A report of the annual meeting of the Society for Research into Higher Education], *Onderzoek van Onderwijs,46*, 19-21.

Conference contributions

Huisman, B., Saab, N., van Driel, J., & van den Broek (2017, April). *Peer feedback on writing: The relation between students' ability match, feedback quality, and essay performance.* Paper presented at the annual conference of the American Educational Research Association, 27 April – 1 May, San Antonio, Texas, USA.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2016, December). *The influence of peer feedback on higher education students' academic writing performance: A meta-analysis.* Poster presented at the annual conference of the Society for Research into Higher Education, 7-9 December, Newport, South Wales, United Kingdom.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2016, December). *The influence of peer feedback on higher education students' academic writing*

performance: A meta-analysis. Poster presented at the Newer Researcher Conference of the Society for Research into Higher Education, 6 December, Newport, South Wales, United Kingdom.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2016, July). *The influence of peer feedback on students' academic writing in higher education: A systematic review.* Paper presented at the Higher Education Conference, 13-15 July, Amsterdam, the Netherlands.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2015, November). *Student ability and –matching with peer feedback: Effects on performance with essay assignments.* Paper presented at the National Fall School of the Dutch Interuniversity Centre for Educational Research, 5-6 November, Utrecht, the Netherlands.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2015, August). *Matching students on ability: Effects on feedback quality and performance.* Round table presented at the biannual conference of the European Association for Research on Learning and Instruction, 25-29 August, Limassol, Cyprus.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2015, June). Peer feedback: Het matchen van studenten bij geschreven opdrachten [Peer feedback: matching students with writing assignments]. Paper presented in W. van der Linden (chair), *Het effect van studentverschillen in leerpotentieel op prestaties in en beleving van het hoger onderwijs [The effect of learning potential differences on students' performance in, and experience of higher education].* Symposium conducted at the Onderwijs Research Dagen, 17-19 June, Leiden, the Netherlands.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2014, November). *Peer feedback on written assignments: Investigating the effect of different ability matching constellations on performance.* Round table presentation at the International Fall School of the Dutch Interuniversity Centre for Educational Research, 10-14 November, Blankenberge, Belgium.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2015, August). *Formative feedback in interactive lectures: The effect of web-based interaction on lecturers' feedback, and students' self-efficacy, engagement and performance.* Research

design forum presentation at the SIG4 & SIG17 conference of the European Association for Research on Learning and Instruction, 20-22 August, Leuven, Belgium.

Huisman, B., Saab, N., van Driel, J., & van den Broek (2014, June). *Het effect van web-based interactie op de feedback van docenten, en de self-efficacy, engagement en prestaties van studenten [The effect of web-based interaction on lecturers' feedback, and students' self-efficacy, engagement and performance].* Poster presented at the Onderwijs Research Dagen, 11-13 June, Groningen, the Netherlands.

Leiden University Graduate School of Teaching

PhD dissertation series

Hoeflaak, A. (1994). *Decoderen en interpreteren: een onderzoek naar het gebruik van strategieën bij het beluisteren van Franse nieuwsteksten.*

Verhoeven, P. (1997). *Tekstbegrip in het onderwijs klassieke talen.*

Meijer, P.C. (1999). *Teachers' practical knowledge: Teaching reading comprehension in secondary education.*

Zanting, A. (2001). *Mining the mentor's mind: The elicitation of mentor teachers' practical knowledge by prospective teachers.*

Uhlenbeck, A.M. (2002). *The development of an assessment procedure for beginning teachers of English as a foreign language.*

Oolbekkink-Marchand, H.W. (2006). *Teachers' perspectives on self-regulated learning: An exploratory study in secondary and university education.*

Henze-Rietveld, F.A. (2006). *Science teachers' knowledge development in the context of educational innovation.*

Mansvelder-Longayroux, D.D. (2006). *The learning portfolio as a tool for stimulating reflection by student teachers.*

Meirink, J.A. (2007). *Individual teacher learning in a context of collaboration in teams.*

Nijveldt, M.J. (2008). *Validity in teacher assessment: An exploration of the judgement processes of assessors.*

Bakker, M.E.J. (2008). *Design and evaluation of video portfolios: Reliability, generalizability, and validity of an authentic performance assessment for teachers.*

Oonk, W. (2009). *Theory-enriched practical knowledge in mathematics teacher education.*

Visser-Wijnveen, G.J. (2009). *The research-teaching nexus in the humanities:*

Variations among academics.

Van der Rijst, R.M. (2009). *The research-teaching nexus in the sciences: Scientific research dispositions and teaching practice.*

Platteel, T.L. (2010). *Knowledge development of secondary school L1 teachers on concept-context rich education in an action-research setting.*

Kessels, C.C. (2010). *The influence of induction programs on beginning teachers' well-being and professional development.*

Min-Leliveld, M.J. (2011). *Supporting medical teachers' learning: Redesigning a program using characteristics of effective instructional development.*

Dobber, M. (2011). *Collaboration in groups during teacher education.*

Wongsopawiro, D. (2012). *Examining science teachers pedagogical content knowledge in the context of a professional development program.*

Belo, N.A.H. (2013). *Engaging students in the study of physics: An investigation of physics teachers' belief systems about teaching and learning physics.*

De Jong, R.J. (2013). *Student teachers' practical knowledge, discipline strategies, and the teacher-class relationship.*

Verberg, C.P.M. (2013). *The characteristics of a negotiated assessment procedure to promote teacher learning.*

Van Kan, C.A. (2013). *Teachers' interpretations of their classroom interactions in terms of their pupils' best interest: A perspective from continental European pedagogy.*

Dam, M. (2014). *Making educational reforms practical for teachers: Using a modular, success-oriented approach to make a context-based educational reform practical for implementation in Dutch biology education.*

Hu, Y. (2014). *The role of research in university teaching: A comparison of Chinese and Dutch teachers.*

Vink, C.C. (2014). *Mapping for meaning: Using concept maps to integrate clinical and basic sciences in medical education.*

De Hei, M.S.A. (2016). *Collaborative learning in higher education: design, implementation and evaluation of group learning activities.*

Louws, M.L. (2016). *Professional learning: what teachers want to learn.*

Moses, I. (2017). *Student-teachers' commitment to teaching.*

Veldman, I.M.J. (2017). *Stay or leave? Veteran teachers' relationships with students and job satisfaction.*

Chen, D. (2017). *Intercultural identities of English language teachers: An exploration in China and the Netherlands.*

Vereijken M.W.C. (2018) *Student engagement in research in medical education.*

Stollman, S.H.M. (2018). *Differentiated instruction in practice: A teacher perspective.*

Day, I.N.Z. (2018). *Intermediate assessment in higher education: Characteristics, perceptions, and effects.*