Available online at www.sciencedirect.com

ScienceDirect

Computers and Composition 43 (2017) 35–54

Computers and Composition

www.elsevier.com/locate/compcom

A Novel Approach to Examine the Impact of Web-based Peer Review on the Revisions of L2 Writers

Djuddah A.J. Leijen

Department of Estonian and General Linguistics, University of Tartu, Jakobi 2, 51005 Tartu, Estonia

Abstract

The use of peer feedback to support student writing is increasingly popular among second language writing teachers. However, the effect peer feedback has on the revision and the writing process remains unclear. Technological advances have made the application of peer feedback in the L2 writing context more accessible. As a result, there is a growing body of research investigating web-based peer review and L2 writing. This study aims to better understand how L2 writers conduct peer feedback activities, by looking at the types and traits of the feedback and how they influence revisions made in subsequent drafts using a web-based peer review system. A new methodology for studying web-based peer review comments is introduced. The results suggest that a specific type of feedback, *alteration*, and specific type of feedback, *recurring*, are important predictors for revision.
© 2016 Elsevier Inc. All rights reserved.

Keywords: peer feedback; revision; L2 writers; web-based peer-review systems; machine learning

1. Introduction

The use of peer feedback to support the development of student writing is now becoming increasingly popular in a wide range of educational contexts around the world. Certain aspects of peer feedback remain controversial. In the context of second language (L2) writing in particular, it remains unclear how peer feedback supports the L2 writing process and does peer feedback support the L2 writing process, and it remains unclear what the immediate and long term effects are on revision and the writing production (Gielen, Peeters, Dochy, Onghena, & Struyven, 2010; Hyland & Hyland, 2006; Lundstrom & Baker, 2009; Min, 2006; Zhang, 1999). Research on peer feedback has generated conflicting reports either confirming or refuting its usefulness and effectiveness. This notwithstanding, peer feedback on L2 writing has generally been viewed positively in higher education (Hyland & Hyland, 2006; Liu & Hansen, 2002; Topping, 2003), as it supports the process approach to writing, encourages collaborative learning, and learning through dialogue with peers (Hirvela, 1999), and is generally more consistent with current popular pedagogic concepts such as Vygotsky's zone of proximal development (Vygotsky, 1978).

This study aims to contribute to the growing body of knowledge regarding the use of web-based technology, peer feedback, and supporting L2 writers and writing instructors. L2 writers in the context of this study are all writers who are writing in a language (English) other than their mother tongue. Technological advances, specifically web-based

E-mail address: djuddah.leijen@ut.ee

http://dx.doi.org/10.1016/j.compcom.2016.11.005
8755-4615/© 2016 Elsevier Inc. All rights reserved.

technology, have made the application of peer feedback in the L2 writing context more accessible. Web-based tools such as blogs, wikis, and more recently web-based peer review systems are readily used in classrooms for the purpose of teaching writing and engaging students in peer review activities. As a result, there is a growing body of research investigating web-based peer review and L2 writing. The results of these studies are as multiple as the settings in which peer feedback and writing is found (Goldin, Ashley, & Schunn, 2012). Consequently, as the amount of data and different settings grow, so is there a need for new research methods able to investigate these differences systematically. This study applies machine learning as a novel method on a large set of web-based generated data of L2 writers using peer feedback using a web-based peer review system (SWoRD) to support the development of their text over a number of drafts. Machine learning is as statistical method that essentially represents patterns in large set of empirical data in order to make predictions about this data when applied in different settings. These predictions help explain how these patterns perform in a new situation (Leijen, 2014; Conway & White, 2012). Previous studies have indicated that machine learning is both a powerful statistical method to study larger sets of authentic data (Leijen, 2014), and shown to offer new perspectives in writing research studies (Leijen & Leontjeva, 2012; Crossley, 2013; Xiong, Litman, & Schunn, 2012). In addition, as Scott A. Crossley (2013) points out, current advances in both computational linguistics as well as methodological advances in L2 writing research allow for more replication studies, which can, for example, be carried out using machine learning techniques. For example, once specific patterns have been found, these can be tested again by improving the model or changing the setting. As a result, these replication studies on writing tasks can provide a richer understanding of the processes, variation in, and development of writing by generalizing from real examples.

Consequently, the main aim of this paper is to better understand how L2 writers conduct peer feedback activities by looking at the types and the traits of the feedback they provide and how these may influence revisions made in subsequent drafts. Additionally, in response to Crossley's (2013) suggestion for replication studies focusing on second language writing using novel techniques, this paper introduces a new methodology for studying web-based peer review comments and their effectiveness on the revision on subsequent drafts. Finally, the paper extents the discussion to include how the results are beneficial to academic writers and writing instructors in general.

2. Web-based Peer Review Systems

Web-based peer review systems are specifically designed to support the reviewing and revision process of writing. For example, SWoRD™ (Cho & Schunn, 2004), Calibrated Peer Review™ (Russell, 2004), MyReviewers (Moxley, 2012), and ELI (Hart-Davidson, McLeod, Klerkx, & Wojcik, 2010) were all developed at north American higher educational establishments. Research on the use of these systems and the effect these systems have amongst L2 writers has so far remained unexplored.

Although different tools, such as wikis, blogs, *Google Docs,* and the traditional Review tool in *Microsoft Word* are readily available to educators and researchers, the development of web-based peer review systems has gained momentum. Although these systems share some communal features with the systems mentioned above, they are fundamentally different. First, most of the systems such as wikis, forums, and blogs are structured around discussion (feedback and commenting) on a specific text (blog), or collaborative text creation (wiki). *Google Docs* and *Microsoft Word* are centered on text creation and offers tools to allow for comments and tracking revisions (Track Changes tool). The disadvantage of these systems is how they collect data. Researchers need to develop a system for collecting individual pieces of work from students, which, when attempted on a large scale, can fast become a daunting task. This is where web-based peer review systems come into their own as a research resource: By virtue of their very design, they provide the researcher with the means to collect large well-structured sets of student-generated data in a corpus for the purpose of analysis. An added advantage of web-based peer review systems is that the writing itself is often supported with additional features such as changeable rubrics and personal peer feedback prompts. Finally, web-based peer review systems are developed and structured around the educational practice of peer feedback on writing and are flexible to fit most of the writing tasks students are already likely to be engaged in.

The basic principles behind the development of web-based peer review systems are to support the development of content knowledge through writing, to assist in organizing writing assignments for instructors to reduce the workload, to provide students an authentic audience for the giving and receiving of feedback, and to enable students to practice writing through constant revision for the purpose of learning how to write (Cho & Schunn, 2004). In practice, additional aspects have been shown to be useful and effective for both students and instructors; in particular, a number of studies

have found that multiple peer feedback seems to improve writing more than feedback from one teacher (Cho, 2004; Cho & MacArthur, 2010; Cho, Schunn, & Charney, 2006; Cho, Schunn, & Wilson, 2006). As well as stimulating learners to engage in critical communication on the writing process, web-based peer review systems allow for this engagement to be done anonymously, which is likely to provide students with additional encouragement both to offer comments and to respond constructively to them (Guardado & Shi, 2007; Wadhwa, Schulz, & Mann, 2006). As previous studies have demonstrated, different features of peer feedback can be attributed to the system itself, and certain aspects of these features are known to influence how students revise their text. Precisely how and why this process of peer feedback influences revision in a subsequent draft is the question that the current study aims to investigate.

3. Peer Feedback in Web-based Environments

While there is a growing body of evidence that peer feedback can have a positive effect on second language writing (Berg, 1999; Lundstrom & Baker, 2009; Tsui & Ng, 2000; Tuzi, 2004), researchers have also found that L2 writers often find it difficult to give peer feedback (Hyland, 2000; Leki, 1990; Lockhart & Ng, 1993; Nelson & Murphy, 1993). The appeal of computer-based peer feedback systems is that they appear to offer new and more student-friendly ways of facilitating the peer feedback process (Davis & Thiede, 2000; Díez-Bedmar & Pérez-Paredes, 2012; Matsuda, Canagarajah, Harklau, Hyland, & Warschauer, 2003; Ware & O'Dowd, 2008; Ware & Warschauer, 2006). However, while these studies all agree that computer mediated communication in general provides a rich learning experience, it is also clear from these reports that different web-based platforms result in different types of peer feedback comments.

Research on second language writing has found that when commenting on each other's work, L2 students often focus on local issues, such as spelling, vocabulary and grammar, and much less on global issues, such as style, and content (Biber, Nekrasova, & Horn, 2011; Ferris, 2004; Leki, 1991). Similar findings were reported in a study directly comparing traditional and technology enhanced feedback delivery methods (Liu & Sadler, 2003). According to Jun Liu and Randall W. Sadler (2003), the use of technology enhanced the production of peer comments in general, and led in particular to the production of comments that contained specific suggestions for revision, usually focusing on small textual changes. Frank Tuzi (2004), investigating the impact of e-feedback on L2 writing, found that technology enhanced peer feedback supported more macro-level changes in the text. This finding led him to argue that the type of feedback peers receive is influenced by the interface of the web-system itself; macro-level problems might be easier to describe when revising a paper online than micro-level problems such as grammar or spelling errors. This is in line with Paul Kei Matsuda et al. (2003), who proposed that different web-based platforms are likely to change the nature of the feedback and strongly influence the language used as well as highlight specific communication features which effect both the peer review comments themselves and the revision of a subsequent draft. The web-based systems themselves, for L2 writers, should support the focus of peer feedback to be oriented to global aspects, and for the purpose of language learning, can support local aspects. Supporting local concerns, despite it being considered a negative finding, can also be considered an opportunity for learning: Specifically for L2 writers.

Within the context of revision in general, focusing on small textual changes has also been linked to L1 writers engaged in feedback. For example, Larry Beason's (1993) study investigating feedback and revision in a WAC setting, discovered that at least half of the comments students made were local changes. This number was fairly consistent across the discipline, but also in comparison to other studies (Dheram, 1995; Faigley & Witte, 1981). More recent studies assumed that comments on local aspects of the text are more likely to be implemented (Nelson & Schunn, 2009). However, Melissa M. Nelson and Christian D. Schunn's (2009) study, which investigated revision of students using a web-based peer review system, did not find evidence for this to be the case. They concluded, as has also been found by other studies (Kluger & DeNisi, 1996) that the presentation of the information in a comment is much more important, rather than focusing on local issues. As such, within this perspective, this conclusion may also be extended to L2 students peer feedback.

4. Effectiveness of Feedback Type

Feedback that is considered useful is not always considered effective and what is considered effective is not always considered useful. Nelson and Schunn (2009) determined that peer feedback, in web-based peer-review systems, are the most effective when feedback comments contain localization statement (specifying where the problem occurs in the writing), when a solution is offered (providing suggestions how to solve the problem), and a summary (restating

the main points) is included. They derived these findings based on an extensive literature review resulting in an initial model of peer feedback (Nelson & Schunn, 2009). Students' peer feedback data were compared with their text and measured in terms of successful implementation. Linking these findings to L2 writing research, it appears that a critical feature in this model does not seem to influence peer feedback effectiveness, the variability of the language use itself, and the effect this variability might have on understanding student's comments. Pragmatic studies concentrating on student (or teacher) language in feedback have suggested that mitigating devices (such as hedges, indirect speech acts or non-specific reference to the addresser, disclaimers, etc.) in comments, or in expressions of criticism, are not always understood by the receiver (Ferris, 1997; Hyland & Hyland, 2006; Nguyen, 2008). As a result, specific expressions produced by peers might be misinterpreted by the receiver due to a lack of understanding of subtle meaning expressed by mitigating devices, for example. Expressions such as "*I don't see a thesis statement in your introduction*" could come off as harmful, and an expression such as *"the author might perhaps consider trying to bring out the thesis statement in the introduction*" being too indirect. In addition, the inability to produce the right register could cloud the intention and therefore affect the effectiveness of peer feedback.

A recent study where a peer-reviewing system was used in a L2 writing environment offered some insightful results (Leijen & Leontjeva, 2012). Author, who tested Nelson and Schunn's model of peer feedback amongst L2 writers in a higher education institute, concluded that the aspect of localization seemed to positively predict implementation. Furthermore, they also confirmed that multiple peer perspectives, or rather, commenting on the same problem, had a positive impact. As such, different types of feedback can be tested to determine their effectiveness. In this case, the setting is L2 writers commenting on each others' written work using a web-based peer review system.

5. Method

5.1. Participants

A group of first, second, and third year full-time Bachelor's students (N = 43; female N = 19; male N = 24) from the department of Chemistry at a University in Estonia participated in an English language academic writing course. The course was largely conducted as a web-based learning module within the curriculum of chemistry accessible only to students majoring in chemistry. The course ran for approximately four months and students participated in one introductory face-to-face classroom meeting, in which the workings of this course were explained. The main aim of the course was to develop students' English writing skills in the field of chemistry. As students could choose to enroll on the course at any stage during their studies, no distinction was made between the students' levels and grades and all students participated in the same activities. The curriculum of chemistry is primarily taught in Estonian, with a few exceptions, i.e. courses specifically aimed to improve the English language skills needed in the field of chemistry.

All students were Estonian nationals. They were either native Estonian language speakers or bilingual Russian and Estonian language speakers. The Estonian higher education system is primarily conducted in the national language, Estonian. Thus, sufficient Estonian language proficiency is required from all students for enrolment to any university program. For English proficiency, the Estonian high school curriculum has as its required outcome an English language proficiency level of B2 for all students in accordance with the Common European Framework CEF (Little & Perclová, 2012). In terms of writing proficiency, students with a B2 level "can produce clear, detailed text on a wide range of subjects and explain a viewpoint on a topical issue given the advantages and disadvantages of various options." (Little & Perclová, 2012). The average age of the students at the time of taking the course was approximately 22 years. The English language proficiency requirement for attending the writing course was set at B2.

This writing course is the only course that students attend where a web-based peer-review system is used as part of the program. None of the students in this group had attended an English language academic writing course prior to this one, nor had they engaged in any official web-based peer-review tasks. The initial face-to-face meeting informed the participants about the course delivery, goals and outcomes of the course, and the writing assignment. It also introduced students to the workings of the web-based peer-review system. For example, students were instructed how to upload text, how to respond to the text of their peers according to the prompts provided by the system, and how to resubmit a revised text. In addition, information about the usefulness of peer feedback and general guidelines of good peer feedback practice was provided. Overall, students had little experience with the 'write and revise' paradigm. Finally,

SWoRD Peer Assessment **FLKE.01. English Academic Writing for Students of Chemistry** Course Info | Students | Assignments | Grades | Stats | Export Data | Contact

All Reviews for Body Text 1 - Draft #1 by (Panda)

View Submitted Document

Reviewer	10-Wrt Conventions	Level of argument
(marbles)	5	7
(Midget)	7	7
(Dixi)	5	5

Green=7s. Red=4 or less.

Reviewer	(marbles)	(Midget)	(Dixi)
Logic/support	The reasoning was good, it was easy to follow. I don't think this part needs much improvement. ***Backevaluation(5)***: Good. Thank you.	Strengths: you have good ideas and you have presented them in a very clear and easily understandable way. Paragraphs are well structured and the issue is very controversial one at the moment, making it very interesting to read. Weaknesses: There was one sentence that did not quite make sense to me: "The safety of the plants can be considered as a pro and also as a con." How can safety be a disadvantage? Suggestions for improvement: hard to suggest anything since the essay is already very well written. Maybe you could add more examples to illustrate your statements better. ***Backevaluation(5)***: Safety in the meaning that actually they are not so safe. Thouhg they are trying to improve the safety, there still is a possibility for huge disasters. Adding examples is a good idea. Thank you	I like that you give a lot arguments. However, there are a lot of arguments supporting the negative sides of power plants and only a few for the power plants. Maybe you could bring out some more positive sides to having a power plant. The structure of your essay is unusual. You give some positive arguments, then say negative things and then return to positive sides. Overall, I really like the theme and i think that you have given good arguments for the essay. ***Backevaluation(5)***: My vision was first to speak about plant's genereal pros and cons and after that what are the pros and cons for the country. And I'll look up for more positive aspects.
Fundamental writing issues	In the second paragraph, I think it should be "!a!relatively small amount of volume contains !a! big amount of energy." 3rd paragraph: I would use "chance" or "possibility" over "opportunity", because "opportunity" sounds too positive. 4th: "Dangerousness" is not a very good word, just use "danger" "there still is not a solution" is also quite bad, use "there is still no solution" 5th paragraph: I would use "it has it's own !source of energy!" ***Backevaluation(5)***: I will consider these remarks. You said "it has it's own.." Actually is is: "it has ITS.." whitout apostrophe. :)	Spelling and grammar is excellent! You use a great variety of words that make your text interesting to read. There were a couple awkward sentence structures. For example: "In addition to this, there still is not a solution to nuclear waste that is extremely dangerous." I think it sounds better if you say "In addtion to this, there is still no solution to..." There were also a few commas missing. For instance, there should be a comma after "For example," However, these are only minor mistakes that do not compromise the undrestanding of the essay. ***Backevaluation(5)***: I corrected the sentences, you mentioned.	"This essay will consider positive and negative sides of power plants and the reasons why every country should or should not have it/them." I would choose it or them and not include them both. "If to compare nuclear energy to solar and wind energy, it is more efficient and it is available 24 hours a day." I would probably use Comparing nuclear energy with... "Letâ€™s say that every step made is correct and no other harmful factors are present." I think that this sentence is unnecessary. Maybe it would be better just to say THE MAIN ADVANTAGES OF... "The safety of the plants can be considered as a pro and also as a con." I don't understand how can you be against the safety? Or maybe you mean that the safety has its pros and cons? Don't make your sentences too long, as they might become difficult to read. You should also check your decimal points. "tens of years"- "several decades" is better To sum up, you have a good word choice and it is really enjoyable to read your work. Well done! ***Backevaluation(5)***: I considered changing the style of the sentences. And definitely will try to make some sentences shorter. And yes, it is better to say that the safety has pros and cons- that makes the sentence more logical. Thank you

Fig. 1. Feedback and back evaluation of the feedback creating a dialogue between peers in SWoRD web-based peer review system.

students watched a web-based tutorial video (offered by the University of Pittsburgh's peerfeedback.net), which advised them to be constructive, specific, and respectful in their feedback.

5.2. Writing Task and Data Collection

Students were required to write an argumentative academic text (containing a thesis statement, evidence based supported arguments, and conclusion/discussion) over a period of 12 weeks (see Appendix A) in the web-based peer review system SWoRD (rebranded as Peerceptiv). SWoRD was developed to support writing instruction. More specifically, SWoRD supports student's critical thinking skills through the process of having them engaged in peer review tasks (see Fig. 1). In general, students join an instructor developed course online. The first step in the process is for students to upload their completed draft onto SWoRD. Once all the works of the students have been collected, instructors can randomly assign multiple peers to review the uploaded texts. The reviewing process is anonymous and guided by task specified feedback prompts and grading rubrics (see Appendix B). Peers provide their feedback and rate the text according to the instructions of the instructor. Next, students receive and review the multiple reviews on their text and make consideration regarding the helpfulness of the feedback by provided a usefulness rating and providing back evaluations on their reviews. As a result, students make changes to their text and, if the course instructor so wishes, the process can be repeated.

The main objective of the writing task was to introduce students to academic writing in chemistry through reading and writing and commenting on domain specific texts. The underlying assumption of the course is that learning to write academic English is genre specific and best learned within the discipline (Feak & Swales, 2004; Kuteeva, 2013). Students had to write about topics discussed in their chemistry courses. Prior to writing the text, students' topics and outlines, including references, were submitted to the course instructor for approval. Every week, students were presented with course material in *Moodle* related to the field of chemistry (e.g. specialized English used in Chemistry), features of academic writing (global aspects of writing, argumentation, thesis statements, organization), readings for writing guidelines (example texts and journal articles), and general English language learning materials related to writing (such as punctuation rules, verb tenses in writing, cohesive devices).

The writing task was broken down into three stages, each involving a cycle of draft submissions; feedback on the draft submission by multiple peers; back evaluation of the feedback by the author to the reviewer; and resubmission of the draft. In the first stage of the process, students were asked to write the introduction of their text. The introduction had to contain a clear thesis statement and outline of the topic. Students submitted their text to SWoRD for peer feedback.

Once submitted, the system allocated five random peers and required each of them to provide peer feedback on the introduction text. Students were given one week to complete and upload this feedback. Once the peer feedback stage was completed, students were given a week to make revisions to their original text and 'back evaluate' the received feedback for usefulness.

Next, students were given three weeks to write the body of their text. This main text had to contain clear arguments concerning the topic and references to sources used to support the arguments; conform to the conventions of a five-point argumentative essay; include written academic language features such as linking devices, the use of an impersonal voice, and well considered verb tense applications in paragraphs. On completion, students submitted their main body of the text including the revised introduction text to SWoRD for peer feedback. Three random peers were selected to give feedback on the text. As before, peer reviewers were given a week to complete the reviewing process. At the end of the reviewing process, students were given another week to make revisions to their text and back-comment on the received feedback. For the final stage, students had to write a discussion/conclusion text. The discussion/conclusion had to link back to the introduction, and summarize the main points of the text. The text was submitted to SWoRD together with all the revisions made to the previously uploaded text. Three peers were randomly selected to give feedback, and once again given a week to complete the peer feedback process. After the peer feedback process had finished, students were given two weeks to revise their texts for submission to the subject teacher for final evaluation. Adhering to this procedure ensured that students would produce at least two comparable versions of the specified parts of the text: A drafted version and a revised post peer feedback version. The instructor of the course assessed the students after the course was completed. Students were assessed according to their level of participation in the web-based peer review systems, and process of their final product. The outcome was a pass or fail.

Students, upon registering to the web-based peer review system, signed a consent form to have their data used for research purposes with complete anonymity guaranteed.

5.3. Feedback prompts

The instructor of the course selected the feedback prompts relevant to the writing task (See Appendix B for examples of the prompts provided by SWoRD to assist students on commenting on both global and local writing issues). Feedback prompts were provided to support peer reviewers during each round of the writing process. As the writing exercises were designed to develop discipline-specific texts following English academic writing conventions, the focus of the prompts was on both global writing issues (topic, logic and support, organization, thesis statement) and local writing issues (spelling, grammar, sentence structure, academic style, punctuation). The prompts served as a general guideline to help students focus their comments on both aspects. Throughout the course, students were not provided with additional guidance on the peer feedback process, nor did the instructor of the course intervene at any stage during the peer feedback process. In order to receive a pass for the course, students were informed that the instructor of the course would review the feedback students gave to each other. Failing to provide feedback would automatically result in a fail.

For the purposes of this study, only the peer feedback comments on the different parts of the texts (introduction, body, conclusion) and the different versions of the texts (draft 2, draft 3, and the final submission, all of which contained revisions) were used in the data analysis. A raw version of the peer feedback comments was collected in tabular file format, containing the participant identifiers of the writer and peer feedback giver, plus the feedback comments.

5.4. Data segmentation

Feedback comments could be associated with one aspect of the text or contain multiple comments. For the purpose of the analysis, the feedback comments were segmented to include only a single reference to an aspect of the text. In other words, the smallest group of words (sentence or clause) meaningfully expressing a statement, question, exclamation, request, command, or suggestion related to the reviewed text was taken as a segment. For example, *"I think it's a bit short for saying all about this topic. To improve that author should include own ideas and thoughts or arguments in the main body of the essay",* was segmented into two separate instances. The first instance being *"I think it's a bit short for saying all about this topic"*, and the second being *"To improve that author should include own ideas and thoughts or arguments in the main body of the essay"*. Another example, *"There is no clear thesis statement and also no clear plan"* was segmented into two instances: *"There is no clear thesis statement"* and *"and also no clear plan"*. As this study

primarily focuses on the visible revisions made in a subsequent draft as a result of concrete revision oriented comments (implicit or explicit) over three iterations of an essay, non-revision oriented feedback instances were excluded from the analysis (e.g. general expressions of praise like *"great job"* or *"I have nothing to comment"*, and more specific positive comments such as *"you have given a clear overview"* or *"there are no verb disagreements"*). Instances containing a revision oriented comment and a non-revision comment, such as: *"Introduction is a bit short, but fluent and coherent",* were segmented into two instances: *"Introduction is a bit short"* and *"but fluent and coherent"*. The non-revision instances were excluded from the analysis. Additionally, as the purpose of this study was to make predictions about revisions in subsequent drafts, only instances that include an implicit or explicit reference for revision were included in the analysis. As a result, after segmentation and exclusion of non-revision oriented comments, the corpus contained a total of 885 peer feedback instances.

5.5. Data coding and features of analysis

The initial peer feedback features selected for this study are presented in Table 1. The features were categorized either as *Type, Trait,* or *Revision.* The *Type* of feedback generally classifies the feedback as *identification, justification*, *alteration,* or *area.* The *Trait* of the feedback refers to specific language related features such as *question, point of view (POV),* and *localization (LOC).* Most features (see Table 1) were adopted from the pre-existing frameworks from former studies and were selected because they were suggested to be important either for the revision process or peer feedback process (Gielen et al., 2010; Leijen & Leontjeva, 2012; Liu & Sadler, 2003; Nelson & Schunn, 2009). Finally, *Revision* measures the visible revisions made in a subsequent draft and categorizes the feedback instances as *revised* or *not revised.* To ensure the robustness of the features for the analysis, most features were coded by two independent coders and had to 1) score higher than .75 in a Cohen's Kappa intercoder reliability test, using Joseph L. Fleiss, Bruce Levin, & Myunghee C. Paik' (2013) benchmark scale for the Kappa (more than .75 is considered as excellent reliability); 2) be transferable to, and identifiable in, datasets obtained from other web-based systems (such as wikis, blog, and forums) to allow for replication studies; and 3) ideally, be coded automatically in future research (either using part-of-speech tagging or syntactic parsing).

Most of the identification of the features categorized in *Trait* of feedback was coded through specific word searches and grammatical subject analysis. Intercoder reliability calculated over 10% of the randomly selected data (*recurring* and *LOC* were considered to be robust enough to be coded by a single coder). As a result, only the features *question* (Cohen's kappa = .75) and *POV* (.83) passed the intercoder reliability test. The features *agent* (.74) and *object* (.53) were eliminated from the analysis.

The descriptions of the categories listed under *Type* of feedback were first discussed and agreed upon between the two independent coders. A first round of coding was conducted on part of the randomly selected data after which the feature descriptions were discussed and adjusted, if needed. Once feature descriptions were found to be sufficiently clear for an independent researcher to code, intercoder reliability tests were conducted on a 10% random sample of the data, and coded by two independent coders for the features *identification* (Cohen's kappa = .84)*, justification* (.88)*, and alteration* (.85). The intercoder reliability test for the feature *area* was conducted on the whole dataset and achieved a reliability score of .90. Finally, the response variable *revised* was dual-coded throughout and obtained a reliability score of .76. In both cases, these two features were considered important for this study to be coded on all the full dataset. For a detailed description of the selected features see Appendix C.

5.6. Statistical Analysis

A machine learning approach was used to predict how the selected peer feedback features, related to the peer feedback comments, performed on the L2 writing task. Applying machine learning algorithms should assist in gaining a better understanding of the effects L2 online peer reviews have on L2 writers by predicting which of these tested features lead to revisions. More specifically, machine learning will allow to develop a model of effective peer feedback by going through the data and find patterns. In this study, machine learning was used to test the performance of eight features related to multiple online peer review on L2 writing. By testing these features, it might be possible to detect patterns that allow for a more advanced predictive analysis of the data. For example, if patterns are found in the data, these can, using the same method, be tested on another set of data with the same features to see if these patterns recur, or whether these patters are context specific. Additionally, changes could be made to the context in order to improve

Table 1
Peer Feedback Feature Description.

Predictor	Short description	Example
Type of feedback (modType)		
Identification	*Identification* categorizes the feedback instances as either a *Problem* (PRB) or an *Action* (ACT). This feedback type is identified as either referring to a (general) problem in the text that needs solving (PRB) without it explicitly stating how the problem is solved or as an action that needs to be taken by the author in a subsequent draft (explicit).	PRB: "*The text is too long for an introduction and over explained.*" ACT: *"Maybe you could try to put the main subject clearly into one sentence."*
Justification	The feature *Justification* (feedback instance includes a justification or explanation yes/no) categorizes feedback instances that have either include or not included a justified (or explained) opinion in the feedback (Gielen et al., 2010; Nelson & Schunn, 2009)	*"I would suggest uniting short sentences into longer ones, it would be easier to read* (justification)*."*
Alteration	*Alteration* (YES/NO) indicates whether a feedback instances points to a specific change (Liu & Sadler, 2003).	*The last sentence (I cannot. . .) I would rephrase for clarity. Maybe, "So, why is it important" (alteration) or something in those lines.*
Area	Feedback instances refer to either *global* areas (e.g. audience and purpose, idea development, topic, logic and support, organization of writing, thesis statement) or *local* areas (e.g. wording, spelling, grammar, sentence structure, academic style, punctuation) (Liu & Sadler, 2003).	Local: *"I also found a spelling mistake. It should be mankind's instead of mankinds. And that's about it I could find. I had to whine about something;)"*
Trait of the feedback (modTrait)		
Question	Does the feedback instance ask, question, and raise doubt or confusion? (YES or NO).	*Just wondering about the use of words here, it sounds like a specialized knowledge.*
Agent[a]	Allocates the person or thing that does the action in the feedback instance and is labeled as *text, author, reviewer*, or *unknown*.	*Text: This sentence in the second paragraph needs to be shortened to improve understanding.*
Point of view (POV)	Is the feedback instance stated from the point of view of *author, reviewer, or neutral* (Morand & Ocker, 2003; Walther, 1992) .	*Reviewer: I strongly suggest that, the author has to reformulate . . .*
Object[a]	Is the action of the type of feedback referring to a *general* action, an action rooted in the *text*, or *unknown*.	*General: Linking words can fix the problem of long sentences. (GEN)*
Recurring	Have other peers referred to the same/similar aspect in their feedback.	*Reviewer1: Seems title is missing* *Reviewer 2: I would suggest the author to include a title, as the task required a title to be included.*
Localization (LOC)	Indicating specific locations in the text that the feedback instance is referring to, often denoted with *page, line, and paragraph*.	*The author has to reformulate the second sentence in the 3rd paragraph to include a . . .*
Revision in a subsequent draft		
Revised	Revised is the measurable outcome of students making a change in a subsequent draft that can be linked to the peer feedback instance. Revised in the context of this study is understood as the act of altering something based on the feedback that has been provided by peers. Revised does not include the measurement of improvement.	

[a] Eliminated features from the analysis after inter-coding reliability testing below 0.75 Cohen's Kappa.

the strength of the prediction of these patterns. Overall, the significance of this method is that no assumptions are made about the features itself, and the output of the analysis is used to test whether these patterns are recurring or whether these patterns are different on a different set of data under different circumstances.

To investigate the relationship between the response variable (*revised*) and the predictor variables (*identification*, *justification*, *alteration*, *area*, *question*, *POV*, *recurring*, *and LOC*), logistic regression was used to test the predictive

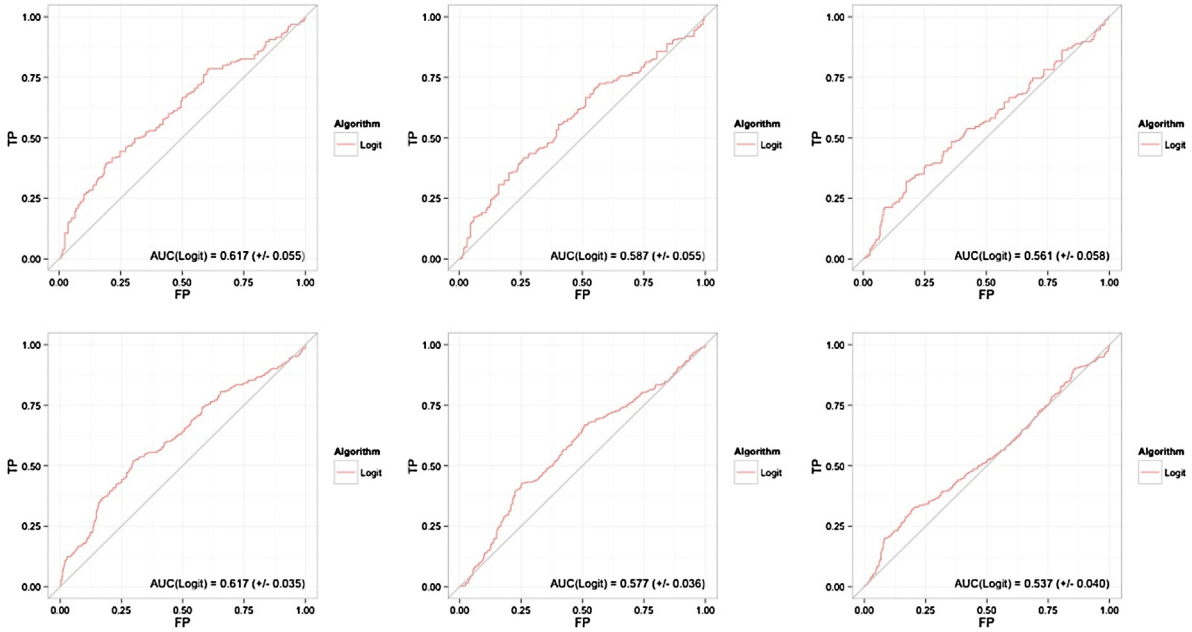

Fig. 2. ROC curves for the logistic regression test on the three models, modAll, modType, modTrait with response feature revised. From left to right, the three feature models (Model 1 containing all features, Model 2 containing the features *Type*, Model 3 containing the features *Trait*. The bottom row, from left to right, was calculated using 10-fold cross-validation to test for reliability.

strength of 3 different models: Model 1 contained all features (henceforth referred to as modAll); Model 2 contained the *Type* features (modType); Model 3 contained the *Trait* features (modTrait). The logistic regression prediction algorithm tested which predictor variables best predicted the response variable, and a statistical model was constructed using a training set (a partition of the whole dataset), and validated on a testing set (the unseen partition of the whole dataset). To increase reliability, a ten-fold cross-validation was also applied. Additional interaction analyses were carried out on the types and traits of feedback to better understand the peer review process. The analyses were conducted using the caret (Wing, Weston, Williams, Keefer, & Engelhardt, 2012), PredictABEL (Kundu, Aulchenko, van Duijn, & Janssens, 2011), and PBSmodelling (Schnute, Couture-Bell, & Haigh, 2006) packages in R open source software (R Core Team, 2013).

6. Results

Of the 885 feedback instances included in the analysis, 458 (52%) were revised, and 427 (48%) were not revised in a subsequent draft. Similar results have also been found in previous studies amongst L2 writers (Leijen & Leontjeva, 2012; Liu & Sadler, 2003; Tuzi, 2004).

The first step of the analysis was to determine the predictive strength of the 3 models: modAll, modType, and modTrait, repeated by a 10-fold cross-validation to test for reliability and avoid overfitting. The results are graphically plotted as receiver operating characteristic (ROC) curves (see Fig. 2). ROC curves map false positives (i.e. false predictions) on the x-axis, and true positives (i.e. correct predictions) on the y-axis. Thus an ideal result, where the trained model perfectly predicts on a testing model, would be graphically displayed as a perfect corner (rising up the y-axis, to the top, carrying across to the right along the x-axis). It is, however, more likely that the results resemble a curve which lies above the middle threshold (i.e. a prediction by chance 50/50 or area under the curve (AUC) 0.500).

The ROC curves of the three different models displayed in Fig. 2 indicated that the overall performance of the modAll was better (AUC of 0.617 ± 0.055) than the performance of modType (AUC of 0.587 ± 0.055) and modTrait (AUC of 0.561 ± 0.058). This indicates that the model containing all 8 features is better at predicting whether a feedback instance will lead to a visible revision. Both modType and modTrait predict slightly higher than 0.500, or random guess. Possibly, a combination of features from modType and modTrait improves the overall predicted outcome. In

Table 2
Predictive Features (all Type and Nature Features) of logistic regression with Response feature Revised.

Predictors	exp (log(A)) = A	SE(B)	C.I. of odds ratio
Identification(Problem)	0.788	0.232	0.499–1.243
Justification(YES)	1.242	0.258	0.749–2.067
Alteration(YES)**	2.247	0.254	1.370–3.714
Question(YES)	1.438	0.348	0.729–2.869
PointofView(Reviewer)	1.572	0.325	0.833–2.993
PointofView(Text)	1.495	0.318	0.805–2.810
Recurring(YES)***	3.382	0.313	2.113–7.263
Location(YES)	0.977	0.311	0.531–1.806
Area(Local)	1.498	0.244	0.930–2.420

** $p < .01$.
*** $p < .001$.

other words, the combination of *Type* features and *Trait* features contribute to a better performance. Finally, assessing the overall performance of modAll, it is likely that not all the selected features are equally accurate in predicting implementation.

To further explore the data, the next step of the analysis was to determine which of the features could be considered the most important when predicting revision in a subsequent draft. This was done by carrying out logistic regression analysis on the training dataset, and confirmed on the testing dataset. The results of this analysis (see Table 2) revealed that two features – *alteration* and *recurring* – are more likely to contribute to revisions than the other features in the model. The odds ratio for the *Alteration* coefficient was 2.247 with a 95% confidence interval of [1.370, 3.714]. This suggests that students are more than twice as likely to revise based on a reviewer's feedback that includes an *alteration* than those that do not include an *alteration*. For the feature *recurring*, the odds ratio was higher, 3.382, with a 95% confidence interval of [2.113, 7.263]. Students were almost three and half times more likely to make a revision in a subsequent draft when another student also commented on the same topic.

To test how the feature *alteration* and *recurring* perform on the unseen testing data, an additional predictive analysis was carried out. A separate model was built for these two features only. The results confirm the results of the logistic regression for these two features (Fig. 3). The model containing only these two predictors improved (AUC of 0.641 ± 0.050). In other words, the model containing only these two features predicted revision better than modAll.

To test whether the features *alteration* and *recurring* interact, i.e. do they predict revision higher when they occur together, a two-way interaction analysis was performed. No combined interaction between these two features was observed. Additional analyses were carried out on these two features to better understand their relevance by adding the interactions to the model one by one to the basic model (modAll). Looking at the feature *recurring,* for example, revealed that only 14% of the feedback instances were *recurring*; the majority of the feedback instances were individual comments referring to a different aspect in the same text. Distributed over the *revised* and non-*revised* instances, nearly 73% of the *recurring* instances lead to a revision. Further analysis did not show any interaction between any of the other features. This might explain why *recurring,* as a feature, was the stronger predictor.

Alteration had a better distribution across the dataset; roughly 35% of the instances included an *alteration* and of these instances, 65% lead to a visible revision. As a feature, *alteration* has been associated with feedback with local changes resulting from feedback, but much less on global changes (Liu & Sadler, 2003) as it would require the reviewer to indicate how to change larger chunks of the text for the author. A cross tabulation of *alteration* and *area* suggested a similar picture. Only three percent of the global feedback included an *alteration*. However, when we make allowance for the fact that they were distributed over the whole dataset, only 34% of the comments actually offered an *alteration*. There was, however, no interaction between the feature *alteration* and *area*.

Alteration did, however, have a positive interaction with another feature: *Justification.* The odds ratio of the interaction of *justification* and *alteration* was 3.488 with a 95% confidence interval of [1.457, 8.772] improving the likelihood of students carrying out revisions if *alteration* is included. The odds ratio for only *alterations* was 1.387 with a 95% confidence interval of [0.946, 2.035]. This suggests that a revision in a subsequent draft is more likely to happen when the feedback type included a justification and an alteration is suggested, as the following example illustrates: *"Positive*

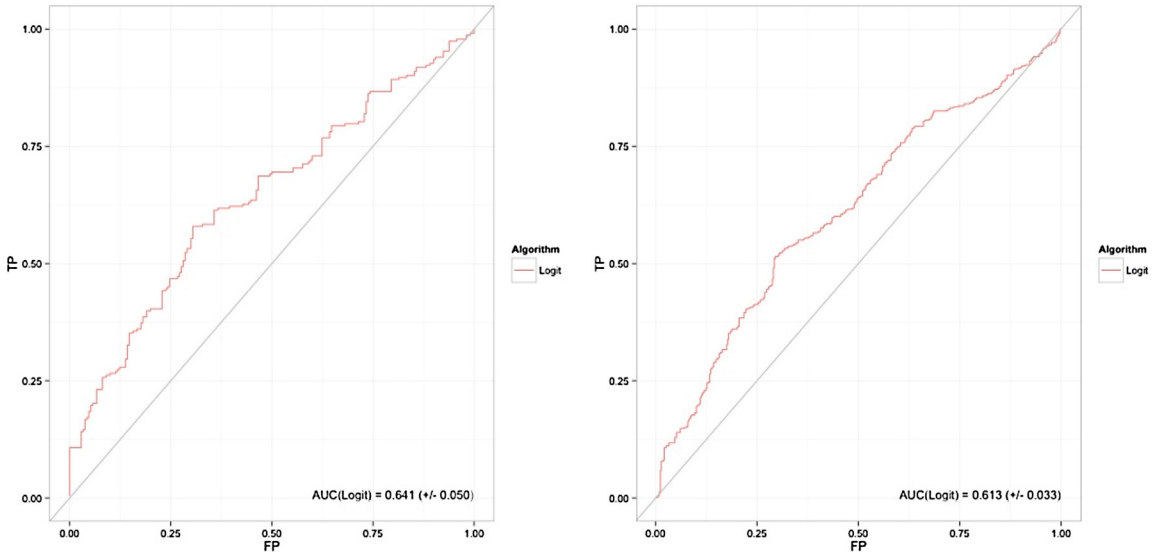

Fig. 3. ROC curves for the logistic regression test on the strongest predictive features *alteration* and *recurring*. Single analysis (left) and 10-fold cross-validation to test for reliability (right).

aspects and disadvantages- In my opinion it would sound better if the author would write positive and negative aspects or advantages and disadvantages". This was confirmed by the improvement in the model: The model containing only *alteration* had an AUC of $.600 \pm 0.051$, whereas the model containing the interaction *justification* and *alteration* had an AUC of $.619 \pm 0.052$. In the interaction model containing *justification* and *alteration*, *recurring* still remained a good predictor for revision. The odds ratio of *recurring* was 3.233 with a 95% confidence interval of [2.109, 5.062].

More interactions were found between features that negatively influenced revision. For example, for *POV* and *justification* the odds ratio of the interaction of *POV and justification* was 0.293 with a 95% confidence interval of [0.108, 0.782]. In other words, when the feedback was written from a neutral point of view, and included a justification, it had a negative influence on revision. This was also confirmed in the model containing the interaction of the features *POV* and *justification* (AUC of $.527 \pm 0.053$). Also a negative interaction was found between the features *justification* and *area*: The odds ratio of the interaction of *justification* and *area* was 0.328 with a 95% confidence interval of [0.141, 0.735]. When the feedback type contained a justification, but was referring to a local surface change, students were less likely to carry out a revision. This was also confirmed in the model containing the interaction of the features *justification* and *area* (AUC of $.540 \pm 0.053$). It would seem, that *justification*, as a feature, both positively and negatively influences revision. Finally, a slight negative influence on revision was found between *identification* and *alteration*: The odds ratio of the interaction of *identification* and *alteration* was 0.360 with a 95% confidence interval of [0.174, 0.738]. In other words, when the feedback type was written as a problem statement and an alteration was suggested students were less likely to revise their text based on that feedback instance. Although *alteration* was an important predictor, when combined with *identification/Problem* the chances of revision were reduced; however, this interaction was stronger than when *alteration* was not included. In other words, for this model, the odds ratio of *alteration* was 3.391 with a 95% confidence interval of [1.943, 6.024]. This suggests that *alteration* in combination with an action statement still predicts the chances for revision to happen quite well. This was confirmed in the interaction model containing the features *identification* and *alteration* (AUC of $.600 \pm 0.053$).

7. Discussion

The aims of this study were twofold. First, to better understand how L2 learner-writers conduct peer feedback activities by looking at the types and traits of feedback they provide and how these may influence revisions of subsequent drafts. The second aim of the study was to apply a novel approach to studying peer feedback, in order to gain a new perspective on the complex process of web-based peer review and its effects on revision for L2 learner-writers, as

enjoined by Crossley (2013). The model containing features identifying both the type and the trait of the feedback was strongest in predicting revisions. This result was mainly due to two features identified as strong predictors of revisions of subsequent drafts, post peer review: *Alteration*, a type of peer feedback; and *recurring*, a trait of peer feedback.

Alteration is a feedback type that points the author to a specific change needed in the text, as the following examples demonstrate:

1) "*I also found a spelling mistake. It should be mankind's instead of mankinds.*"

2) "*. . . For some, chocolate is just a tasty confection, but a considerable number of people consider themselves to be chocoholics and have cravings for it. Maybe you could replace 'considerable number of people' with some other phrase (a lot of people, a large number of people, etc.) It avoids repetition of words*.

The type of feedback that includes alteration is inherently revision oriented, as was also pointed out by Liu and Sadler (2003). Dana Ferris (1997) referred to this type of feedback as 'being explicit', and explicitness in feedback has been found to be an effective type of feedback (Tseng & Tsai, 2007). In the context of training writing teachers, Ferris (1997) recommended adding explicit suggestions in order to make the feedback comments more effective. It is not surprising, therefore, that comments which included specific *alterations* were twice as likely to lead to a revision than comments which did not, and three and a half times more likely when the alteration was accompanied with a justification. In other words, student peer reviewers not only need to be explicit, but also to provide an argument to justify the suggested change. This might also suggest L2 writers' reluctance to accept the feedback of a fellow language learner.

The other feature that predicted implementation was *recurring. Recurring* in this context indicated whether two or more peers independently commented on the same of similar aspect of a peer's text. For example, one peer comments: '*The introduction is way to long.*', and another student comments: "The text is too long for an introduction and over explained." As students received multiple peer reviews, it was likely that a student would receive comments on similar aspects in their text from at least two peers. As a result, similar comments, given by at least two reviewers, resulted in a three and half times greater likelihood that a revision would occur in a subsequent draft, compared to a feature that was only commented on by a single peer. This finding is broadly consistent with a previous study conducted on L2 writers by Leijen & Leontjeva (2012), and confirms the consensus that multiple peer feedback is more beneficial than single peer feedback (Cho & Schunn, 2007). Of particular interest is the finding that *recurring* did not have an interaction with any of the other features, and was also not affected when there were interactions to be noted in other features. It would seem that *recurring* by itself is an important predictor of revision in a subsequent draft after the receiving of multiple peer feedback. Within the specific web-based peer review system (SWoRD), peer reviews are always given independently from each other. Considering the importance of multiple comments on the same issue, it would be interesting to compare this feature to web-based systems where peer reviews are visible to other reviewers such as blogs and wikis and Google Docs. Opening up feedback for all peers to see might support the shared online communication that is often missing in asynchronous peer review in a web-based environment. For L2 writers this might be specifically helpful. If agreements occur between reviewers, this should have a positive influence on revisions made in subsequent drafts, regardless of the *Type* or *Trait* of the feedback. Replication studies are needed to confirm this assumption based on this finding, however.

Further analysis of the type and trait of the feedback also suggests that certain aspects of peer feedback, whether categorized as *Type* or *Trait*, influences revision. In the case of this study, these interactions seem to be more negative, rather than positive. For example, as the results showed, the feedback feature *POV* interacted with *justification,* as the following three examples illustrate:

(1) "*The introduction is too hectic, jumps too much from one argument to another.*"

(2) "*The point of the essay is also very hard to understand, because there is no clear line of thought.*"

Authors who receive such feedback are less likely to revise their text. It appears that feedback instances including a *justification* and written from the *point of view* of the text seem to create distance, which in turn reduces the strength and conviction of the comment. Additionally, referring to the text itself as the subject of the comment might be a strategy on the part of the reviewer for softening the justification, which, as the above examples highlight, are fairly direct

statements. This 'softening' effect is perhaps better explained by another negative interaction between the feedback types *justification* and *area,* as demonstrated by the following three examples:

(1) "*Also, I believe its inappropriate to say: Finally, Last together, because they mean the same.*"

(2) "*I would suggest uniting short sentences into longer ones, it would be easier to read.*"

(3) "*As in the second sentence i would write . . .as ammonia, bleaches and so one, because without and so one i felt this sentence was unfinished.*"

The feedback type *Justification* shares similarities to the feedback type *explanation* (Nelson & Schunn, 2009). Explanations clarify and give motive to the purpose of the feedback, but were found unhelpful and may even be harmful to the feelings of the author. This observation also finds support in a study conducted by Sheng-Chau Tseng and Chin-Chung Tsai (2007) among high school students: Explanation produced negative relationships to students' performance. In these examples that refer to local revision, the *justifications,* or explanations, seemed to have been intended as softening devices, but actually turn the clear suggestions into an utterance which can be interpreted as hurtful and face threatening. This might clarify the negative influence on revision of these two interactions between *justification* and *POV* and *justification* and *area.* In this case, if we consider feedback instances to be harmful only when they are formulated from the perspective of the 'neutral' text and includes a justification, informing students to write the feedback from the perspective of themselves (as reviewer), or from the perspective of the author (as writer and owner), might make the difference between comments being harmful or not. Similarly, when students comment on local changes, it might be better to exclude a justification for that change. The local change itself might be obvious enough (like a grammatical error) and informing the author about the 'simple' error might be too obvious and as a result cause hurt. This assumption would need to be further tested on large datasets.

Another interaction between features that had a negative influence on revision was observed between the features *identification* and *alteration*, as the following examples highlight:

(1) "*As the population grows, more food is needed and less land is available to support this. I think that the word SUPPORT is a little odd in this context. Maybe 'land is available to provide it'.*"

(2) "*A significant error was in the 3rd paragraph, Because of nitrogen, alkaloids have weak acidic properties. Nitrogen has basic properties.*"

If the *identification* of feedback was labeled as a problem statement, rather than an action statement, plus included an *alteration*, it was less likely to lead to a revision. In general, *alterations* accounted for a positive influence on revision; however, it seems that *alterations* which were not formulated as a simple '*do this*' instruction caused some confusion, which L2 writers might have difficulties with. Explicit comments are probably more effective, whilst comments such as the ones illustrated above that have a high degree of implicitness, might require more understanding from the reviewer to revise.

Observing the peer feedback performance from these negative interactions might suggests the importance to include features that account for the interpersonal aspects of feedback, as suggested in a previous study by Leijen & Leontjeva (2012). Specifically, those features that greatly vary amongst students, disciplines, countries, L2 language levels, and the web-based tool used. One of the aims of this study was to develop robust features that are replicable for different settings and other studies, thereby allowing for the observation of cultural differences in relation to phenomena such as politeness and facework during the giving and receiving of peer feedback. For any large dataset generated from a web-based peer feedback system, these features outlined in this study, specifically the ones which seem to influence revision both positively and negatively, can be coded for and tested using the same or a similar approach. In this case, these studies should not aim to confirm or reject these findings, but offer additional insights as to why the findings in these studies had different results. This supporting that conflicting results in studies are inherent to the complexity of the peer feedback situations.

Owing to the context of this study there are several considerations that should be highlighted. The course itself aimed to introduce and train English language writing skills in Chemistry using a web-based peer review system: SWoRD. The advantage of this system, as the results seemed to confirm, is that the prompts provided by the system, and included in the peer feedback process, helped guide the students to focus not only on local problems, but on global aspects of the text, a problem that frequently is associated with peer feedback offered by L2 writers. The importance

of prompts is confirmed by other studies (Gielen et al., 2010). There is no evidence that suggests feedback on local issues more frequently leads to revision than feedback on global issues. This in comparison to earlier research about the effect peer feedback has on revision for L1 writers. Beason (1993), for example, in his study found that most students focused on local revisions and the least to global revision. The use of a web-based system, including prompts, has possibly resulted in students being more on task in their review and revise tasks.

What the data does suggest is that the specific types of feedback, *alteration*, and the specific trait of feedback, *recurring*, are important predictors for revision. In terms of L2 students learning to write, this result might not be considered the most optimal result. *Alteration* is perhaps the most 'simple' revision to implement as a clear suggestion is provided on how to change the text. *Alteration*, therefore, might not always lead to an improvement of the text, if students do not consider them carefully, but blindly implement them. As the quality of the peer feedback (in terms of whether the feedback made correct or appropriate comments about the text to be revised) was not measured, it was not possible to predict, in the context of this study, whether *alterations* led to improvement or not. *Recurring* might actually be the better feature for predicting learning, as recurrence is probably closely linked with a negotiation of meaning, which is much more commonplace in a language learning environment. The disadvantage of SWoRD was that the system itself did not allow for a more open discussion amongst the reviewers and the authors; however, one additional aspect of the web-based peer review provided by SWoRD was not included in this study: The back evaluations given to the reviewers by the author post peer feedback. There might be evidence in these comments that provide agreement and disagreement concerning the provided feedback. Further studies should include this 'conversation' in their data analysis. In addition, web-based peer review systems might consider opening up the discussion for all reviewers to see all reviews during the peer feedback process. More research is needed to see if less anonymity actually supports revision and increases learning.

Perhaps the most interesting feature is the one that interacted both positively and negatively upon revision: *Justification*. The feature *justification*, in combination with *alteration* increased the chance of revision taking place. However, *justification* in combination with *area/*local and *POV/*text reduced the chance for revision to take place. As also claimed by Nelson and Schunn (2009) and Tseng and Tsai (2007), intuitively one would assume that providing a *justification* or explanation should facilitate the understanding of the problem or action. In the present study, however, this was only found to be the case when an *alteration* was given; it was not the case when it was on a local issue and written from the perspective of the text. This feature would need to be further investigated in combination with other features on larger and different datasets. Such studies would benefit from an additional qualitative analysis to better understand how students interpret *justifications* and in what instances they are considered harmful or helpful.

Overall, the results of this study are both valuable and applicable to all writing teachers who make use of a web-based application for peer review on writing. Knowing how specific contexts alters the effectiveness of peer feedback, better guidance can be given to feedback practice and training. For example, knowing that justifications are major contributors to effective peer reviews in web-based systems where the peer review is anonymous allows the instructor to pay specific attention to training this prior to the peer feedback tasks. Also, knowing that affective language use in the feedback does not effect the effectiveness of the feedback at all, would allow the instructor to inform their students to be more direct and concrete in their comments towards each other. This can be extended to cover differences between novice and expert writers and how their feedback differs and the effect that difference has on the effectiveness of the feedback. This study highlights the patterns L2 writers commenting in a web-based peer review system SWoRD respond when testing existing models.

Acknowledgement

The author would like to thank Anna Leontjeva from the Institute of Computer Science at the University of Tartu for guiding me through the statistical analysis of this paper.

Appendix A. Students writing assignment

Course assignment prompt.

Take a topic in Chemistry (from your Introduction to Chemistry course) which interest you, or you wish to further investigate. Find at least 3 research articles/texts which relate to this topic and write an argumentative essay on this topic, using the common 5 paragraph structure, introduction, Body (Point 1, 2, 3), and conclusion. Please make use of the following guidelines when developing your text.

1. taking an informed position on the basis of reading and research.
2. the logical and persuasive development of arguments to support this position.
3. the appropriate use of evidence to support these arguments.
4. clear cohesion between the individual components of the essay (the introduction, arguments and conclusion).

In addition;

1. identify the position you are taking.
2. predict how your arguments are going to unfold.
3. see logical relationships between the different arguments and pieces of evidence present in the body of your essay.
4. see how each argument relates to the overall position you have taken.

If you use these strategies well, your readers will never be confused about 'where your writing is going'. It does not matter whether they agree with your arguments or not – they will still see your writing as logical, well presented and cohesively argued.

Appendix B. Peer review prompts for the writing assignment provided by SWoRD

Concerning	Example Prompts
Global writing issues (higher order)	1. Logic/support. Identify the main strengths and weaknesses of the document in terms of the reasoning/support that was provided for the main claims or arguments or proposed actions. Be specific. Provide clear suggestions for improvement. 2. Conclusion. How well did the author include the appropriate content in the conclusion (i.e., from the introduction and theory, experimental setup, and data analysis)? **** The conclusion should capture all the important points from each of the sections without having any superfluous information (i.e., it was the perfect length). The conclusion should connect information from each section of the report in order to summarize the final results and sources of uncertainty. 3. Level of argument. How well does the author of the paper present their arguments? 4. Evaluation of topic. Do you think the author of this paper has given a clear overview of the topic he/she wishes to write about? If yes, why do you think so, if no, what are its shortcomings? 5. Document organization. Identify the main strengths and weaknesses in the organization of the document. Be specific. Provide clear suggestions for improvement. 6. Did the author include a clear, specific thesis in his or her introduction?
Local writing issues (lower order)	1. Fundamental writing issues. If there were some common problems at the basic writing level (spelling/grammar, poor word choice, awkward sentence structures), please describe the type of problem including the location of one instance. 2. How well did the author demonstrate a good grasp of standard writing conventions (e.g., spelling, punctuation, capitalization, grammar, usage, paragraphing) and use conventions effectively to enhance readability? 3. How appropriate are the writing style and vocabulary for an academic essay? 4. Word choice: Words convey the intended message in a precise, interesting, and natural way. The words are powerful and engaging. Please comment.

* https://go.peerceptiv.com/.

Appendix C. Detailed description of the 8 selected peer feedback features, examples, and inter-rater reliability score

Predictor	Description	Examples	Additional information and references
Feedback Type (modAll)			
Identification Problem (PRB) Action (ACT)	*Identification* categorizes the feedback instances as either a Problem (PRB) or an Action (ACT). Problem statements tend to state that a part of the paper or sentence or word is incorrect, problematic, difficult to understand, misplaced, etc. In general, these instances identify themselves by their implicitness and indirectness. Action statements tend to give the receiver of the feedback a clear instruction or clear action to change a word, improve the spelling, alter the organization, include a phrase, etc. Generally, these instances identify themselves by their explicitness and directness.	PROBLEM: "*The text is too long for an introduction and over explained.*" *"The second part of the statement beginning with .what are the main characteristics. . . doesn't suit there"* *"I don't see a traditional structure of a paragraph which should begin with a general statement and end with a little conclusion."* ACTION: *"Maybe you could try to put the main subject clearly into one sentence."* *"You should put in more discussion."* *"Try to bring out more the pros and cons of steroid use/abuse."*	In general, previous research is not conclusive about the explicit or implicitness or directness or indirectness of peer feedback. For example, directive feedback versus nondirective feedback provides different results, often culturally determined. In some studies directive feedback is considered more effective, whereas other studies consider nondirective more effective (Leijen & Leontjeva, 2012). Categorizing feedback instances to refer to a problem or state an action should be more definitive.
Intercoder reliability score of 0.84 (Cohen's Kappa) (Landis & Koch, 1977)			
Justification Yes No	The feature *Justification* (feedback instance includes a justification or explanation yes/no) categorizes feedback instances that have either include or not included a justified (or explained) opinion in the feedback. This type of feedback characterizes itself by the added phrases or sentences to the identified issues that need to be address by the receiver of the feedback. A simple or complex explanation why as reviewer we think something needs to be changed, included, revised, reorganized, etc.	Justification/Yes *"I would suggest uniting short sentences into longer ones, it would be easier to read* (justification)*."* *"You should make it more clear because it took some time for me to really figure out what you want to write about* (justification)*."* *"The only mistake I found was the last word of the essay: don't. I'd use do not instead to make it more formal* (justification)*."*	Gielen et al. (2010) and Nelson and Schunn (2009)
Intercoder reliability score of 0.88 (Cohen's Kappa).			
Alteration Yes No	Feedback that is more specific is considered to be more helpful in comparison to feedback that is more general. In the context of this study, *alteration* (adapted from Liu & Sadler, 2003) indicates whether a feedback instance points to a specific change. Alteration characterizes itself by phases such as, *for example*, or visual recasts, such as the incorrect word/phrase and the corrected word/phrase following.	*The last sentence (I cannot. . .) I would rephrase for clarity. Maybe, "So, why is it important" (alteration) or something in those lines.* *Little to primitive->* *A little toO primitive (alteration)* *"The first thing that caught my attention was the use of the word-pair rather that, which I think should be in this context rather than. The sentence is:Hydrocarbons rather that ethers can be used as a solvent."* *"The spelling of the word fumingnation isn't correct, the right spelling is fumigation."*	Ferris(1997) (1997) Liu and Sadler (2003)

Intercoder reliability score of 0.85 (Cohen's Kappa).			
Area Local Global	Feedback instances refer to either *global* areas (e.g. audience and purpose, idea development, topic, logic and support, organization of writing, thesis statement) or *local* areas (e.g. wording, spelling, grammar, sentence structure, academic style, punctuation).	Local: *"I also found a spelling mistake. It should be mankind's instead of mankinds. And that's about it I could find. I had to whine about something;)"* *"Also you already list too many notions that should be used in the main body instead."* Global: *"I think, you should support your ideas more. It is true, that no poison can kill you, if you take one molecule of it, but I think there are some, which can kill you after inhaling."*	Liu and Sadler (2003) categorize global and local changes as an area and not as a type of feedback. As this study focuses on revision-oriented comments, area (global/local) is considered to be a type of feedback.
Intercoder reliability score of 0.90 (Cohen's Kappa).			
Feedback Trait (modTrait)			
Question Yes No	Feedback instances can either include or not include a sentence, phrase or word formulating a *question.* As a feature, *question* does not need to be a direct question marked by a question mark, but can also include rhetorical questions, or statements which questions because there is doubt or confusion of something that is being referred to in the text or being stated by the reviewer.	*Can this text be considered an essay?* *Just wondering about the use of words here, it sounds like a specialized knowledge.*	The nature of question statements in communication – often used to mitigate statements – is more often associated to negative effects (Ferris, 1997; Leijen & Leontjeva, 2012; Nelson & Schunn, 2009).
Intercoder reliability score of 0.75 (Cohen's Kappa)			
Point of View (POV) AUT REV NEU	*Point of view* takes a more general perspective of the feedback instance. *POV* looks at who is the reporting subject in the feedback instance. Generally, POV can be found by identifying the subject of the sentence. In cases where there are more subjects, the most prominent subject represented in the feedback instance determines the POV, which can also be the implicit subject.	*The author has to reformulate the second sentence in the 3rd paragraph to include a . . . (AUT)* *I strongly suggest that, the author has to reformulate . . . (REV)* *Try to think of a problem statement, what do you want to discuss, argue about. . . (AUT)* *As the topic was not given, it's hard to evaluate the execution of the overview properly (NEU)*	The assumption with *POV* is that the perspective taken in the feedback comment refers to the distances reviewers take in their comments. Studies in computer mediated communication (CMC) report that communication in an online environment displays similar measures of distance, as would be in face to face communication (Morand & Ocker, 2003; Walther, 1992) which can be measured through specific features of politeness.
The initial agreement test for inter-rater reliability of 10% of the data resulted in a reliability score of 0.83 (Cohen's Kappa).			
Recurring Yes No	Recurring (YES) indicates that another reviewer has mentioned the same topic highlighted in that instance.	*Seems title is missing* *I would suggest the author to include a title, as the task required a title to be included.*	For example, two reviewers of the same paper can both comment (in different ways) that the text needs a title. Recurring is likely when multiple peer feedback is given. Research indicates the advantages of multiple peer feedback (Cho & Schunn, 2007).

Location (LOC) Yes No	Statements, words, phrases, which locate. Prepositions and nouns such as "on", "in", "page", "paragraph", "sentence", "phrase", "before", "after", "line", "beginning", "middle", "ending" etc. *But also:* Instances which are on a more conceptual level, such as *in the introduction, in the discussion, in the text*	*For example:* *"The last sentence (I cannot. . .) I would rephrase for clarity. Maybe So, why is it important or something in those lines."* *2nd paragraph, 2nd sentence – that?*	Research suggests that an important characteristic of successful feedback (i.e. feedback that leads to implementation) is that it contains language features that explicitly locates the problem in question (Leijen & Leontjeva, 2012; Nelson & Schunn, 2009; Van der Pol, Admiraal, & Simons, 2006; Xiong et al., 2012).
Revision in a subsequent draft (response variable)			
Revised Yes No	Revised is the measurable outcome of students making a change in a subsequent draft that can be linked to the peer feedback instance. Revised in the context of this study is understood as the act of altering something based on the feedback that has been provided by peers. Revised does not include the measurement of improvement.		A feedback instance is regarded as having been revised when the changes in the revised text can be directly linked to the content of the feedback. Partial changes in revised texts and changes made in later revised versions of the text were also considered having been revised.
Intercoder reliability score of 0.76 (Cohen's Kappa)			

Djuddah A.J. Leijen is a final year PhD student at the department of Estonian and General Linguistics (expected graduation year 2015), Lecturer of Academic Writing at the College of World Languages and Cultures, and Head of the Centre for Academic Writing and Communication at the University of Tartu, Estonia. His research interests include second language writing, web-based peer review, writing research methodologies, corpus linguistics, and machine learning.

References

Beason, Larry. (1993). Feedback and revision in writing across the curriculum classes. *Research in the Teaching of English*, 395–422.

Berg, Cathrine E. (1999). The effects of trained peer response on ESL students' revision types and writing quality. *Journal of Second Language Writing*, *8*(3), 215–241. http://dx.doi.org/10.1016/S1060-3743(99)80115-5

Biber, Douglas, Nekrasova, Tatiana, & Brad, Horn. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *ETS Research Report Series*, *2011*(1), i–99. http://dx.doi.org/10.1002/j.2333-8504.2011.tb02241.x

Cho, Kwangsu. (2004). *When multi-peers give better advice than an expert: The type and impact of feedback given by students and an expert on student writing*. University of Pittsburgh.

Cho, Kwangsu, & MacArthur, Charles. (2010). Student revision with peer and expert reviewing. *Learning and Instruction*, *20*(4), 328–338. http://dx.doi.org/10.1016/j.learninstruc.2009.08.006

Cho, Kwangsu, & Schunn, Christian D. (2004). The SWoRD is mightier than the pen: Scaffolded writing and rewriting in the discipline. *Proceedings. IEEE International Conference on Advanced Learning Technologies, 2004. IEEE*, pp. 545–549. http://doi.org/http://doi.ieeecomputersociety.org/10.1109/ICALT.2004.1357474

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

Cho, Kwangsu, Schunn, Christian D., & Charney, Davida. (2006). Commenting on writing typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication*, *23*(3), 260–294. http://dx.doi.org/10.1177/0741088306289261

Cho, Kwangsu, Schunn, Christian D., & Wilson, Roy W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology*, *98*(4), 891. http://dx.doi.org/10.1037/0022-0663.98.4.891

Conway, Drew, & White, John Myles. (2012). *Machine learning for hackers*. O'Reilly Media, Inc.

Crossley, Scott A. (2013). Advancing research in second language writing through computational tools and machine learning techniques: A research agenda. *Language Teaching*, *46*(02), 256–271. http://dx.doi.org/10.1017/s0261444812000547

Davis, Boyd, & Thiede, Ralf. (2000). Writing into change: Style shifting in asynchronous electronic discourse. *Network-Based Language Teaching: Concepts and Practice*, 87–120.

Dheram, Prema Kumari. (1995). Feedback as a two-bullock cart: a case study of teaching writing. *ELT Journal*, *49*(2), 160–168.

Díez-Bedmar, Maria B., & Pérez-Paredes, Pascual. (2012). The types and effects of peer native speaker's feedback on CMC. *Language Learning & Technology*, *16*(1), 62–90.

Faigley, Lester, & Witte, Stephen. (1981). Analyzing revision. *College Composition and Communication*, 400–414.

Feak, Christine, & Swales, John. (2004). *Academic writing for graduate students*. Ann Arbor, MI: University of Michigan Press.

Ferris, Dana R. (1997). The influence of teacher commentary on student revision. *Tesol Quarterly*, *31*(2), 315–339. http://dx.doi.org/10.2307/3588049

Ferris, Dana R. (2004). The grammar correction debate in L2 writing: Where are we, and where do we go from here? (and what do we do in the meantime. . .?). *Journal of Second Language Writing*, *13*(1), 49–62. http://dx.doi.org/10.1016/j.jslw.2004.04.005

Fleiss, Joseph L., Levin, Bruce, & Paik, Myunghee C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Gielen, Sarah, Peeters, Elien, Dochy, Filip, Onghena, Patrick, & Struyven, Katrien. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*(4), 304–315. http://dx.doi.org/10.1016/j.learninstruc.2009.08.007

Goldin, Ilya M., Ashley, Kevin D., & Schunn, Christian D. (2012). Redesigning Educational Peer Review Interactions Using Computer Tools: An Introduction. *Journal of Writing Research*, *4*(2), 111–119. http://dx.doi.org/10.17239/jowr-2012.04.02.1

Guardado, Martin, & Shi, Ling. (2007). ESL students' experiences of online peer feedback. *Computers and Composition*, *24*(4), 443–461. http://dx.doi.org/10.1016/j.compcom.2007.03.002

Hart-Davidson, William, McLeod, Michael, Klerkx, Christopher, & Wojcik, M. (2010). A method for measuring helpfulness in online peer review. *Proceedings of the 28th ACM International Conference on Design of Communication. ACM*, http://dx.doi.org/10.1145/1878450.1878470, pp. 115–121

Hirvela, Alan. (1999). Collaborative writing instruction and communities of readers and writers. *TESOL Journal*, *8*(2), 7–12. http://dx.doi.org/10.1002/j.1949-3533.1999.tb00169.x

Hyland, Fiona. (2000). ESL writers and feedback: Giving more autonomy to students. *Language Teaching Research*, *4*(1), 33–54. http://dx.doi.org/10.1177/136216880000400103

Hyland, Ken, & Hyland, Fiona. (2006). Feedback on second language students' writing. *Language Teaching*, *39*(02), 83–101. http://dx.doi.org/10.1017/S0261444806003399

Kluger, Avraham N., & DeNisi, Angelo. (1996). The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254.

Kundu, Suman, Aulchenko, Yurii S., van Duijn, Cornelia M., & Janssens, Cecile A. (2011). PredictABEL: an R package for the assessment of risk prediction models. *European Journal of Epidemiology*, *26*(4), 261–264. http://dx.doi.org/10.1007/s10654-011-9567-4

Kuteeva, Maria. (2013). Graduate learners' approaches to genre-analysis tasks: Variations across and within four disciplines. *English for Specific Purposes*, *32*(2), 84–96. http://dx.doi.org/10.1016/j.esp.2012.11.004

Leijen, D. A. J., & Leontjeva, A. (2012). Linguistic and review features of peer feedback and their effect on the implementation of changes in academic writing: A corpus based investigation. *Journal of Writing Research*, *4*(2) http://dx.doi.org/10.17239/jowr-2012.04.02.4

Leijen, D. A. J. (2014). Applying machine learning techniques to investigate the influence of peer feedback on the writing process. In D. Knorr, C. Heine, & J. Engberg (Eds.), *Methods in Writing Process Research* (pp. 167–183). New York: Peter Lang: Frankfurt am Main.

Leki, Ilona. (1990). Coaching from the margins: Issues in written response. *Second Language Writing: Research Insights for the Classroom*, 57–68.

Leki, Ilona. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals*, *24*(3), 203–218. http://dx.doi.org/10.1111/j.1944-9720.1991.tb00464.x

Little, David, & Perclová, Radka. (2012). *The European Language Portfolio: Teacher's Guide*. Council of Europe.

Liu, Jun, & Hansen, Jette G. (2002). *Peer response in second language writing classrooms*. Ann Arbor: University of Michigan Press.

Liu, Jun, & Sadler, Randall W. (2003). The effect and affect of peer review in electronic versus traditional modes on L2 writing. *Journal of English for Academic Purposes*, *2*(3), 193–227. http://dx.doi.org/10.1016/S1475-1585(03)00025-0

Lockhart, Charles, & Ng, Peggy. (1993). How useful is peer response. *Perspectives*, *5*(1), 17–29.

Lundstrom, Kristi, & Baker, Wendy. (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. http://dx.doi.org/10.1016/j.jslw.2008.06.002

Matsuda, Paul K., Canagarajah, Suresh A., Harklau, Linda, Hyland, Ken, & Warschauer, Mark. (2003). Changing currents in second language writing research: A colloquium. *Journal of Second Language Writing*, *12*(2), 151–179. http://dx.doi.org/10.1016/S1060-3743(03)00016-X

Min, Hui-Tzu. (2006). The effects of trained peer review on EFL students' revision types and writing quality. *Journal of Second Language Writing*, *15*(2), 118–141. http://dx.doi.org/10.1016/j.jslw.2006.01.003

Morand, David A., & Ocker, Rosalie J. (2003). Politeness theory and computer-mediated communication: A sociolinguistic approach to analyzing relational messages. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, 2003. IEEE*, http://dx.doi.org/10.1109/HICSS.2003.1173660 (p. 10–pp)

Moxley, Joseph M. (2012). Aggregated assessment and objectivity 2.0. In *Proceedings of the Second Workshop on Computational Linguistics and Writing (CLW 2012): Linguistic and Cognitive Aspects of Document Creation and Document Engineering* , pp. 19–26.

Nelson, Gayle L., & Murphy, John M. (1993). Peer response groups: Do L2 writers use peer comments in revising their drafts? *Tesol Quarterly*, *27*(1), 135–141. http://dx.doi.org/10.2307/3586965

Nelson, Mellisa M., & Schunn, Christian D. (2009). The nature of feedback: How different types of peer feedback affect writing performance. *Instructional Science*, *37*(4), 375–401. http://dx.doi.org/10.1007/s11251-008-9053-x

Nguyen, Thi T. M. (2008). Modifying L2 criticisms: How learners do it? *Journal of Pragmatics*, *40*(4), 768–791. http://dx.doi.org/10.1016/j.pragma.2007.05.008

R Core Team. (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

Russell, Arlene A. (2004). Calibrated peer review-a writing and critical-thinking instructional tool. *Teaching Tips: Innovations in Undergraduate Science Instruction*, 54.

Schnute, Jon T., Couture-Bell, Alex, & Haigh, Rowan. (2006). *PBS Modelling 1: User's Guide*. Fisheries and Oceans Canada, Science Branch, Pacific Region, Pacific Biological Station.

Topping, Keith. (2003). Self and peer assessment in school and university: Reliability, validity and utility. In *Optimising new modes of assessment: In search of qualities and standards*. Springer., pp. 55–87.

Tseng, Sheng-Chau, & Tsai, Chin-Chung. (2007). On-line peer assessment and the role of the peer feedback: A study of high school computer course. *Computers & Education*, *49*(4), 1161–1174. http://dx.doi.org/10.1016/j.compedu.2006.01.007

Tsui, Amy B., & Ng, Maria. (2000). Do secondary L2 writers benefit from peer comments? *Journal of Second Language Writing*, *9*(2), 147–170.

Tuzi, Frank. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition*, *21*(2), 217–235. http://dx.doi.org/10.1016/j.compcom.2004.02.003

Van der Pol, Jacco, Admiraal, Wilfried, & Simons, P. Robert-Jan. (2006). The affordance of anchored discussion for the collaborative processing of academic texts. *International Journal of Computer-Supported Collaborative Learning*, *1*(3), 339–357.

Vygotsky, Lev. (1978). Interaction between learning and development. *Readings on the Development of Children*, *23*(3), 34–41.

Wadhwa, Gunita, Schulz, Henry, & Mann, Bruce L. (2006). Effects of Anonymity and Accountability During Online Peer Assessment. *Selected Styles in Web-Based Educational Research*, 302.

Walther, Joseph B. (1992). Interpersonal effects in computer-mediated interaction a relational perspective. *Communication Research*, *19*(1), 52–90. http://dx.doi.org/10.1177/00936509201900100

Ware, Paige D., & O'Dowd, Robert. (2008). Peer feedback on language form in telecollaboration. *Language Learning & Technology*, *12*(1), 43–63.

Ware, Paige D., & Warschauer, Mark. (2006). Electronic feedback and second language writing. *Feedback in Second Language Writing: Contexts and Issues*, 105–122.

Wing, Max K., Weston, Steve, Williams, Andre, Keefer, Chris, & Engelhardt, Allan. (2012). *caret: Classification and Regression Training*. Retrieved from http://CRAN.R-project.org/package=caret.

Xiong, Wenting, Litman, Diane, & Schunn, Christian. (2012). Improving research on and instructional quality of peer feedback through natural language processing. *Journal of Writing Research*, *4*(2), 155–176. http://dx.doi.org/10.17239/jowr-2012.04.02.3

Zhang, Shuqiang. (1999). Thoughts on some recent evidence concerning the affective advantage of peer feedback. *Journal of Second Language Writing*, *8*(3), 321–326. http://dx.doi.org/10.1016/s1060-3743(99)80119-2