Making the Most Out of It: Maximizing Learners' Benefits from Expert, Peer and Automated Feedback across Domains

Astrid Wichmann (co-chair), Ruhr University Bochum, astrid.wichmann@rub.de,
Danielle S. McNamara (co-chair), Arizona State University, danielle.mcnamara@asu.edu

Markus Bolzer, Ludwig-Maximilians-Universität, Markus.Bolzer@psy.lmu.de
Jan-Willem Strijbos, Ludwig-Maximilians-Universität, Jan-Willem.Strijbos@psy.lmu.de
Frank Fischer, Ludwig-Maximilians-Universität, Frank.Fischer@psy.lmu.de
Moshe Leiba, Holon Institute of Technology, Moshel@hit.ac.il
Alexandra Funk, Ruhr University Bochum, Alexandra.Funk@rub.de
Nikol Rummel, Ruhr University Bochum, Nikol.Rummel@rub.de
Michaela Ronen, Holon Institute of Technology, Ronen@hit.ac.il
Olaf Peters, Technische Universität Dresden, olaf.peters@tu-dresden.de
Susanne Narciss, Technische Universität Dresden, susanne.narciss@tu-dresden.de
Hermann Körndle, Technische Universität Dresden, hermann.koerndle@tu-dresden.de
Rod D. Roscoe, Arizona State University, rod.roscoe@asu.edu
Laura K. Varner, Arizona State University, laura.varner@asu.edu
Erica L. Snow, Arizona State University, erica.l.snow@asu.edu

Discussant: Chris Quintana, University of Michigan, quintana@umich.edu

Abstract: Across a variety of domains, formative feedback is often regarded as beneficial, if not crucial to learning. Yet studies show that this assumption does not always hold true: some types of feedback do not benefit learners. This symposium brings together researchers investigating how feedback can be optimized to maximize potential benefits. The four papers include studies investigating the effectiveness of feedback from various sources including expert, peer and automatically generated feedback in the domains of writing and math. The studies use a variety of methodological approaches including behavioral studies, eye tracking, and data mining. The discussion emanating from the results to be reported during the symposium will focus on how these empirical findings can help to inform feedback delivery in the classroom and how to more effectively design automated feedback.

Symposium Description

Across a variety of domains, formative feedback is assumed to enhance learning. Yet learners often have difficulties capitalizing on feedback. A critical educational issue and a core challenge involve finding ways for students to more effectively benefit from feedback such that feedback helps students to improve performance and learning outcomes. This interdisciplinary symposium brings together research on feedback effectiveness from a variety of perspectives that examine various qualities and sources of feedback including peer, expert and automated feedback. We focus on (1) understanding students' challenges in benefiting from feedback, and (2) provide recommendations for optimizing conditions to maximize potential benefits.

Formative feedback refers to information that is provided to the learner in order to change the learner's behavior for the purpose of improving learning (Shute, 2008). Formative feedback includes information on how to direct learners' attention to response errors, response quality, or misconceptions. It can be provided by various sources, such as peers, experts, or automatically generated by computer software.

The complexity of conditions and contexts points toward the need for a multidimensional perspective on feedback to better understand students' challenges in taking advantage of feedback while learning. When exploring feedback effectiveness, there are a number of factors to consider, including (a) feedback quality and source, (b) learner characteristics, and (c) instructional settings (Narciss, 2013). Depending on the feedback quality, students may respond to feedback with skepticism and are likely to reject feedback upfront (Roscoe & McNamara, 2013). Furthermore, although feedback may potentially help learners to focus their attention on errors, it can sometimes decrease opportunities to process key information (Eva et al., 2012). Immediate feedback, for instance, has shown to be effective for enhancing performance in the short term but can be detrimental for long-term retention or skill acquisition (Goodmann & Wood, 2004). Additionally, feedback can be more or less effective depending on learner characteristics (Mathan & Koedinger, 2002). Indeed, one factor that may contribute to the inconsistency of findings across studies on feedback is the wide variability in students' prior knowledge, skills, or attitudes (Narciss, et al. 2014). Other important factors to take into account are the instructional setting (Narciss, 2013), and the amount or specificity of support provided through feedback messages (e.g., Goodman, Wood & Chen, 2011).

A further issue regards the challenge of considering assessment both in terms of online process data as well as post-task learning outcomes. A better understanding of the factors that impact feedback uptake will require considering how students interpret and understand feedback. Various methodological approaches such as eye tracking and data mining enable investigations of students' responses to feedback during the learning process. In addition, automated essay scoring enables researchers to investigate factors that influence feedback uptake at a larger scale.

The four presentations included in this symposium address and discuss these issues from a variety of perspectives. The authors' findings have both theoretical and practical implications. On a theoretical level, the results improve our understanding of the factors that influence the effectiveness of formative feedback from various sources. On a practical level, the findings inform practitioners and teachers regarding techniques to optimize students' use of feedback and guidelines for designing automated feedback.

Markus Bolzer, Jan-Willem Strijbos, and Frank Fischer investigated the impact of competence level of the peer feedback sender and content of the peer feedback on perception and essay revision performance; they took mindful (cognitive) processing of peer feedback into account and included an eye tracking measurement.

Astrid Wichmann, Moshe Leiba, Alexandra Funk, Nikol Rummel, and Miky Ronen present two studies investigating feedback effectiveness, particularly feedback uptake in the domain of academic writing. In the first study, the authors explored the impact of the assessor's expertise level (peer vs. expert), while the second study focused on the impact of sense making support. In addition to feedback uptake, trust (study 1) and revision skills (study 2) were assessed.

Olaf Peters, Susanne Narciss, and Hermann Körndle developed and evaluated a formative assessment and feedback script to support feedback generation in the domain of vocational education. The authors were particularly interested in comparing the effects of generating feedback to a peer vs. to one's own. As part of an iterative design cycle, script effects on (a) student's perceptions of assessment activities and (b) their revision activities were investigated.

Rod Roscoe, Laura Varner, Erica Snow, and Danielle McNamara evaluated the implementation and effectiveness of automated formative feedback in the Writing Pal. Feedback uptake in the Writing Pal has been iteratively improved in design experiments by reducing obstacles of unhelpful, overwhelming, or threatening feedback while maintaining a focus on formative assistance. In this study, the authors examined the quality of students' original versus revised essays, using automated tools to detect patterns of revisions implemented by students.

Each of the four presentations will include discussions of the research objectives, research questions, methods, results and implications. The four presentations will be followed by a discussion with **Professor Chris Quintana** whose renowned contributions and extensive experience in computer-based scaffolding, along with his interest in design across a variety of domains make him an excellent choice for integrating the work in this symposium. The discussion will draw on the symposium contributions to provide recommendations on how to maximize benefits from feedback.

Effects of Peer Feedback Content and Senders' Competence on Perceptions and Mindful Cognitive Processing of Written Peer Feedback: An Eye Tracking Study

Markus Bolzer, Jan-Willem Strijbos and Frank Fischer, Ludwig-Maximilians-Universität, Germany

In academic settings, writing and revising texts is a daily business. In many cases – increasingly so within university courses – students often receive feedback from a fellow student after producing a text (i.e., an essay). Peer feedback can be provided more frequently and more quickly than feedback given by one person, such as the instructor of the course (Falchikov & Goldfinch, 2000). However, there are important aspects to be taken into account. In peer feedback, many students are concerned about fairness and doubt their own and peers' skill to provide peer feedback (Van Gennip, Segers, & Tillema, 2009). Feedback quality strongly depends on feedback content, form, and function (Narciss, 2008). Strijbos et al. (2010) observed that perceptions of peer feedback were influenced by an interaction between feedback content and the competence level of the feedback sender, but no direct correlation was found between feedback perceptions and revision performance.

One important mechanism could be *mindful cognitive processing*, i.e. how deeply the presented peer feedback has been processed and understood. Several authors have emphasized the importance of mindful cognitive processing for feedback efficiency (Bangert-Drowns et al., 1991; Narciss, 2008; Poulos & Mahony, 2008; Gielen et al., 2010). A systematic investigation of mindful cognitive processing combined with the impact of feedback content and senders' competence level is still lacking. Based on the eye-mind-hypothesis of Just and Carpenter (1980), which states that what a person consciously looks at is also cognitively processed, eye tracking enables us to measure the exact amount of time (i.e., fixation duration) a person consciously views

written peer feedback. Eye tracking thus provides the means to investigate the reading process of the peer feedback and obtain more insight into mindful cognitive processing.

The present study investigates the impact of peer feedback content and competence level of the feedback sender on feedback perception, revision performance, and feedback recall. Furthermore, mindful cognitive processing will be investigated, as (a) the relationship between fixation duration and revision performance and (b) the relationship between fixation duration and feedback recall. We expect that more elaborated peer feedback and higher competence of the feedback sender lead to a more positive perception of the peer feedback, better revision performance, better feedback recall, and to more mindful cognitive processing. We expect a positive relationship between fixation duration, revision performance, feedback recall.

Method

Forty-five psychology students (10 male, 35 female) participated in a laboratory study. In a 2 × 2 factorial design, participants received a scenario that varied in feedback content (concise general feedback [CGF] vs. elaborated specific feedback [ESF]) and competence level of the feedback sender (high vs. low). In the beginning, the participants received information on academic writing including four writing criteria (simplicity, structure, conciseness, stimulation) followed by an essay with the task of imagining that this essay was their own product. After reading the essay, they received on-screen written peer feedback by a fictional peer together with information about the competence level of that peer. The peer feedback was either CGF or ESF, and was based on the four writing criteria. During the feedback reading phase, data about how the peer feedback was read – i.e., fixation duration – was gathered via a head-mounted eye tracker, to infer mindful cognitive processing. Data provided by the eye tracker included how much time the participants spent on reading criteria or content, together with the overall fixation duration on the peer feedback.

After reading the peer feedback, participants completed a questionnaire assessing peer feedback perceptions. The questionnaire consisted of 18 items referring to four scales with three items each: (a) fairness (Cronbach's $\alpha = 0.90$), (b) usefulness ($\alpha = 0.94$), (c) acceptance ($\alpha = 0.84$), and (d) willingness to improve ($\alpha = 0.84$), and one scale with six items: (e) affect ($\alpha = 0.77$). The fairness, usefulness, and acceptance scales constitute 'Perceived Adequacy of Feedback' (PAF) ($\alpha = 0.93$). During the next step, participants received the essay again and were asked to use the peer feedback to revise the essay. Each correct improvement to a correctly identified error received 1 point. A maximum of 29 errors, which were artificially inserted into the essay, could be identified by the participants. Revision performance for each participant was calculated as the total number of points divided by time needed for revision. After a 10 minute distraction phase of solving Sudoku puzzles, participants engaged in a free recall task during which they were asked to write down everything they recalled of the on-screen written peer feedback. In total there were 11 aspects to recall for CGF and 24 aspects for ESF. Each correctly recalled aspect received 1 point (maximum in CGF: 11; maximum in ESF: 24). Recall for each participant was calculated as the total number of points divided by the maximum amount of points to achieve comparability of both conditions. Mindful cognitive processing was operationalized as the correlation between fixation duration (total time focused on the on-screen peer feedback) with (a) revision performance and (b) feedback recall.

Results

We found a significant main effect for competence level of the peer feedback sender on PAF: $F(1, 44) = 9.25$, $p = .004$, $\eta^2 = .18$, i.e. feedback from a high competent peer ($M_{high} = 6.35$; $SD = 1.79$) was perceived as more adequate than feedback from a low competent peer ($M_{low} = 4.49$; $SD = 2.20$). Furthermore, we found a significant main effect for peer feedback content on affect: $F(1, 44) = 5.49$, $p = .024$, $\eta^2 = .12$., i.e. ESF leads to a more positive affect ($M_{ESF} = 5.24$; $SD = 1.67$) than CGF ($M_{CGF} = 4.16$; $SD = 1.31$). Secondly, no significant effects were found for competence level and/ or peer feedback content on revision performance or feedback recall. Descriptives, however, showed a tendency towards better feedback recall in the CGF condition: $M_{CGF} = 0.38$ ($SD = 0.26$), $M_{ESF} = 0.26$ ($SD = 0.16$), $d = .56$. Thirdly, no significant effects were found for competence level and/ or peer feedback content on fixation duration. With respect to mindful cognitive processing, a significant negative correlation was found in both conditions between fixation duration and revision performance: $r = -.51$*, $p = .012$ (ESF), and $r = -.54$*, $p = .010$ (CGF), whereas fixation duration was uncorrelated to feedback recall in both conditions. Nevertheless, revision performance shows a significant positive correlation to feedback recall across conditions: $r = .31$*, $p = .040$.

Discussion and Outlook

The aim of this study was to obtain more insight into the perceptions and mindful cognitive processing of written peer feedback. In line with Strijbos et al. (2010), ESF was perceived as more adequate than CGF, and feedback from a more competent peer was perceived as more adequate than feedback from a less competent peer. In contrast, where Strijbos et al. found an interaction between peer feedback content and competence of the sender for affect, this study only found a main effect for peer feedback content, i.e. that ESF leads to more

positive affect. No significant effects were found for revision performance or feedback recall, although descriptives showed a tendency towards better recall in the CGF condition. There were no significant differences in fixation duration while reading the peer feedback. Counter intuitively, with respect to mindful cognitive processing, we found a significant negative correlation between fixation duration and revision performance in both conditions and no correlation between fixation duration and feedback recall. The correlation between revision performance and feedback recall was significant and positive. On the one hand, spending more time reading the peer feedback seems to inhibit mindful cognitive processing and results in low application of the peer feedback. On the other hand, participants were still able to recall some aspects of the peer feedback, which shows that at least a basic amount of mindful cognitive processing occurred. A possible conclusion is that the participants suffered from an overload. Shute (2008, p. 177) claims it is important to "provide elaborated feedback in small enough pieces so that it is not overwhelming and discarded". Additional insights into the perceptions and mindful cognitive processing of written peer feedback could add to the efficiency of peer feedback practice in university courses.

Based on these findings, we recently completed the data collection of a follow-up study with a slightly different design: the on-screen peer feedback is presented simultaneously with the essay and contains justifications in one condition. Justifications were added to the peer feedback, as their presence appears to increase revision performance (Gielen et al., 2010). Simultaneous presentation allows for integrative transitions between essay and peer feedback, which serve as an additional measure to infer mindful cognitive processing (Mason, Pluchino, Tornatora, & Ariasi, 2013). Finally, given the possible overload observed in the initial study, a self-report measure for perceived cognitive load was included after reading the peer feedback, essay revision, and feedback recall. Results will be presented at the conference.

Investigating Feedback Uptake by Looking at the Assessor's Level of Expertise and Providing Sense-Making Support

Astrid Wichmann, Alexandra Funk and Nikol Rummel, Ruhr University Bochum, Germany
Moshe Leiba and Michaela Ronen, Holon Institute of Technology, Israel

In this paper, we report results of two studies investigating feedback effectiveness by examining students' feedback uptake in the domain of academic writing. Drafting and revising texts is a challenging task for students especially in the beginning of their academic careers (Wichmann & Rummel, 2013). Peer feedback from an assessor can help the assessee to meet the challenges of successful revision, because feedback can be provided more quickly and frequently (Falchikov & Goldfinch, 2000). However, students often exhibit poor feedback uptake and thus fail to capitalize on feedback they receive (Van der Pol, Van den Berg, Admiraal, & Simons, 2008). Feedback uptake refers to changes made to the assessee's product during revision. On the one hand, students' problems of feedback uptake might be related to available information regarding the assessors' expertise. Several studies have investigated the impact of the assessors' expertise, comparing peer feedback with teacher feedback (Leki, 1991) or comparing different levels of expertise (Strijbos, Narciß & Dünnebier, 2010). There is agreement that the level of the assessors' expertise affects assessees' perceptions of the feedback and their trust. In other words, depending on whether feedback comes from a peer or a teacher, students might trust the assessor to different degrees. Trust might then affect feedback uptake. On the other hand, problems of feedback uptake might be related to deficits in students' understanding of the feedback and lack of reflection. Writers often reject feedback upfront without engaging in sense-making processes or have problems with managing the feedback (Boero & Novarese, 2012). However, sense-making processes are crucial because understanding the problems the feedback relates to is necessary for improving performance (Nelson & Schunn, 2009). Support is needed for students to make sense of feedback with the goal to prevent feedback rejection, to increase feedback understanding, to organize and plan steps to correct detected problems, and thus to improve feedback uptake. Thus, supporting sense making should improve students' revision skills. Based on the students' problems of feedback uptake in academic writing, we conducted two studies: In Study 1, we explored the impact of the assessor's expertise level (as labeled in the feedback) on feedback uptake and on students' trust. In Study 2, we investigated the influence of sense-making support on feedback uptake and revision skills. We expected that providing assessees with sense-making support as they received feedback would improve feedback uptake and revision skills.

Methods

Study 1 was conducted in an authentic setting with first year undergraduate students who were enrolled in an Academic Literacy course in Israel (52 students 21 male, 31 female). As part of the course requirements (20% of the course overall grade), students wrote a short essay according to criteria of academic writing. After having submitted their first version, each student was asked to review two of the peers' essays (presented anonymously) and to provide constructive feedback for improving these essays. This resulted in two feedback comments for each essay, which differed in quality and specificity. Next, one of the two feedback comments for each essay

was randomly labeled as "expert" or "peer". Each student received the two labeled feedback comments and was asked to revise his or her essay accordingly and to submit a final version. The study was conducted online over a period of 21 days. The CeLS environment (Ronen & Kohen-Vacs, 2010) was used for orchestrating the activity: selectively administering the essays and feedback comments and presenting the instructions and research instruments. Feedback uptake and trust were assessed as dependent variables. Feedback uptake was measured by counting the instances of text change divided by the instances of concrete comments given. Trust was assessed using a self-report questionnaire including 8 items on a 1 to 5 Likert scale. In addition, feedback specificity was measured because it differed across expertise levels. It was assessed by counting the instances of concrete comments given divided by the instances of all the comments given (specific and general).

Study 2 comprised 67 (13 male, 54 female) participants. The students were recruited from three courses of the bachelor program in educational sciences at a major German university. Students participated in the study as part of their regular course activities. The study followed an experimental design with sense-making support as independent variable. Students were randomly assigned to one of two conditions: Sense-Making Support (SMS+) condition and No-Sense-Making Support (SMS-) condition. Sense-making support aimed at encouraging the student to reflect on the feedback. Students were asked to rank and to judge received feedback and to plan their corresponding revisions. The academic writing task was conducted over a period of 10 days in which students created a draft in MS Word, received feedback from an assessor, and revised the essay. The feedback included 12 comments on errors that frequently occur. Students were informed that the feedback was given to them by a peer; however, the feedback was given by trained tutors in order to control for the amount and kind of feedback. We used Moodle (Moodle, 2013) to distribute instructions and questionnaires. Feedback uptake and revision skills (Pre-Post) were assessed as dependent variables. Feedback uptake was assessed by counting the instances in which a text change was successfully made. Revision skills were assessed using counterbalanced pretest and posttest versions. The pretests and posttests assessed two distinct skills related to academic writing: problem detection and problem correction.

Results

Study 1 (focus on assessor's level of perceived expertise): As expected, students trusted the experts more than they trusted their peers *($t(51) = 2.89$, $p < 0.01$)*. In general, students who trusted the feedback more also considered feedback comments to be more helpful (r(expert) = .39, r(peer) = *.66, $p < .01$)*. Students' trust of the expert was not related to the specificity of the feedback. Concerning feedback uptake, students who received and reacted to specific feedback (N = 40) tended to correct more feedback labeled as "expert" than feedback labeled as "peer" ($t(39) = 2.02$, $p < .05$). In general, students' feedback uptake was higher if the specificity of the feedback was higher (r(expert) = .45, r(peer) = .46, $p < .01$).

Study 2 (focus on sense-making support): Unexpectedly, we did not find significant differences between the conditions SMS- and SMS+ concerning feedback uptake, $F = 1.88$, $p = .18$. In general, participants used a large amount of the feedback comments (81.43%) to make changes. Participants made successful changes with respect to 50.16% of the feedback comments. For revision skills, we did not find a significant difference between conditions concerning revision skills (problem detection skill: M = 41.72, SD = 16.97, F = .08, p = .78, problem correction skill: M = 72.1, SD = 15.35, F = .74, p = .39).

Conclusion

The main goal of both studies was to investigate feedback effectiveness by examining students' feedback uptake as a function of different levels of assessor's perceived expertise (peer and expert) and different levels of sense-making support. In study one, students showed higher trust towards feedback comments labeled as expert feedback. The expert is trusted due to declared authority (the expert title) while the peer is trusted according to his/her professional authority (more specific feedback comments). One reason might be that an expert is viewed as more capable of detecting problems than a peer and thus is more likely to help improve the essay. We found that students' feedback uptake was higher when the specific feedback was attributed to an expert source than when attributed to a peer. In line with research on feedback content, students' feedback uptake increased with feedback specificity (Narciss, 2013). In study two, we attempted to improve feedback uptake by providing sense-making support. We cannot be sure that our sense-making support helped students during feedback uptake or with revision skills acquisition. One reason may be that feedback was easy to take up because feedback comments were very specific and students had little problems understanding it. This might have made sense-making support redundant. In general, we found interesting results on feedback uptake and revision skills. Feedback uptake was surprisingly high, given relatively low uptake rates in other studies of peer assessment (e.g. Van der Pol et al., 2008). In the revision skills posttest, we found that students scored low on problem detection and high on problem correction. In other words, it seemed easy for students to correct errors based on feedback during the activity and during the posttest. Yet, results from the posttest indicate that students had difficulties with detecting errors.

From these studies we can conclude that feedback effectiveness depends on the way feedback is presented. Besides looking at feedback uptake (how students change text based on feedback they receive), future studies should assess how well students detect errors during writing and explore how students can be supported to better detect errors.

Development and Evaluation of a Formative Assessment and Feedback (FAF) Script to Support Generating Feedback to a Peer versus to One's Own Performance

Olaf Peters, Susanne Narciss and Hermann Körndle, Technische Universität Dresden, Germany

Numerous studies have documented the benefits of giving peer feedback (e.g., Cho & Cho, 2011) and its impact on self-assessment processes (Topping, 1998). This study aims at transferring these findings to vocational education, namely apprenticeship of metal cutting mechanics. In order to carry out the cutting mechanic's job efficiently, craft skills and planning skills are required – planning the production procedures and creating the programs for Computerized Numerical Control (CNC) machines is a core task for cutting mechanics. Novices in this field often struggle with this complex task (Berner, 2009) because there can be more than one correct planning solution. Thus, it might be a powerful strategy to implement systematic peer feedback to support the acquisition of planning skills. By comparing the approach of their peers with their own planning processes, students can reflect on the strengths and weaknesses of different planning solutions. Such comparisons may improve their own subsequent planning.

Importantly, peer assessment activities do not always improve learning (Kollar & Fischer, 2010). One reason might be that, in particular for complex tasks, students have difficulties assessing their own performance, or assessing peer performance and generating useful formative feedback. Thus, they may consider the assessment activities as useless and inadequate. Following the argumentation of Falchikov (2005) and Dochy, Segers and Sluijsmans (1999), the implementation of a feedback script might support students in assessing and generating feedback. If this is the case, providing a feedback script should also have a positive effect on students' perceptions of their assessment activities, and help them to use the assessment activities for revising and improving their solution. However, these assumptions may only hold true if the feedback script provides information that is perceived as useful and necessary for the assessment or for performance improvement on the given task.

The goal of this project is to develop a formative assessment and feedback script (FAF-script) to support feedback generation. We are particularly interested in comparing the effects of generating feedback to a peer vs. to one's own performance. As part of an iterative design cycle, we investigate script effects on (a) student's perceptions of assessment activities and (b) their revision activities.

Method

In a first step, we iteratively developed two versions of a feedback script in collaboration with experienced vocational teachers using insights from Gan's work on assessment scripts (Gan, 2011). The first script comprised general guidelines on how to provide the feedback. The second script offered general guidelines and specific hints regarding assessment criteria and standards for feedback generation. The acceptance of the two feedback script versions was evaluated with 21 vocational students (male, second year of apprenticeship). Results of this study revealed that the script version that included specific hints was perceived as more useful than the more general version. However, students stated that they found it difficult to keep all of the criteria in mind because the feedback script with both general guidelines and specific hints was presented on a separate sheet of paper.

Based on these findings, a tabular version of the assessment task was developed that included prompts on how to proceed for assessing a cutting-mechanic plan. Cues mentioning the specific assessment criteria script were provided on a note card (hereafter referred to as formative assessment and feedback (FAF) script. The FAF-script was evaluated with regard to feedback quality and feedback perception by the peer assessors. Moreover, we examined if there were differences in students' implementation of the FAF-script when used to assess and generate feedback to either a peer vs. their own planning approach. Furthermore, we explored (a) students' perceived usefulness of generating peer feedback with the FAF-script, (b) students' perceptions of adequacy of the feedback they have generated with the help of the FAF-script, and (c) correlations between perceived usefulness of generating feedback and features of the peer's planning approach (i.e., number of planning ideas).

Participants, design and procedure

18 apprentice cutting mechanics students (1 female and 17 male students in third year of apprenticeship) participated in the evaluation study. In the first session, all subjects planned the manufacturing process of a typical work piece and produced a working plan for all relevant steps. In the second session, students were

randomly assigned to two groups (generating feedback to a peer vs. one's own working plan – peer feedback vs. internal feedback). In the peer feedback group, students were provided with the working plan of a fictitious peer that included typical errors and were asked to generate feedback on this plan. To do so they were offered the FAF-script that included assessment criteria such as the selection and order of operations, the selection of tools and the specification of technological data (e.g. cutting speed or tool adjusting). Afterwards, students had to assess and generate internal feedback on their own planning, and were asked to revise their own initial plans on the basis of what they learned in the peer feedback process. Under the internal feedback condition, students also received the FAF-script but only assessed and generated internal feedback on their own planning, and were asked to revise their own initial plans if they thought it would improve them.

Measures

Students' perceptions of generating and providing peer feedback were measured by an adapted version of the *feedback perception questionnaire* (Strijbos, Narciss, & Dünnebier, 2010). The adapted items measured how students perceived the peer feedback adequacy they had generated in terms of the scales fairness (Cronbach's α = .82), usefulness (Cronbach's α = .61), and acceptance (Cronbach's α = .67). Students had to respond to these items on a 10 centimeters bi-polar scale from 0 "I fully agree" to 10 "I fully disagree". Furthermore, we assessed the perceived usefulness of giving peer feedback to improve assessors' own planning with 7 items (e.g., generating peer feedback was very helpful for revising my own plan; Cronbach's alpha .83). *Peer feedback quality* was measured (a) by the number of errors detected and (b) how it addressed the criteria of the FAF-script. *Revision performance and activities* were measured by the number of errors detected and corrected.

Results

Students in the peer feedback group stated that providing feedback on a peer's working plan was helpful for revising their own planning (median = 4.0, *IQR* = 2.0). Most apprentices stated that they used the criteria offered by the feedback script to generate the peer feedback (median = 4.0, *IQR* = 1.0). On average students stated that they used these criteria to generate internal feedback and to revise their own planning, but the variance between students regarding their use of criteria was high (median = 4.0, *IQR* = 3.0). A preliminary analysis of peer assessors' internal and peer feedback quality revealed that peer assessors detected and revised most errors in the peer plan (M = 15.56, SD = 5.59) but rarely detected errors in their own planning (M = 1.89, SD = 2.42). This difference was statistically significant (Z = -2.67, p = .008, φ = -.88). The detected errors mostly related to the criteria selection and order of operations. These impressions need to be confirmed in an in-depth analysis of the individual planning and feedback quality in both groups with regard to the type of revised working steps and errors that had or had not been detected.

Peer assessors perceived that the feedback they generated based on the feedback script was useful (M = 1.91, SD = 1.13) and fair (M = 3.50, SD = 2.53), and that they would accept the given peer feedback (M = 2.08, SD = 1.13). A correlation analysis revealed a significant correlation between the use of planning ideas offered in the peer working plan and the perceived usefulness of providing feedback on the peer working plan to improve the assessors' own planning (r = .80, p = .01). Additionally a significant correlation was found between the use of the feedback script for providing peer feedback and the perceived fairness of the given peer feedback (r = .61, p = .03).

Conclusions and Outlook

The results of the studies above indicate that the students consider it helpful to generate feedback to a peer because in doing so they realized potential improvements they could implement at in their own work. The FAF-script was not perceived as a redundant burden but as a helpful guide for providing useful peer feedback. When using the FAF-script, students felt confident that their given feedback was perceived as fair, which was important for overcoming fears of being unable to provide high-quality peer feedback.

As a next step of this iterative design cycle, we initiated another study with 107 apprentices of cutting mechanics. Within this study, a quasi-experimental two-factorial design was used to investigate the influence of two factors, (a) generating feedback on a peer's plan vs. one's own plan, and (b) feedback generation with vs. without the FAF-script. Students had to accomplish a slightly different planning task. Analogous to the exploratory study, the effects of the conditions were investigated with regard to feedback quality, feedback perception, and revision performance in order to investigate the issue of what are the benefits of generating peer vs. internal feedback for a complex vocational task. The results of this study support the findings of the pilot study. Again students considered it helpful to use the FAF-script for peer- and self-revision activities. Students with the FAF-script perceived their feedback quality as significantly more fair, useful and acceptable than students without the script. Most students stated that they adopted ideas from the peer draft to revise their own plans. Results of both studies will be presented and discussed at the ICLS 2014.

Designing Usable Automated Formative Feedback for Intelligent Tutoring of Writing

Rod D. Roscoe, Laura K. Varner, Erica L. Snow and Danielle S. McNamara, Arizona State University, USA

Student Resistance to Automated Writing Feedback

Many educators have voiced doubts about the validity of automated writing evaluation (Deane, 2013). Students have expressed similar concerns, which hinder their implementation of automated feedback when revising (Grimes & Warschauer, 2010). For example, in the Writing Pal (W-Pal) tutoring system (Roscoe & McNamara, 2013; Roscoe, Varner, Weston, Crossley, & McNamara, in press), students can write prompt-based essays and receive automated scores and feedback. Prior feasibility research revealed student doubts about the feedback system. Over 140 high school students used an early version of W-Pal in their English classes for a school year. Although students' writing improved and most students rated the system as easy to use, some students reported difficulty with reading the feedback (23.7% of students), using the feedback (24.6%), understanding the feedback (38.3%), or quantity (50.4%). Students' critiques related to specificity ("the feedback needs to be more helpful for us on our own personal essay") and usable recommendations ("W-Pal never really tells you what you need to improve on"). Such challenges are exacerbated by students' reluctance to make substantive, document-level revisions rather than surface, word-level edits (Crawford, Lloyd, & Knoth, 2008; Fitzgerald, 1987).

Responding to Students' Concerns about Automated Feedback

To create software that supports writing instruction and promotes substantive revising, students' doubts about automated feedback must be allayed via careful design. Guided by research on revising and formative feedback (Fitzgerald, 1987; Shute, 2008), W-Pal feedback focuses on high-level writing goals and recommendations (e.g., elaboration of ideas); W-Pal provides no lower-level feedback on spelling, grammar, or punctuation errors. In response to student reactions to W-Pal, feedback messages were rewritten to ensure that *problem identification* statements (i.e., issues to address) were made in an impersonal and suggestive manner. Instead of stating, "*Your* essay *does* not" the feedback states "*This* essay *may* not" In contrast, recommendations for *problem resolution* (i.e., actions to improve the text) were rewritten to be personal and specific. For instance, rather than stating, "*A* good strategy for linking ideas" the feedback states "One way that *you* can link *your* ideas" Yet another change granted students more *control* over feedback quantity. All students receive one message on the most critical issue (Initial Topic). Students can then request more feedback on that Initial Topic and/or feedback on the next most serious problem (Next Topic). Such changes sought to improve feedback uptake by reducing obstacles of unhelpful, overwhelming, or threatening feedback while maintaining a focus on formative assistance.

The revised W-Pal (Roscoe et al., in press) was evaluated with a new sample of high school students. In this report, we consider (1) the quality of students' original versus revised essays and (2) the nature of students' revising. An innovative contribution of this work is that revision patterns were assessed via automated computational tools rather than time-consuming annotation by human raters.

Method

High school students (n = 87) participated in a 10-session program using W-Pal. The first and final sessions collected data on individual differences and overall writing proficiency. Students began each of eight training sessions by writing a prompt-based, persuasive essay. A different argument topic was assigned each day, in the following order: *Planning*, *Winning*, *Patience*, *Heroes*, *Perfection*, *Uniformity*, *Beliefs*, and *Fame.* Students were allotted 25 minutes to draft their persuasive essays and 10 minutes to revise after receiving feedback.

Essay quality was assessed using scoring algorithms that rate essays on a 6-point scale similar to the SAT exam. Linguistic features of essays were assessed via Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, in press) and focused on indicators of *word-level* and *document-level* revisions. Word-level revisions tend to capture superficial edits, such as replacing short words with long words (e.g., increase *average syllables per word*). In contrast, document-level revisions may capture more substantive changes, such as improving semantic cohesion across paragraphs (e.g., increase *LSA scores among paragraphs*).

Results

A 2 (draft) x 8 (prompt) repeated measures ANOVA was conducted to examine scores across original and revised drafts and across prompts. A main effect of draft indicated that students, on average, made small (d = .12) but significant improvements when revising their essays, $F(1,74) = 15.42$, $p < .001$. Original drafts earned a mean score of 2.56 (SD = 1.04) whereas revised drafts earned a mean score of 2.68 (SD = 1.03). The quality of drafts also increased gradually over time, $F(1,74) = 9.24$, $p = .003$. For example, mean scores for students' final practice essay (M = 2.82, SD = 1.10) were higher (d = .47) than those on the first essay (M = 2.34, SD = .92).

In what ways did students revise? Did students focus on surface word-level revisions or implement more substantive document-level revisions? Results indicate that word-level revisions were less common than document-level revisions. For a few prompts (e.g., *Planning*, *Uniformity*), students made several revisions that

replaced common words with less frequent words or removed first-person pronouns. However, for many essays (e.g., *Heroes*, *Perfection*, *Fame*) our automated tools detected minimal word-level revising. In contrast, students frequently implemented revisions at the document-level. For almost every topic, students added substantive content (e.g., examples), improved organization (e.g., paragraph structure), and improved essay cohesion (e.g., linking ideas across paragraphs). Such results run counter to traditional findings that students avoid revising or implement mainly superficial edits (Crawford et al., 2008; Fitzgerald, 1987).

Conclusion and Outlook

Supporting students' use of feedback to revise in computer-based settings faces key challenges: students doubt the validity of usefulness of the feedback, and students seem to naturally resist making revisions rather than superficial edits. To address such obstacles, formative feedback in W-Pal emphasizes higher-level processes and strategies rather than lower-level concerns, which was echoed in the observed patterns of revisions. Students appeared to use W-Pal feedback to implement revisions that improved the document as a whole (e.g., developing deeper text cohesion) rather than word choice (e.g., including more rare words). In the field of automated writing evaluation, these findings suggest that formative feedback designed to be (1) strategy-oriented, (2) actionable, (3), specific, and (4) student-controlled, may be an effective means of supporting feedback uptake. In future analyses, we will examine user surveys to reveal students' specific perceptions of the updated W-Pal system, such as whether the feedback is viewed as more helpful or understandable.

Additional research is ongoing to expand the types and content of feedback offered in W-Pal and related systems. For example, the work presented here demonstrated that aspects of student revising could be captured via automated tools. Such revision patterns can be incorporated in novel feedback algorithms. That is, instead of providing feedback only on the discrete *products* of student writing (i.e., individual essay drafts), we can now begin to give automated feedback on the writing *process* (i.e., the transformation of drafts via revising).

References

Bangert-Drowns, R., Kulik, C., Kulik, J., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research*, *61*(2), 213-238.

Berner, B. (2009). Learning Control: Sense-Making, CNC machines, and changes in vocational training for industrial work. *Vocations and Learning, 2(3)*, 177-194.

Boero, R., & Novarese, M. (2012). Feedback and learning. Encyclopedia of the sciences of learning (pp. 1282-1285).

Cho, Y. H. & Cho, K. (2011). Peer reviewers learn from giving comments. *Instructional Science*, 39, 629-643.

Crawford, L., Lloyd, S., & Knoth, K. (2008). Analysis of student revisions on a state writing test. *Assessment for Effective Intervention*, *33*, 108-119.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7-24.

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

Eva, K. et al. (2012). Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv in Health Sci Educ, 17*, 15-26

Falchikov, N. (2005). *Improving assessment through student involvement: practical solutions for aiding learning in higher and further education*. London/New York: RoutledgeFalmer.

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

Fitzgerald, J. (1987). Research on revision in writing. *Review of Educational Research*, *57*, 481-506.

Gan, M. (2011). *The effects of prompts and explicit coaching on peer feedback quality.* (Unpublished doctoral dissertation), University of Aukland. Retrieved from https://researchspace.auckland.ac.nz/handle/2292/6630 (05/11/2013).

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction*, *20*, 304-315.

Goodman, J. S., & Wood, R. E. (2004). Feedback specificity, learning opportunities and learning. *Journal of Applied Psychology, 89*, 809–821.

Grimes, D., & Warschauer, M. (2010). Utility in a fallible tool: A multi-site case study of automated writing evaluation. *Journal of Technology, Learning and Assessment*, *8*, 1-43.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329-355.

Kollar, I. & Fischer, F. (2010). Peer assessment as collaborative learning: A cognitive perspective. *Learning and Instruction, 20*(4), 344-348

Leki, I. (1991). The preferences of ESL students for error correction in college-level-writing classes. *Foreign Language Annals, 24*, 203-218.

Mason, L., Pluchino, P., Tornatora, M., & Ariasi, N. (2013). An eye-tracking study of learning from science text with concrete and abstract illustrations. *The Journal of Experimental Education*, *81*(3), 356-384.

Mathan, S. A., & Koedinger, K. R. (2002). An empirical assessment of comprehension fostering features in an intelligent tutoring system. In S. A. Cerri, G. Gouarderes, & F. Paraguacu (Eds.), Intelligent tutoring systems, 6th international conference ITS 2002: Vol. 2363. Lecture notes in computer science (pp. 330–343), New York: Springer-Verlag.

McNamara, D. S., Graesser, A. C., McCarthy, P., Cai, Z. (in press). *Automated evaluation of text and discourse with Coh-Metrix.* Cambridge: Cambridge University Press.

Moodle Pty. Ltd. (2013). Moodle (Version 2.3.3) [Learning environment software]. Retrieved from https://moodle.org/.

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

Narciss, S. (2013). Designing and evaluating tutoring feedback strategies for digital learning environments on the basis of the interactive tutoring feedback model. *Digital Education Review, 23,* 7-26.

Narciss, S., Sosnovsky, S., Schnaubert, L., Andrès, E., Eichelmann, A., Goguadze, G., & Melis, E. (2014). Exploring feedback and student characteristics relevant for personalizing feedback strategies. *Computers & Education, 71*, 56–76.

Nelson, M.M. & Schunn, C.D. (2009). The nature of feedback: how different types of peer feedback affect writing performance. *Instructional Science, 37*, 375-401.

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: The students' perspective. *Assessment & Evaluation in Higher Education*, *33*, 143-154.

Ronen, M. & Kohen-Vacs, D. (2010). Modeling, enacting sharing and reusing online collaborative pedagogy with CeLS. In: Persico, D. & Pozzi, F. (Eds.), Techniques for Fostering Collaboration in Online Learning Communities: Theoretical and Practical Perspectives, IGI Global.

Roscoe, R. & McNamara, D.S. (2013). Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*(4), 1010-1025.

Roscoe, R. D., Varner, L. K., Weston, J. L., Crossley, S. A., & McNamara, D. S. (in press). The Writing Pal Intelligent Tutoring System: usability testing and development. *Computers and Composition*.

Shute, V. J. (2008). Focus on Formative Feedback. Review of Educational Research, 78(1), 153 -189

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*, 291-303.

Topping, K. (1998). *Peer assisted learning*. Mahwah, NJ: Erlbaum.

Wichmann, A., & Rummel, N. (2013). Improving revision in wiki-based writing: Coordination pays off. *Computers & Education, 62*(0), 262-270.

Van der Pol, J., Van den Berg, B. A. M., Admiraal, W. F., & Simons, P. R. J. (2008). The nature, reception, and use of online peer feedback in higher education. *Computers and Education, 51*, 1804-1817

Van Gennip, N. A. E., Segers, M. S. R., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review*, *4*, 41-51.